Development of Immunoinformatics Methods for Improved Rational Identification of T cell Epitopes

Povlsen, Helle Rus

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Development of Immunoinformatics Methods for Improved Rational Identification of T cell Epitopes

Helle Rus Povlsen

PhD Thesis

July 2022

Preface

This PhD thesis was prepared in the Immunoinformatics and Machine Learning group, the section of Bioinformatics, Department of Health Technology at the Technical University of Denmark (DTU), as a requirement for obtaining the PhD degree.

The work presented in this thesis was done from April 2019 to July 2022 under the supervision of Professor Morten Nielsen and co-supervision of Associate Professor Leon Eyrich Jessen. This PhD was funded with a grant from Immune Epitope Database.

Helle Rus Povlsen
Kongens Lyngby, July 2022

Publications Included in the Thesis

PAPER I

SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial $CD8^+$ T cell activation in COVID-19 patients

Sunil Kumar Saini, Ditte Stampe Hersby, Tripti Tamhane, Helle Rus Povlsen, Susana Patricia Amaya Hernandez, Morten Nielsen, Anne Ortved Gang, Sine Reker Hadrup

Published in: Science Immunology, Volume: 6, Issue: 58, Pages: 1-15, Year: 2021

PAPER II

ATRAP - Accurate T cell Receptor Antigen Pairing through data-driven filtering of sequencing information from single-cells

Helle Rus Povlsen*, Amalie Kai Bentzen*, Mohammad Kadivar, Leon Eyrich Jessen, Sine Reker Hadrup*, Morten Nielsen*

Submitted to: eLife, Year: 2022

PAPER III

Benchmarking data-driven filtering approaches for single-cell screening of T cell specificity

Helle Rus Povlsen, Morten Nielsen

Data from an on-going project

* Contributed equally to the work

Publications not included in the Thesis

NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data
Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentze, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, Sine R. Hadrup, Ole Winther, Bjoern Peter, Leon Eyrich Jessen, Morten Nielsen
Published in: Communications Biology, Volume: 4, Issue: 1, Pages: 1-13, Year: 2021

* Contributed equally to the work

Summary

The research projects presented in this thesis are centered around T cell specificity. T cells play a crucial role in maintaining health by eliminating intruding pathogens and malignant cell changes. This ability is granted via the T cell receptor (TCR), which interacts with peptides presented by MHC molecules on the surface of host cells. To ensure broad protection against any potential pathogen, the immune system has evolved to generate highly diverse TCRs which may recognize a wide range of targets. However, such a complex system is inevitably very challenging to study. Nevertheless, this thesis has been dedicated to investigate T cell specificity via popular experimental methods and develop immunoinformatic tools and analyses to enhance the yield of such methods.

A commonly used method for assaying T cell specificity is peptide-MHC (pMHC) multimer staining, which procures the distribution of T cells responding to given peptides of a panel. This method was applied to map SARS-CoV-2 epitopes across cohorts of infected and healthy individuals, in the first project of this thesis. We identified several immunodominant epitopes even in healthy individuals, which suggest strong influence of cross-reactive T cells primed for other, perhaps similar, antigens.

However, multimer staining only provides shallow insight into the complexity of TCR recognition of pMHCs. In order to truly understand the rules that govern T cell specificity, we employed single-cell sequencing, enabling the capture of TCR$\alpha\beta$-chains, the cognate pMHC provided by DNA-barcoded multimers, and hashing antibodies in the second project of the thesis. As single-cell data is polluted with multiple confounding factors, the key aim was to develop a method to efficiently remove noise and retain accurate pairing of TCR-pMHC.

In the third and final project, we benchmarked the previous project against a recently released method to learn the advantages and disadvantages of each approach. The two methods distinctively differ by their prioritization between specificity and sensitivity of detecting TCR-pMHC pairs.

Resumé

Studierne præsenteret i denne afhandling, er alle centreret omkring T celle specificitet. T celler spiller en afgørende rolle i forbindelse med opretholdelsen af den raske krop ved at eliminere indtrængende pathogener og ondartede celleforandringer. Evnen til dette er givet via T cellens receptor (TCR), som interagerer med peptider præsenteret af MHC molekyler på overfladen af kroppens celler. For at sikre bred beskyttelse mod enhver given pathogen, har immunforsvaret udviklet sig til at kunne generere mange forskelligartede T celle receptorer, som hver især kan genkende en bred vifte af peptider. Kompleksiteten af dette system gør dog også studiet heraf mere problematisk. Ikke desto mindre er denne afhandling dedikeret til at undersøge T celle specificitet ved hjælp af populære eksperimentelle metoder samt at udvikle immunoinformatiske værktøjer og analyser for at fremme udbyttet af sådanne metoder.

En gængs metode til at måle T celle specificitet er peptid-MHC (pMHC) multimer farvning, hvilket tilvejebringer fordelingen af T celler, som responderer på et givent peptid. Denne metode blev anvendt til at kortlægge SARS-CoV-2 epitoper på tværs af kohorter af inficerede og raske individer, i det første projekt. Vi identificerede immunodominante epitoper i self raske individer, hvilket indikerer tilstedeværelsen af krydsreaktive T celler, som oprindeligt blev aktiveret mod et andet og måske lignende antigen.

Dog giver multimer farvning kun et overfladisk indblik i kompleksiteten af T celle receptor genkendelsen af en pMHC. For i sandhed at forstå mekanismerne bag T celle specificitet, benyttede vi os af enkelt-celle sekventering, som muliggør detektion af TCR$\alpha\beta$-kæder, et kognat pMHC, og celle hashing vha. antistoffer i det andet projekt. Fordi enkelt-celle data er forurenet pga. en række egenskaber ved teknikken, var hovedopgaven at udvikle en metode to effektivt at fjerne støj og bevare korrekte observationer af TCR-pMHC.

I det tredje og sidste projekt testede vi den førnævnte metode mod en nyligt publiceret metode for at forstå fordele og ulemper

ved begge. De to metoder viste sig at være meget forskellige ved hver især at prioritere modsatrettet i forhold til balancen mellem specificitet og sensitivitet når man måler TCR-pMHC parring.

Acknowledgements

First, I would like to express my deepest appreciation to my supervisor Professor Morten Nielsen; you have been a lighthouse throughout my time at DTU. Your patience and understanding has meant a lot to me and has served as a rock in the inevitable chaos of a 3-year dissertation. Equally important, is your ability to cut trough academic clutter with your honesty - it has helped to keep me focused and allowed me to trust you completely. I could also not have undertaken this journey without the help and support of co-supervisor Asssociate Professor Leon Eyrich Jessen. Besides being a great source of knowledge, you truly see people and pay attention to their needs. In a world of academia, those are rare skills and I want to applaud you for exercising them. You have made me feel seen in times of great pressure.

Second, I would also like to give special thanks to Postdoc Amalie Kai Bentzen for our great collaboration. Your innate curiosity and subsequent eagerness to push the limits of known and unknown methods within your own field has, in turn, helped me push my creativity in the application bioinformatic's methods. Thanks should also go to Professor Sine Reker Hadrup and her research group for always presenting me with new perspectives and for showing me the potential of interdisciplinary cooperation.

Finally, I would like to thank all my colleagues for making it not only educational to come into work, but also social and fun. Without you the coronavirus pandemic would have hit a lot harder. Even though, I am glad to be moving on to the next adventure, my feelings are mixed as I will be leaving you. I wish you all the best.

Abbreviations

ACC	Accuracy
APC	Antigen-Presenting Cell
ATRAP	Accurate T cell Receptor and Antigen Pairing
AUC	Area Under the rocCurve
BLOSUM	Blocks Substitution Matrix
CEF	Influenza virus
CD4	Cluster of Differentiation 4 (T helper cell)
CD8	Cluster of Differentiation 8 (cytotoxic T cell)
cDNA	Complementary DNA
CDR	Complementarity-Determining Region
CMV	Cytomegalovirus
CNN	Convolutional Neural Network
CV	Cross-Validation
DC	Dendritic cell
EBV	Epstein-Barr virus
ELISPOT	Enzyme-Linked Immunospot
ER	Endoplasmic Reticulum
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GEM	Gel-bead in Emulsion
HLA	Human Leukocyte Antigen
IEDB	Immune Epitope Database
MCC	Matthews Correlation Coefficient
MHC	Major Histocompatibility Complex
MHC I	Major Histocompatibility Complex class I
MHC II	Major Histocompatibility Complex class II
ML	Machine Learning

NGS	Next-Generation Sequencing
PCC	Pearson's Correlation Coefficient
PCR	Polymerase Chain Reaction
pMHC	peptide-MHC
PPV	Positive Predictive Value
RNN	Recurren Neural Network
ROC	Receiver Operating Characteristic
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
scRNA-seq	Single-Cell RNA sequencing
TAP	Transporter associated with antigen Processing
T cell	Thymus-dependent lymphocyte
TCR	T Cell Receptor
TN	True Negative
TNR	True Negative Rate (specificity)
TP	True Positive
TPR	True Positive Rate (sensitivity)
TSO	Template Switch Oligo
UMI	Unique Molecular Identifier
VAE	Variational Auto-Encoder
V(D)J	Variable, (Diversity), Joining genes of the TCR
VDJdb	VDJ Database

Contents

Introduction

Our immune system is a vast, intricate and sensitive interplay between different types of specialized cells and molecules of various functions. It keeps a delicate balance to rapidly eliminate pathogenic infections and malfunctioning cells without damaging healthy tissue [1]. It does so by recognizing both broad and highly specific patterns of foreignness. This task is delegated between the innate and the adaptive immune system. T cells, which are part of the adaptive arm, recognize their target via their T cell receptor (TCR). The receptor interacts with peptides presented on the Major Histocompatibility Complex (MHC) located on the cell surface [2, 3]. Upon interacting with a cognate peptide-MHC (pMHC) an immune response can be initiated. If the T cell belonged to the subclass of $CD8^+$ T cells, the eventual effect of the response would be cytotoxic killing [4] of cells presenting that specific peptide, coined an epitope. The adaptive feature resides in T cells being individually selected based on their specificity towards the given set of MHC presented epitopes. Thus, a T cell repertoire is determined by the history of pathogenic encounters and malignant cell transformations, resulting in both unique and highly diverse repertoires between individuals. Adding to the repertoire diversity is the feature of genetic polymorphism of the MHC, ensuring different peptides being presented from the same pathogen by different individuals. The great diversity makes it challenging to accurately describe T cell recognition, both in terms of determining the peptide specificity as well as understanding the structural features that constitutes the TCR interaction with its epitope. Gaining better understanding of the TCR-pMHC interaction and cohesion would aid monitoring of infectious disease progression and pave

the way for improved T cell based immunotherapy and rational design of vaccines [5].

Scope of the thesis

The identification of T cell epitopes is a complicated task due to high biological variability in T cell genetics, the TCR interaction mode, and the plethora of presented peptides. The identification is moreover challenged by the available assay techniques, which generally do not provide sufficient resolution to capture the exact TCR-pMHC binding requirements which are embedded within the amino acid sequences of the interacting partners. Recently, the single-cell platform has developed to facilitate high-throughput screening of T cell specificity, which holds the promise of detailed interrogation of T cells paired with cognate targets. However, an advent of next-generation technologies is typically followed by the next-generation of confounding factors.

The work of this thesis aims to interrogate T cell specificity and develop immunoinformatic methods for improved analysis and yield. The primary focus has been on extracting reliable TCR-pMHC pairing from single-cell data to enrich the field and enable the enticing goal of clinically applicable TCR-pMHC models.

Contents of the thesis chapters

The thesis is structed in the following way:

Chapter 1 covers the background of the biology determining T cell specificity.

Chapter 2 covers the background of various assay techniques of screening T cell specificity.

Chapter 3 covers the background of existing methods for modeling T cell specificity.

Chapter 4 introduces the first scientific paper. The main aim of the project was to map epitopes of the SARS-CoV-2 virus and to determine immunodominant epitopes.

Chapter 5 introduces the second scientific paper. The main aim of the project was to develop a framework to de-noise data from single-cell screening of T cell specificities.

Chapter 6 introduces the third scientific paper, which is an ongoing project. The main aim was to benchmark the de-noising framework presented in chapter 5.

Chapter 7 closes the thesis with an epilogue, discussing the key lessons from the presented work as well as the future prospects.

CHAPTER 1

T cell-mediated immunity

When a pathogen enters the human body a cascade of events will be effectuated to clear the infection. One of the key events leading to a T cell mediated immune response is the activation of dendritic cells (DCs), which are part of the innate immune response. Dendritic cells ingest the materials of their surroundings (known as phagocytosis, pinocytoses, or generally endocytosis), when triggered to do so [6–8]. Captured in a phagolysosome the pathogen is degraded into peptide fragments, a pathway known as antigen processing. Some of these peptides will be displayed by MHC molecules on the surface of the dendritic cell, known as antigen presentation (see sec. 1.1). The dendritic cells become activated if the exogenous ingestion is complemented with other signs of infection, such as cytokines or binding of its pathogen-recognition receptors [9]. The activated dendritic cell travels to the nearest lymphoid tissue to raise a T cell response by either priming naive T cells into mature effector cells (see sec. 1.4) or by reactivating memory T cells which continually recirculate the lymphoid organs [10, 11]. The T cells will each transiently contact the antigen-presenting dendritic cell via T cell receptor (TCR) probing of the pMHC and upon proper interaction, the T cell becomes activated via co-stimulatory signals [12–14]. The activated T cells will proliferate and migrate to the site of infection guided by signaling molecules exuding from the innate immune cells already at work

[15, 16]. Upon complete clearance, a fraction of the effector T cells will develop into memory T cells ready to fight off the pathogen in case of reinfection [17, 18].

1.1 Antigen processing and presentation

Antigen-presenting cells (APCs), such as dendritic cells, have different means of antigen processing. Generally, pathogens can be categorized as intracellular or extracelluar, indicating their preferred site of infection and replication. Typically bacteria, protozoa, fungi and worms are considered extracellular, while viruses are intracellular, as well as mutations leading to cell transformation. The site of infection hugely impacts the type of immune response needed: intracellular pathogens and malfunctioning cells must be eliminated via cytotoxic killing of the affected cell while extracellular pathogens can be targeted more directly with antibodies. The different responses are raised through different pathways of antigen processing: the cytosolic pathway, the endocytic pathway, the cross-presentation pathway, and the autophagy pathway.

The cytosolic pathway leads to MHC class I (MHC I) presentation of intracellular antigenic peptides as well as host proteome peptides and can be performed by any nucleated cell [2]. The pathway consist of TAP restricted transportation of proteasome degraded cytosolic proteins into the endoplasmic reticulum (ER) [20, 21], where the peptides are loaded onto MHC I molecules by a loading complex composed of several ER chaperons, including tapasin, calnexin, calreticulin [22, 23]. Once loaded, the MHC I molecules are rapidly transferred through the Golgi apparatus en route to the cell surface [24, 25].

The endocytic pathway leads to MHC class II (MHC II) presentation of extracellular antigenic peptides and is solely performed by APCs and B-cells [26, 27], because the pathway requires phagocytosis or macropinocytosis. In this pathway, the final ligand landscape is highly influenced by the cleavage motifs of the lysosomal proteases such as cathepsins [28, 29]. The MHC II molecules arrive in the late endosome where peptides compete with the MHC

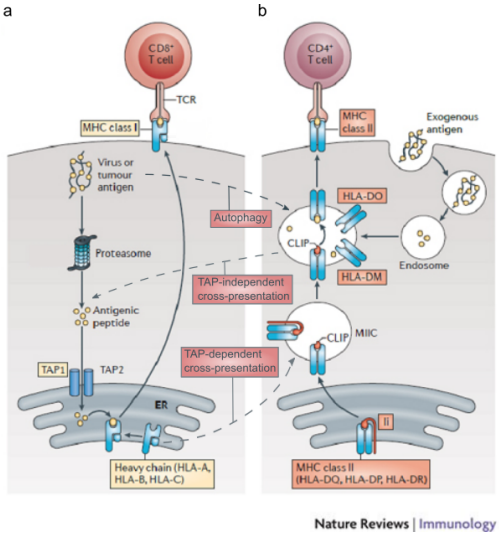


Figure 1.1: Antigen processing and presentation pathways, adapted from Kobayashi et al [19]. (a) The cytosolic pathway processing endogenous antigens to be presented on MHC I molecules. Intracellular proteins, such as virus or tumour antigens, as well as host proteins, are processed into peptides by the proteasome. The peptides are transported into the endoplasmic reticulum (ER), where they are loaded onto the MHC I molecules. On the cell surface, the pMHC I complexes present the cellular internal state to CD8 T cells. (b) The endocytic pathway processing exogenous antigens to be presented on MHC II molecules. Antigens from extracellular sources are engulfed via phagocytosis or phagopinocytosis and contained in an endosome where they are lysed into peptides. The late endosome describes the stage where MHC II molecules arrive and the peptides compete for binding against the class II-associated invariant chain peptide (CLIP). The peptide-loading process is regulated by HLA-DO and HLA-DM. On the cell surface, the pMHC II complexes presents the state of surrounding tissues to CD4+ T cells indicating the presence of bacteria, opsonized viruses or cancerous cells. Alternative pathways enable cross-presentation. The intracellular proteins may also be presented on MHC II molecules. Via autophagy, the cytosolic peptides may be captured in an endosome, indicated by the dashed line between the endogenous antigens in (a) and the late endosome in (b). Likewise, extracellular proteins may be presented on MHC I molecules. Either endocytic peptides are translocated to the cytosol via an unknown transporter mechanism, as indicated by the dashed line between panel (b) endosome to panel (a) cytosol, or MHC I molecules exist in endosomes outside of the ER.

II inhibitor, CLIP, for the binding groove [30–32]. Peptide-loaded MHC II molecules are finally shuttled to the cell surface of the APC.

A third antigen processing pathway leads to MHC I presentation of extracellular antigenic peptides, known as cross-presentation. Two mechanisms have been proposed: TAP-dependent and TAP-independent. The TAP-dependent pathway relies on a membrane transport pathway translocating proteins of endocytic compartments to the cytosol where they may enter into the cytosolic pathway [33–35]. The TAP-independent pathway requires the presence of MHC I molecules in the endocytic pathway [36–38]. The cross-presentation enables presentation of intracellular antigens from neighboring cells. Macropinocytosis allows cross-presentation of soluble antigens [33] while phagocytosis allows cross-presentation of bacteria [39, 40] or apoptotic cells due to viral infection [41, 42] or tumor state [43–45].

Finally, the autophagy pathway enables MHC II presentation of cytosolic proteins [46]. An autophagosome is the capture of cytoplasm in a vesicle, which is then fused with endocytic vesicles and lysosomes where the contents are degraded. The concept is a naturally occurring alternative to proteasome-mediated degradation, where both processes maintain a well-controlled balance between anabolism and catabolism [47]. However, the degraded proteins are kept in endosomes which are part of the endocytic pathway leading to MHC II loading.

The designated pathways are important for different functions, but at its core they all result in peptide presentation by either MHC I or II. The two classes of MHC molecules differ in their subunit composition, in ligand interaction, and in the type of T cell that might bind. MHC I ligands are short peptides of typically 8-11 residues [48] and are recognized by the $CD8^+$ cytotoxic T cells, while MHC II ligands are longer, 13-25 residues [49, 50], and pair with the $CD4^+$ T helper cells. The MHC ligands fit into a cleft between the two subunits of the MHC molecule, known as the binding groove. The mode of binding is characterized by few anchor residues defining the interaction of the ligand with the MHC binding groove [51, 52], allowing high variability in the

remaining ligand residues. This enables each MHC molecule to present a large amount of highly diverse peptides. The ligand residues not in the anchor positions generally face out of the MHC binding groove and are part of the interaction with the TCR [52].

1.1.1 MHC diversity

Beyond high diversity within ligands of an MHC molecule, broad immunological protection is also ensured by large population-wide variations in MHC molecules amongst individuals. Diversity of an individual's HLA repertoire is attained in at least three ways: polygeny, polymorphism, and allele co-dominance, and in addition, the binding cleft of MHC II is composed of two different subunits. In humans, MHC molecules are encoded by the human leukocyte antigen (HLA) locus. MHC I is encoded by HLA-A, -B, and -C, while MHC II is encoded by HLA-DR, -DQ, and -DP. The polymorphism amounts to more than 15,000 different molecules [53–55], which ensures heterozygosity in most individuals, resulting in individual expression of up to 6 different MHC I and 12 different MHC II molecules. The composition of alleles on each parental chromosome has the biological term HLA haplotype, however, in practice the haplotype often refers to the complete set of alleles accounting for the total HLA profile. A pathogen may evolve to evade any of those MHC molecules in a single individual, but across the entire human population the risk is greatly reduced [56]. Evidence of the importance of HLA heterozygosity for enabling presentation of different peptides comes from studies showing how most of the allelic variation resides in regions corresponding to the peptide-binding groove [57, 58].

1.2 T cell receptor interaction with peptide-MHC

Several X-ray crystallographic studies have provided detailed overview of the TCR:pMHC binding site as well as the TCR structure itself, see figure 1.2a. The TCR is a covalently linked heterodimeric protein, most often composed of an α- and a β-chain. A small subset of T cells express γ- and δ-chains [59], which also produce a functional TCR, however in this thesis TCRs will exclusively

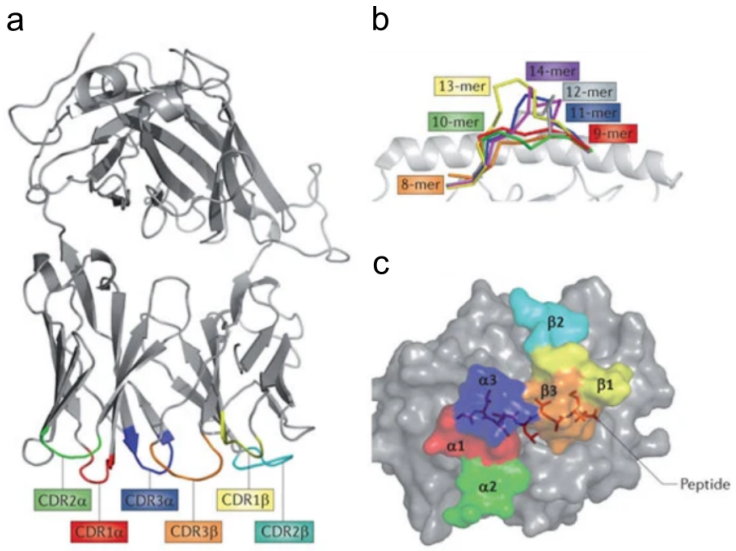


Figure 1.2: T cell interaction with peptide-MHC from X-ray crystallography adapted from Sewell et al. [60]. (a) A ribbon model of the αβ dimerized TCR highlighting the positions of the three CDR loops on each chain. (b) Illustration of how an MHC I molecule can accommodate peptides of different lengths. The longer the peptide the more it appear to 'bulge' out of the binding cleft. (c) Projection of CDR loops interacting with the pMHC onto the pMHC. In the given example the MHC is HLA-A*0201 (in grey) presenting the known epitope GLCTLVAML (stick model) from Epstein–Barr virus to the AS01 TCR96 receptor. The colored segments on the MHC indicate the contact points of each CDR loop. In this example, the CDR1 and CDR2 loops interact with the MHC molecule, while only the CDR3 loops interact with the peptide as well. This appears to be the general rule across TCR-pMHC interactions, albeit with some exceptions [61, 62].

refer to αβ-dimers. Both chains consist of a constant domain and a variable domain, see figure 1.3c. The TCR is anchored to the T cell via the transmembrane region in the constant domain of both chains. Variable domain contains three sets of loops, known as complementarity-determining regions CDR1, CDR2, and CDR3, closest to the pMHC.

The CDR1 and CDR2 loops are predominantly in contact with the MHC facilitating the initial binding [60–62]. Both the CDR3 loops are most often in close contact with the peptide, and dictate

the final outcome - binding or dissociation [63]. This dual specificity towards both peptide and MHC is a remnant of the TCR maturation process elaborated in section 1.3.

The flexibility of the loops allows different residues of the CDR3 to interact with different residues of the peptide [64], indicated in figure 1.2c, and allows the CDR3 to interact differently with an other peptide [65–67]. Since only a few residues at the interface may be essential for the specificity and binding strength, a TCR may recognize related, but different peptides [68, 69], although hypothesized to favor a certain epitope length [70]. It has been estimated that a single TCR can bind at least 10^6 different MHC-bound peptides, and perhaps even more [71]. However, the potential diversity of all possible 9-11mer peptide sequences combined exceeds $2 \cdot 10^{14}$ [60]. This diversity does not even account for post-translational modifications, which further expands the landscape of peptides. Fortunately, the number does not accurately reflect the peptidome of MHC ligands since antigen processing and presentation imposes strong restrictions on peptides, and thus serves as a strict initial selection step [20, 21]. However, assuming that 1% of all possible peptides are presented by an MHC molecule, the peptide landscape still exceeds the estimated human T cell repertoire of 10^{11} T cells, which are not all unique [72]. Thus, cross-reactivity is essential for sufficient protection, and even though 10^6 different targets of one T cell sounds extreme, the functional recognition translates into a frequency of 1 in 100,000 if the total pool is $10^1 1$ peptides [60]. This small example illustrates that observing cross-reactive binding should still be fairly rare, which is in good accord with an experimental attempt to directly measure this parameter [73]. Although rare, it ensures overlapping specificities within repertoires of both individuals and populations. Cross-reactivity reduces the risk of pathogens evading T cell mediated immunity by introducing critical mutations in epitopes and is thereby essential for overall survival [64, 74].

1.3 T cell development

Lymphopoiesis is the process of generating a diverse repertoire of T cell receptors which enables individuals to raise T cell mediated

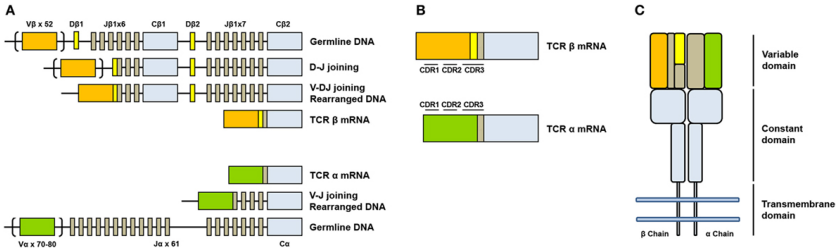


Figure 1.3: A schematic of V(D)J recombination of TCR αβ-chains adapted from Simone et al. [75]. (a) A map of the genomic organization and somatic recombination of TCR αβ-loci. Note how the D and J genes of the β-chain are split in two groups. This ensures the T cell two attempts of assembling a functional gene. (b) Map of the location of the three complementary-determining regions (CDRs) within each of the two chains. CDR1 and CDR2 are germline-encoded by the V-gene whereas the CDR3 spans the highly variable junctions consisting of N- and P-nucleotides. The CDR3s of α- and β-chain differ in that the CDR3 of the β-chain spans the gene junction of both D-J and V-DJ, which allows additional variability. (c) The final αβ-dimerized receptor anchored in the cell membrane. The CDR regions (not visible) are located at the top, in the variable domain of the receptor in short, flexible loops.

immune responses against the wide range of pathogens and cell transformations encountered during a lifetime. The maturation process starts in the thymus where progenitor cells migrate to and become thymus-dependent (T) lymphocytes, or T cells. In the thymus, the precursors commit to the T cell lineage by initiating TCR gene rearrangements which are illustrated in figure 1.3.

To begin with the thymocyte undergoes somatic recombination of the β-chain genes consisting of 52 variable (Vβ), 2 diversity (Dβ), 13 joining (Jβ), and 2 constant (C) regions, resulting in a VDJ-C transcript [1]. During the joining of D-J and the V-DJ segments, random nucleotides are added to form highly variable junctions. The joining process first involves addition of palindromic sequences (P-nucleotides) to each segment followed by addition of non-template-encoded (N-) nucleotides [76, 77]. Nucleotides can also be deleted during joining, which might erase traces of the introduced palindromes and even result in complete deletion of the D segment. The outcomes are highly variable joining regions of also variable length, which is what constitutes the CDR3 [78–80]. Typically, CDR3 sequences of both chains consist of 10-19

amino acids, most often of length 13 and 15, α and β respectively. The CDR1 and CDR2 are shorter sequences in the range of 4-10 residues [81–85]. Both CDR1 and CDR2 reside within the V-gene and thus only offer the variability of the number of gene segments: 70 Vα and 52 Vβ, respectively. However as the α- and β-chains are paired the number of combinations amounts to $5.8 \cdot 10^6$ [1].

Since the total number of added nucleotides is random, the reading frame is often disrupted leading to non-productive rearrangements and therefore non-viable thymocytes. In the surviving thymocytes the productive β-chain is paired with a surrogate pre-TCR α-chains (pTα) to form a pre-TCR. Ligand-independent dimerization of pre-TCRs induces both proliferation and expression of CD4 and CD8 co-receptors while the gene rearrangement of the β-chain is arrested [86]. The α-locus rearrangement does not begin before reaching this double-positive ($CD4^+CD8^+$) state, thereby ensuring that only a single β-chain is associated with the many different α-chains in the progeny cells. The somatic recombination of the α-locus follows the same model as the β-locus, only there is no D-gene, so the 70 Vα-genes may be joined with any of the 61 Jα-genes [78]. Non-productive VJ joining can be rescued by successive rearrangements which occur simultaneously on both chromosomes and continues until positive selection or cell death. Thus, in the strict sense, the α-locus is not subject to allelic exclusion as the β-locus, and therefore some mature T cells may express two productive α-chains. The dual receptor property does not inflict dual specificity, since only one of the productive TCRαβ-pairs is likely to have responded to the positive selection and recognize an MHC molecule [86–88].

Positive selection is the process of identifying TCRs that bind appropriately to an MHC of the haplotype of either class I or II, resulting in either $CD8^+$ or $CD4^+$ single-positive thymocytes. The MHC will present self-peptides processed by either the direct or the autophagy pathway [89]. The positive selection is followed by a negative selection, eliminating thymocytes that bind too strongly to self-peptides, thus promoting self-tolerance. The mature, naive, co-receptor specific T cell is now ready to enter the T cell repertoire and circulate the peripheral lymphoid tissue [1].

1.4 T cell activation

T cell responses are initiated when a mature naive T cell encounters an activated APC in a peripheral lymphoid organ, e.g. a lymph node [12–14]. Figure 1.4 illustrates the general features of $CD8^+$ T cell priming that will be described in the following. Dendritic cells express an array of co-stimulatory molecules that locks the naive T cell in place to properly test its specificity toward the range of presented pMHCs [90–93]. In the rare case of a TCR-pMHC match, the cell-cell adhesion is enforced and T cell differentiation is promoted via co-stimulatory signals and cytokine secretion [94]. The association persist throughout T cell proliferation, encouraging the progeny cells to adhere as well [13]. CD8 cytotoxic T cells often require excessive stimulation, perhaps due to their destructive effector actions and probably also to distinguish a foreign peptide presented on MHC I from self [13]. The additional stimulation is obtained from cross-primed $CD4^+$ effector T cells interacting with the same APC [95, 96]. Differentiation into effector T cell alters the expression of surface proteins, enabling the cell to locate and enter sites of inflammation and attach to host cells to sample their pMHC repertoire ($CD8^+$ T cells) [1]. The $CD8^+$ cytotoxic T cells are of main interest in this theses, and therefore, any unspecified phenotype can be assumed to refer to the $CD8^+$ subclass, for the remaining part.

Even though a T cell is primed by one pMHC, it can still recognize a palette of other peptides, as described in sec 1.3. However, it is important to distinguish between biochemical and immunological recognition between TCR and pMHC, since not every interaction will lead to a response [98]. As described, a T cell mediated immune response requires an activated APC that efficiently signals inflammation both for priming and for stimulating effector functions. Moreover, low affinity of pMHC complexes and a threshold requirement of TCR-pMHC interactions means that biochemical recognition can only be translated into immunological recognition if the source protein of the peptide is highly expressed [99]. Finally, several peripheral mechanism, such as immune suppression by regulatory T cells and cell signalling via surface expressed or secreted molecules, continually maintain self-tolerance of T cells

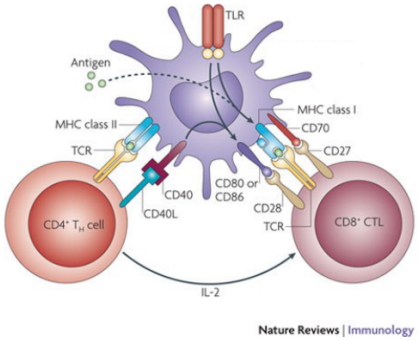


Figure 1.4: Mechanisms in T cell priming, adapted from Kurts et al. [97]. A dendritic cell has become activated and presents peptides on both MHC I and MHC II via direct and cross-presentation pathways (not shown). The MHC II peptide is being recognized by a CD4 T helper cell while the MHC I peptide is being recognized by a CD8 T cell. The priming of the CD8 T cell is enforced by sustained binding of TCR-pMHC as well as binding of signalling receptors like CD70/CD27 and CD80/CD28. Further, the activated CD4 T helper cell stimulates priming of the CD8 T cell with IL-2 cytokines. The dendritic cell is licensed to continue stimulating the CD8 T cell via the CD40/CD40L signalling receptors of the CD4 T helper cell. As a result, dendritic co-stimulatory molecules (CD70, CD80 and CD86) are upregulated, while inhibitory molecules, such as programmed cell death ligand (PDL1) are downregulated to promote priming and proliferation.

cross-binding to self-peptides [100–102]. Still, it remains unresolved exactly how most individuals are exempt from broad immune reactions that target self.

CHAPTER 2

Assaying T cell specificity

What is known today about T cell specificity is deduced from a range of experimental assays, each with their own advantages and disadvantages. The assays are designed for distinct purposes and therefore measure different aspects of the TCR-pMHC binding at different levels of resolution. Identifying epitopes, and especially immunodominant epitopes, can guide rational vaccine design and provide an initial step toward understanding the rules governing specificity. However, to truly decode specificity, it is essential to also capture the information embedded in the T cell receptor and link it to its cognate epitope. This chapter presents the current experimental methods of interrogating T cells for their specificity and elucidates the critical limitations that must be addressed to progress the field.

2.1 Common methods of screening T cells

The assays for probing the nature and quality of the TCR-pMHC interaction can roughly be grouped into four categories: cytokine production assays, proliferation assays, kinetic assays, and qualitative binding assays.

The two first methods rely on features reflecting immune responses of activated T cells, namely cytokine production and cell proliferation. Cytokine production can be used to measure both the

frequency of antigen-responding T cells and the type of cytokines they released, reflecting the effector function of the T cell [103, 104]. The most widely used technique is ELISPOT (enzyme-linked immunospot). T-cell proliferation can be assessed via tracers, such as the radioactive $[^3H]$-thymidine or the fluorescent dye BrdU (bromodeoxyuridine), which intercalates into replicating chromosomes and thereby quantifies the total amount of synthesized DNA in a bulk culture [105, 106]. To additionally estimate precursor frequency, flow cytometry can be utilized by staining cells with CFSE (carboxyfluorescein diacetate succinimidyl ester), which binds to amino groups of intracellular proteins [107, 108]. For each cell cycle, the amount of dye in a cell is halved, and at the end of the experiment, the distribution of dye reveals the initial frequency of the T cell precursor. Kinetic assays estimate binding affinity of the TCR-pMHC, which previously was thought to determine T cell signalling strength [109–114]. Still, binding affinity provides rich information regarding the specificity, especially when the affinity is relative to multiple different peptides, thus demonstrating how a single TCR binds a range of targets, known as TCR fingerprinting [69]. The technique dissects TCR specificity by sequentially substituting each position of the original TCR target with either a small amino acid (alanine or glycine) or each of the remaining 19 amino acids, for a more extensive analysis [68, 69, 115]. The binding affinity or functional response to each substitution is measured to produce a hierarchy of peptide preferences which essentially reveals cross-reactivity. This interrogation of a TCR by effectively masking out individual positions helps identifying “hot-spots” of interaction [116].

Common to the above listed methods is they depend on the TCRs of the responding cell culture being monoclonal. Monoclonal populations can be obtained from hybrids of T cells, cloned T cell lines, and limiting-dilution culture [117–119]. Although feasible, cell culturing is time consuming. Further the relevant assays only yield specificity indirectly and typically, only few antigens can be assayed at a time. Essentially, unveiling T cell specificity requires high-throughput qualitative binding assays.

The most detailed qualitative binding assay is via X-ray crystal-

lography. Not only does the method provide the sequence of both α- and β-chain, but it also provides the three-dimensional structure of the TCR-pMHC interaction. Despite great contribution to the field, the number of solved unique structures is still less than 100 due to the high cost and low throughput [120]. The key advancement for assaying specificity is the invention of synthetic conjugates of pMHC molecules which enable *in vitro* screening of T cells [121]. The conjugates consist of a multimer backbone carrying four to eight pMHC molecules and a label such as a fluorochrome, a metal tag, and/or a DNA barcode [121–125]. The label enables quantification of responding T cells, similar to cytokine production assays, and facilitates sorting of polyclonal cultures into antigen-specific T cell sub-populations ideal for sequencing of T cell receptors, as shown in figure 2.1.

Combinatorics with fluorescent labels or metal tags enables a library of 28-109 pMHCs [122, 124], whereas DNA barcoding allows high-throughput screening of >1000 pMHCs [125]. However, with such a large-scale specificity screening, the assay limitations are simply redefined from low-throughput to potentially capturing unspecific interactions [125]. The advantage is that the screening will resemble the *in vivo* scenario much closer, where each T cell will be presented to a plethora of pMHCs of which only few will qualify for binding. Hence, the scale of multimer staining has increased the scope of known interactions from a few model antigens to rare personal disease-associated antigens.

The assay is limited by only providing specificity distributions of the sampled repertoire. We do not know how big the complete repertoire is, if there is a dominant clone, nor whether certain clones exhibit cross-reactivity. This can be mitigated by sequencing the T cell receptor. Unfortunately, the genes encoding the α-chain and the β-chain reside on different chromosomes and can therefore not be sequenced in tandem using bulk sequencing. Since the CDR3β holds the greatest potential for variability due to the joining of both D-J and V-DJ it has been hypothesized that the β-chain conveys the most information of TCR specificity. Therefore, the majority of catalogued data only contains the β-chain. Pairing α- and β-chains has been approached from the assump-

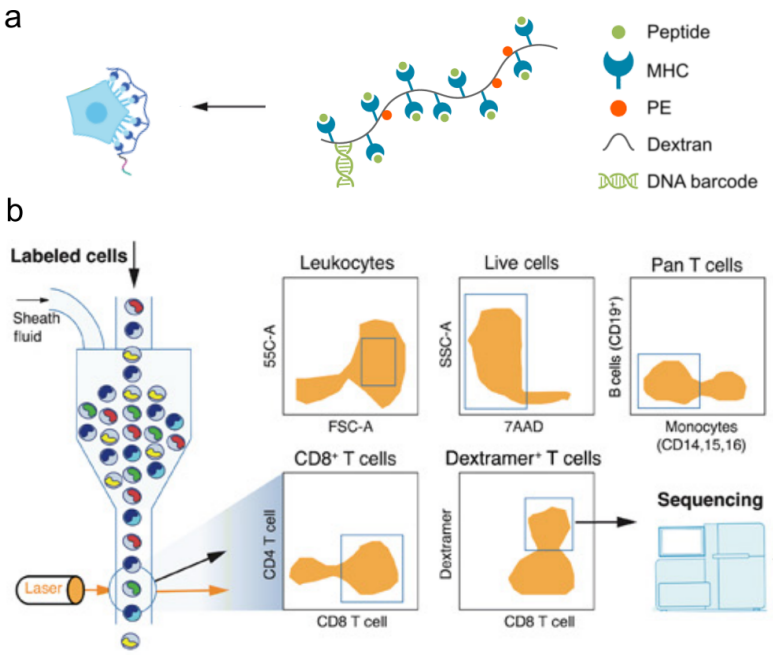


Figure 2.1: Cell sorting adapted from 10x Genomics, Immudex, and Zhang et al., 2021 [126, 127]. (a) Schematic overview of the DNA barcode-labeled pMHC multimers which are excellent for detection of antigen-specific T cells in complex cellular suspensions. The multimer scaffold is a dextran backbone carrying biotinylated pMHC molecules, a fluorochrome such as PE (phycoerythrin), and a DNA barcode. The DNA barcode serves as a tag for the corresponding pMHC. In *ex vivo* screening of TCR specificity the TCR-pMHC interaction is stabilized by multimerization of pMHC complexes on the scaffolds [121, 128–130]. The DNA barcodes are sequenced to determine the composition of antigen-responsive T cells in the sample. (b) Based on the fluorescent label, multimer bound T cells can be sorted using FACS (fluorescence activated cell sorting) [131] and visualized by flow cytometry [132]. When screening samples from buffy coat a set of gating-strategies are used to capture $CD8^+$ T cells only.

tion that naturally occurring pairs derived from the same cell will be observed with same relative frequency of transcriptomic reads [133, 134]. However, this indirect method of pairing chains becomes unreliable when sequencing highly diverse populations of T cells. Therefore, the next step for improved resolution is single-cell sequencing which captures the T cell receptor in direct association with a bound pMHC.

2.2 Single-cell sequencing elucidating TCR specificities

Single-cell sequencing is the frontier of next-generation sequencing (NGS) because it enables the study of complex and rare cell populations [135, 136], regulatory relationships between genes [137, 138], and trajectories of cell lineages during development [139–141]. Specifically, single-cell RNA sequencing (scRNA-seq) is paramount for truly understanding the link between genotype and phenotype. In the context of this thesis, scRNA-seq holds the promise of high-throughput screening of several thousand T cells against libraries of >1000 individual peptide-MHC complexes, thus connecting the specificity directly to the amino acid sequence of the TCR$\alpha\beta$. Beyond analysing specificity, the technology also enables screening of surface markers and sample hashing via a library of antibodies. In other words, any analyte with a conjugated DNA barcode can be included in an assay and if the analyte is associated with a cell the relation will show in the data.

Different platforms utilize different technologies to isolate individual cells either by droplet-based microfluidics [142, 143], microwells [144], or by *in situ* barcoding [135, 145]. Initially, the protocols did not provide 5' sequencing nor full-length coverage of transcripts, which excludes the CDR3 of the TCR transcripts. The commercial droplet-based platform, 10x Genomics, was the first to provide barcoded 5'end sequencing and hence became widely deployed especially for immune profiling.

2.2.1 The immune profiling platform by 10x Genomics

Figure 2.2 is a schematic overview of how the 10x Genomics droplet-based single-cell sequencing works. Cells in limiting dilution and gel beads in emulsion (GEMs) are pulsed into an aqueous stream and are captured in a droplet of emulsion when the stream is flushed with oil. By systematically pulsing cells with a lower rate than gel beads most droplets will contain only gel beads, while many droplets will contain a gel bead and a cell, and some will contain a gel bead and two or more cells. The doublet/multiplet rate is a trade-off for the capture rate, i.e. the fraction of cells introduced to the system which are recovered [142].

Sub-optimal capture is problematic when working with low frequency cell populations especially when searching for rare neoepitope specificities. Therefore, some cases might call for an increased capture rate at the expense of increased multiplet rate. Another challenge is that ambient transcripts in the cell suspension from apototic cells or from analyte barcodes may be randomly captured along a cell or even captured in an otherwise empty droplet, only containing the gel bead. The result is that unrelated sequences will, erroneously, be associated with the cell-of-origin.

In the droplets, the cells are lysed and the gel beads are dissolved, releasing reverse transcription (RT) reagents and poly(dT) primers (fig. 2.2b). The mRNA is reverse transcribed into full-length cDNA via priming of the poly(A)-tail and elongated with a template switch oligo (TSO). The TSO enables priming of another set of gel bead primers which contain a GEM barcode, a unique molecular identifier (UMI) [146], and a sequencing primer. The GEM barcode is identical for all the primers in the GEM, and unique for each gel bead, thus labeling all transcripts within the GEM such that the downstream sequencing reads can be traced back to the cell-of-origin. The UMI is a unique set of random nucleotides which identifies each transcript. The concept removes the count bias generated from nonuniform PCR amplification. Hence, downstream analyses can directly count the number of captured transcripts of a certain gene as a proxy for the expression profile. Any analyte DNA barcode must contain a capture-sequence which matches the TSO of the gel bead primer. The analyte barcodes

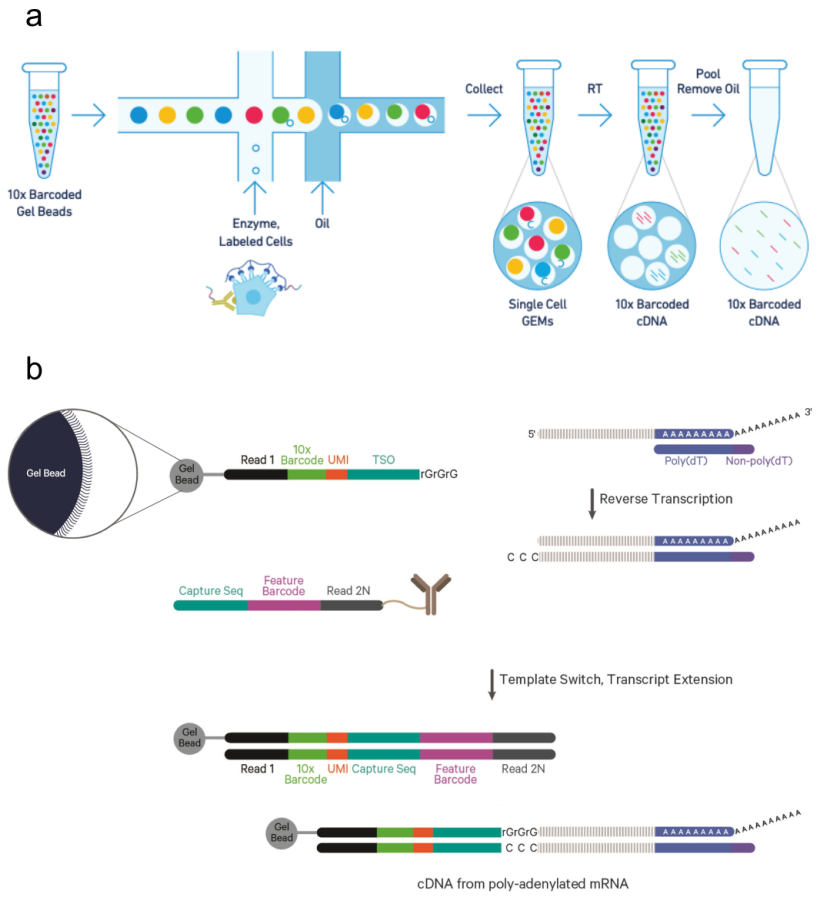


Figure 2.2: 10x Genomics Chromium Single Cell Immune Profiling Solution with Feature Barcode technology, adopted from 10x Genomics [126]. (a) T cells from donors are labeled with pMHC multimers and may also be labeled with DNA-barcoded antibodies for cell hashing or surface marker analysis. The $CD8^+$ T cells are sorted and introduced to the aqueous stream along with gel-beads in emulsion (GEMs). (b) During sequencing the cellular transcripts and DNA barcodes are captured by GEM primers such that all downstream reads are prefixed with a GEM barcode and a unique molecular identifier (UMI).

will then also be preceded by the GEM barcode to trace the cell-of-origin. In the next phase GEMs are broken and the contents are pooled for bulk DNA amplification and then separated into cDNA from poly(A) mRNA and DNA from barcoded analyte by size. To ensure that TCR transcripts are sufficiently sequenced they are amplified in an additional step with primers specific to the TCR constant regions. Next the amplified cDNA is enzymatically fragmented and sequencing primers as well as sample indexes are added for batch multiplexing. The final sequence output can be processed with standard NGS techniques and then converted into count matrices of genes or other features per GEM. Software for mapping and annotating GEMs is provided by 10x Genomics.

2.2.2 Challenges in single-cell data pertaining to T cell specificity

To ensure that scRNA-seq data is fully exploited and interpreted correctly, it is important to apply appropriate computational and statistical approaches. Some methods originating from bulk RNA sequencing can be reused, however, scRNA-seq poses several novel confounding factors that require adapted analytical strategies. Computational methods must account for biological confounders such as cell cycle variations, apoptosis, and stochastic gene expression, as well as technical artifacts such as cell multiplets, cross-contamination, dropout, and batch effects when compiling different experiments.

Both artificial dropout and stochastic gene expression result in zero-inflated gene counts. Genes are expressed transiently, even when accounting for cell cycle variations [147, 148]. Hence, transcripts for all genes are not present at all times which falsely render some genes with a count of zero. An artificial dropout can refer to multiple steps during scRNA-seq which cause genes to be underrepresented in the final output data. Causes may be inefficient initial priming of the poly(A) mRNA, PCR amplification bias, or insufficient sequencing depth [149–151]. Generally, counts of genes with low expression magnitude are more likely to be zero-inflated [152]. Thus, dropout of an otherwise highly expressed gene is more indicative of true expression differences than of stochastic variabil-

ity, which had been the conclusion if the gene had been generally lowly expressed.

Random capture of ambient mRNA or DNA barcode can affect all GEMs as illustrated in figure 2.3c+d. In case of random capture in a cell-void GEM, the GEM will naively appear to contain a cell with sparse gene expression. In case of random capture in a GEM that actually contains a cell, the GEM will be observed with a high expression profile. In both cases the expression profile would be inconsistent with the expected expression profile [153]. For example, ambient mRNA contamination is likely when highly expressed genes are observed at low levels in a few cells within a homogeneous population. Similarly, contamination of analyte DNA barcodes may also be an issue when concentrations are high. Advanced methods exist to eliminate contaminated GEMs by modelling the expected expression profiles and filtering out GEMs that deviate. A more simple approach is to filter away GEMs where expression across all genes is generally too low or too high, thus excluding abnormal expression profiles. It is a bit more problematic to rule out analyte contamination because the same extend of background observations does not exist.

A similar confounder is multiplet capture, i.e. capture of multiple cells in one GEM as in 2.3b. Depending on the protocol, multiplets may constitute up to 40% of GEMs [154]. Again, the GEM will appear with abnormally high expression values because transcripts from two cells are captured in the GEM. As before, such GEMs may simply be removed by filtering on a maximum threshold for expression values. However, it is probable that the captured cells are completely orthogonal in their expression profiles which would not result in extreme outliers to filter on. Fortunately, more advanced methods have been developed to discern a single cell from a multiplet. DoubletFinder [155] and Scrublet [156] filters data based on a similarity score of artificially constructed doublets to the dimensionality-reduced data. Demuxlet [157], scds [158], and scSplit [159] identifies doublets based on expression of genes that are likely to not occur simultaneously, e.g. transcripts associated with mutually exclusive sets of single-nucleotide polymorphisms (SNPs). DecontX [153] specifically identifies contaminated cells

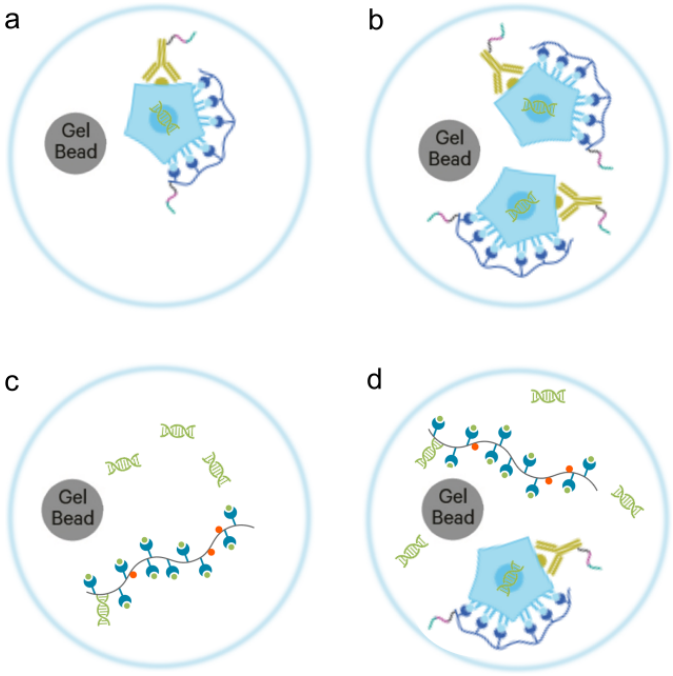


Figure 2.3: Challenges in single-cell data, adopted from Immudex and 10x Genomics [126]. (a) The expected capture of a single cell in a droplet. (b) The calculated risk of capturing cell multiplets. (c) The event of capturing ambient DNA and/or pMHC multimer in an otherwise empty droplet. (d) The event of capturing ambient DNA and/or pMHC multimer together with a cell. In examples (c) and (d), both native mRNA from the cell and contaminating ambient mRNA/DNA will be barcoded and counted within a droplet.

using a Bayesian mixture model from the assumption that transcripts from the native cell are distributed differently than the contaminating transcripts. Several experimental techniques have also been developed to assist detection of doublets. Example techniques include species mixture [142, 143], lipid-tagged indexing [160], and cell hashing [161, 162]. The two latter introducing benefits such as hashing by origin-of-donor which is highly desirable in patient studies.

The methods developed to filter scRNA-seq data each have their own advantages and disadvantages, however, common to them all is that this field of research lacks a ground truth dataset. A

comprehensive benchmarking of these methods on high quality of data would be desirable. The major concerns when screening for TCR specificity are nonspecific binding, incomplete TCR annotation, and T cell multiplets. These issues arise from the above stated caveats of single-cell sequencing. Incomplete TCR annotation and T cell multiplets may be mitigated applying some of the above mentioned methods, however, unspecific binding is a relatively unresolved issue. Unspecific binding may completely dilute the signal from actual interactions and impede the assay and applicability of the data.

2.3 Currently available data

In order to collect the data produced by individual groups, several databases have been established. Each aim to collect and curate published TCR specificities to propel research forward for clinical use. The majority, with some overlap, of the publicly available TCR specificity data resides in the Immune Epitope Database (IEDB) [163], VDJdb [164], the manually curated catalogue of pathology associated TCRs (McPAS-TCR) [165], TCR3d [166], and ImmuneCODE [167]. The latter, ImmuneCODE, was created in the wake of the covid-19 pandemic to catalogue disease relevant specificities and repertoires [167].

The most commonly described antigens are derived from viral infections of cytomegalovirus (CMV), Epstein-Barr virus (EBV), and influenza viruses (CEF). Virus-responsive T cells are easily detected because they are often present at high frequencies in infected individuals. Moreover, virus-derived peptides are likely to be dissimilar to self-peptides, which may yield high affinity TCRs which are good candidates for multimer staining [168, 169]. Unfortunately, the opposite can be said about neoantigens. They are likely to be highly similar to self, towards which, T cells have been selected to have low affinity, resulting in only transient binding [170]. Further, cancer cells are typically heterogeneous across patients and even within tumors, thus minimizing the chance of shared neoepitopes that would elicit a broad T cell response, also observable in a cohort. Instead, T cells specific for a given neoepitope are rare and difficult to detect. To top it off, T cells recogniz-

ing neoantigens may be difficult to culture and stimulate ex vivo due to the suppression mechanisms of immunoediting [171, 172].

Before the covid-19 pandemic the majority of bulk-sequenced data was centered on a few hundred peptides, some of the most abundant being YVL, GLC, NLV, GIL, extracted from table 2.1. These four peptides are also the most thoroughly investigated when requiring both TCR chains. The top 5 most abundant peptide specificities are all restricted by the same HLA allele, HLA-A*02:01, which limits the potential for understanding how TCRs might differentiate between HLA molecules. During the covid-19 pandemic the interest for pathogen specific TCRs spiked which of course expanded the databases drastically, however, these investigations have primarily resulted in sequencing of the TCRβ-chain only. Recently, several TCR-pMHC interaction prediction models have been published, unanimously stating that including both chains of the TCR improved performance [127, 173–178]. Montemurro et. al further demonstrated how peptide-wise performance was particularly dependent on the number of unique TCRs [173]. It was estimated that a minimum of 150 specificity observations were a requirement for robust predictions on independent test datasets. From the table below, it is evident that only 3 peptides meet that requirement. In order to fully understand TCR specificity, we must be able to study it from many more different types of interactions, consisting of more peptides across more HLA alleles. The need for both TCR chains as well as high amounts of different TCRs per peptide specificity, highlight the importance of assays that enable screening of large peptide libraries against many thousand T cells from which both TCR chains can be obtained. In 2019, the commercial single-cell RNA sequencing platform 10x Genomics published such a dataset containing 50 library peptides [126, 179]. The assay yielded 55,221 uniquely paired TCR sequences. The drawback of this rich source of information is again the many unspecific interactions which are not trivial to discern from true binding events. Investigating solutions to this challenge is one of the focus points of this thesis.

Organism	Peptide	Length	MHC Allele	# CDR3b	# CDR3ab
Human herpesvirus 4 (EBV)	RAKFKQLL	8	HLA-B*08:01	187	1
Human herpesvirus 5 (CMV)	VTEHDTLLY	9	HLA-A*01:01	274	1
Human herpesvirus 5 (CMV)	TPRVTGGGAM	10	HLA-B*07:02	2292	1
Homo sapiens	EAAGIGILTV	10	HLA-A*02:01	214	16
Hepatitis C virus	CINGVCWTV	9	HLA-A*02:01	114	28
Homo sapiens	ELAGIGILTV	10	HLA-A*02:01	558	79
Human herpesvirus 4 (EBV)	YVLDHLIVV	9	HLA-A*02:01	8488	115
Human herpesvirus 4 (EBV)	GLCTLVAML	9	HLA-A*02:01	7032	128
Human herpesvirus 5 (CMV)	NLVPMVATV	9	HLA-A*02:01	4886	210
Yellow fever virus	LLWNGPMAV	9	HLA-A*02:01	2173	410
Influenza A virus (CEF)	GILGFVFTL	9	HLA-A*02:01	4539	438
Human herpesvirus 5 (CMV)	NEGVKAAW	8	HLA-B*44:03	117	
Hepatitis C virus	ATDALMTGY	9	HLA-A*01:01	131	
Hepatitis B virus	KTAYSHLSTSK	11	HLA-A*11:01	476	
Hepatitis B virus	STLPETAVVRR	11	HLA-A*11:01	925	
Hepatitis B virus	LVVDFSQFSR	10	HLA-A*11:01	1875	
Influenza A virus (CEF)	LPRRSGAAGA	10	HLA-B*07:02	2142	

Table 2.1: Counts of unique CDR3β and CDR3αβ per epitope from IEDB [163]. The table contains the 17 epitopes catalogued with at least 100 unique CDR3β sequences. The first column describes the source organism of the epitope. The epitope is given in column 2, and the length of the epitope is registered in column 3. The restricting HLA allele of the pMHC complex is found in column 4. In column 5 and 6 are listed the counts of unique CDR3β and CDR3αβ chains, respectively.

CHAPTER 3

Immunoinformatic approaches for characterizing T cell specificity

The field of immunoinformatics has long been dominated by predictions of MHC I and II ligands which has now progressed into a plateau at very high performance [180–183]. Given the growing availability of T cell specificity data, attention has shifted towards modeling TCR-pMHC interaction [127, 173–178, 184–192], which holds great clinical potential. A substantial amount of data is required in order to produce robust models that can generalize beyond the scope of the currently available observations. Hence, the publication of the large 10x Genomics specificity data has further paved the way for many new models in recent years [127, 173, 174, 177]. Many different types of models have been proposed, suggesting that the quest of identifying the best suited method is still ongoing. In order to evaluate and compare these models, the field makes use of several performance metrics, which can also be applied when assessing the quality of a data processing or other types of analyses. In this chapter, the most common performance metrics will be introduced, followed by presentation of some of the most prominent models of TCR-pMHC specificity.

3.1 Performance metrics

A vital element in developing data-driven models is fair and unbiased evaluation of performance. Many different metrics exist, however, not all metrics are suited for any problem. Most metrics rely on quantification of observations defined as true positives (TP), true negatives (TN), false positive (FP) and false negatives (FN). This definition is straight forward when the evaluation variable is binary. In the situation where the evaluation variable is continuous, the scale can be divided into "true" and "false" bins by setting a discrimination threshold or by testing the impact of continuously increasing the threshold. The positive values binned above the discrimination threshold will be quantified as true positives, while the positive values binned below the threshold will be quantified as false negative. Vice versa for the negative values which are quantified as truly negative below the threshold and falsely positive above the threshold. The further apart the distributions of the two categories are, the easier it is to set a good threshold and the fewer false calls made. The decision of where to place the threshold may be affected by what types of mistakes we can and cannot accept. The trade-off is specificity versus sensitivity. Sensitivity refers to the probability of correctly classifying a positive observation as positive, also known as the true positive rate (TPR):

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

Specificity is the complementary metric which refers to the probability of correctly classifying a negative observation as negative, known as the true negative rate (TNR):

$$TNR = \frac{TN}{TN + FP} \tag{3.2}$$

Choosing high sensitivity over specificity allows the detection of most true positive values at the expense of including many false positive values. This is optimal if it is important that the method does not miss any true positive instances. An example could be a

medical screening of a patient where the medical staff would rather detect a benign tumor than miss a malignant cancer. Choosing high specificity over sensitivity allows the method to be highly selective at the expense of excluding many positive instances. This prioritization is relevant when including a false negative is too costly. In the context of screening TCR specificities, high sensitivity would enable detection of rare and low affinity interactions while falsely including transient unspecific pMHC probing as a binder. Alternatively, high specificity would aid elimination of unspecific binding and instead only present credible binding events.

3.1.1 The receiver operating characteristic and area under the curve

A popular performance metric that incorporates the trade-off between specificity and sensitivity is the receiver operating characteristic (ROC). A ROC curve is created by plotting TPR against $1 - TNR$, also known as the false positive rate (FPR) or fall-out, as the discrimination threshold changes from $\infty : -\infty$. The area under the ROC curve (rocAUC or AUC) is a single metric that summarizes the balance between sensitivity and 1-specificity. An AUC value of 1 indicates complete separation of the positive and negative instances, while a value of 0.5 or below shows no separation. The AUC metric can be modified to highlight methods that favor specificity over sensitivity by only taking the integral of the ROC curve up until FPR of 0.1, coined AUC 0.1.

3.1.2 Matthews correlation coefficient

When the data that must be evaluated is strictly binary or if the discrimination threshold is pre-determined a ROC curve is undue. Conveniently, Matthews' Correlation Coefficient (MCC) measures the association between two binary variables [193], which is a special case of Pearson's Correlation Coefficient (PCC) [194]. The correlation is given by:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3}$$

The MCC is robust to class imbalance as it takes all classes (TP, FP, TN, FN) into consideration. A perfect correlation results in a value of 1, while 0 reflects random association, and -1 indicates an inverse correlation.

3.1.3 Accuracy

Another commonly used metric is accuracy (ACC). In the situation where the evaluation variable is continuous, accuracy is a measure of observational error. In a binary classification problem accuracy is the ratio of true labels (TP & TN), given by:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Accuracy is widely popular as it is applicable for evaluation of both regression and classification models. However, the metric is unable to account for sensitivity and specificity and it is easily affected by class imbalance, falsely favoring the largest class. Despite its caveats accuracy is sometimes the only option because real-life cases do not always hold a clear definition of what is a false positive and what is a false negative, and perhaps true negatives do not even exist in the problem at hand. An example is investigating specificity of many clones of a T cell. Ideally they would all bind the same pMHC (TP), but as described earlier, TCRs are expected to display some promiscuity which would result in multiple observed binding events (TPs). Similarly, screening via single-cell platforms are prone to artifacts which would introduce false binding events (FPs). The remaining pMHCs from the screening library that did not elicit a binding may all be truly negative (TN), however, some might also just be missed observations (FN) due to dropout for example. Since the labels are not clearly defined all metrics that rely on FPs and FNs are unfit for use, whereas accuracy is unaffected.

3.2 Similarity of TCRs

Based on the assumption that T cells responding to the same epitope share more similar TCRs than compared to T cells of other specificities, many have attempted to cluster TCRs to investigate the patterns of specificity [178, 184, 190, 191]. Clustering requires a numerical representation of the sequences or a distance/similarity metric between all pairs of TCRs. There is no universal consensus of which metric is preferred, since we have yet to learn the rules of TCR similarity, both in regards to sequence as well as specificity. Moreover, simple models can be implemented to make predictions of TCR specificity based on the distance metrics.

Figure 3.1 represents a sketch of how similarly color-coded TCRs are clustered together and appear to bind the same epitope.

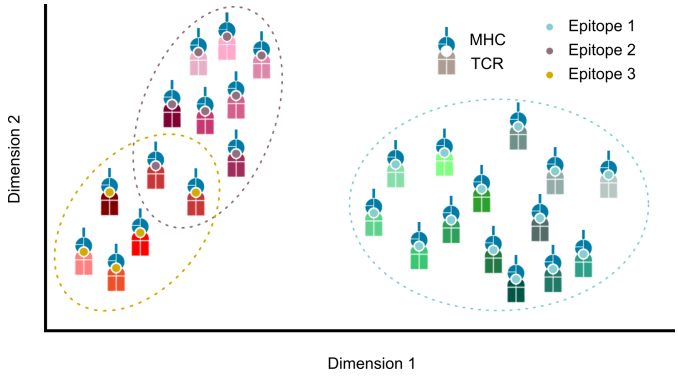


Figure 3.1: TCR similarity within and across epitopes, adapted from Immudex. The figure illustrates the concept that TCRs (in some dimensionality-reduced space) cluster based on their α- and/or β-chains. Given that a TCR often will be more similar to other TCRs of same specificity than to TCRs of different specificities, the clusters will each correspond to an epitope.

3.2.1 Hamming and Levenshtein distances

Naively, one may compute distance between sequences using the Hamming distance. This metric measures the edit distance as the number of substitutions required to transform one string into the other. However, this metric requires that the compared strings are of equal length. The Levenshtein distance metric accounts for

this shortcoming by measuring the number of required insertions and/or deletions in addition to substitutions [195].

This similarity score has been used for unsupervised clustering of TCRs from repertoire sequencing [165, 196, 197]. Despite enabling sequence comparison of unequal length, Levenshtein is still biased by length. If the distance between a sequence pair is normalized by the longest sequence, the normalization will render two short sequences with very few mismatches as similar as two long sequences with many more mismatches.

3.2.2 BLOSUM scoring

Counting a uniform edit penalty may be sub-optimal, as substitutions of different amino acids confer various types of alternation to a sequence. A distance measure that accounts for that can be based on BLOSUM. BLOSUM (blocks substitution matrix) scores alignments based on relative frequencies of amino acids and their substitution probabilities within highly conserved sequences [198]. The score reflects the evolutionary pressure to preserve the folding and, thereby, function of proteins, and hence, serves as a similarity metric.

This scheme is deployed by TCRdist which is defined as a similarity-weighted mismatch distance between αβ-pairs of the CDR2, CDR2.5, and CDR3 sequences [178]. The distance is calculated as an alignment score, however with inverted BLOSUM62 scores. In the TCRdist publication, the scores of all-against-all TCRs were utilized for a k-nearest neighbor (kNN) prediction based on the nearest 10 percent of the repertoire with a weight that linearly decreases from nearest to farthest neighbors.

3.2.3 Physico-chemical profiling

Since CDR3 sequences are not a product of evolution, but arise from random insertion and deletion of nucleotides (see section 1.3), one can argue that BLOSUM scoring is not optimal. Instead, amino acids can be represented via their physico-chemical properties such as basicity, helicity and hydrophobicity to construct a sequence profile.

Physico-chemical profiling has been used both for unsupervised clustering of $CD4^+$ TCRs from repertoire sequencing [199], for sequence encoding [188, 189], and for a benchmark of distance scoring schemes [200]. In the benchmark, the distance between two sequences is calculated by the sum of position-weighted Euclidean distances for each of the normalized physico-chemical profiles. Weights are highest for central residues and linearly decreases toward the edges of the sequence. In this case, insertions/deletions were accounted for by truncating the longest sequence on either end, which skews the distance to favor sequences of comparable lengths.

3.2.4 K-mer scoring and kernel similarity

An alternative to sequence alignment is comparing pairs of sequences by k-mer subsets. The distance measure in GLIPH calculates the motif frequency of k-mers ($k = 2, 3, 4, 5$) relative to a distribution of expected frequencies [184]. This approach can be combined with BLOSUM scoring to account for residue similarity. Such a method was presented by Shen et. al [201], and is the underlying method in both MAIT Match [190] and TCRMatch [191]. The algorithm takes two sequences as input: s_1 and s_2. Both sequences are split into sets of k-mers, where k takes on values from 1 up to the length of the shortest of the sequences, s_1 and s_2. For each value of k, all possible combinations of k-mer pairs between s_1 and s_2 are aligned to compute a score based on an amino acid substitution matrix such as BLOSUM62. For each k-mer, the scores for each of the aligned amino acids are multiplied. All k-mer products are then summed and normalized to yield a value between 0 and 1, where 1 is a perfect match:

$$\hat{K}(s_1, s_2) = \frac{K(s_1, s_2)}{\sqrt{K(s_1, s_1)K(s_2, s_2)}} \tag{3.5}$$

The kernel method utilized in TCRmatch [191] was evaluated by precision and recall, defining a true positive when $\hat{K}$ is above a set threshold and the epitopes of s_1 and s_2 are identical. A mismatch between the epitopes would be considered a false positive.

3.2.5 Predictions

These similarity or distance scoring schemes of sequences have in few cases been used to predict binding of novel TCRs. TCRmatch and TCRdist relied on adapted implementations of k-nearest neighbors [178, 191], while other distance schemes have been benchmarked using the DBSCAN clustering algorithm [200]. The few publications conclude that none of the tested distance schemes is superior to the others and that the models in general are limited by the scope of the available data [191, 200]. Further, clustering demonstrated the complexity of TCR specificity, by accentuating how TCRs sharing specificity exist in a wide spectrum and may internally be as dissimilar as TCRs targeting widely different epitopes. Thus, the hypothetical clustering of figure 3.1 is extremely idealized. This observations clearly exemplifies the intricate task of modeling TCR specificity. Finally, these types of models cannot generalize to novel epitopes beyond the training set, which reduces their prospects and applicability. The ability to generalize arises from learning the underlying mechanisms that define a match between a TCR and a pMHC.

3.3 Modeling T cell specificity

Investigations of TCR similarity suggests that TCR specificity is a complex problem, which might require complex solutions. Selected machine learning (ML) frameworks facilitate detection of hidden features of paired TCR-pMHCs. Learning the underlying mechanisms of binding provides the ability of models to generalize beyond the experimentally obtained measurements. Thus in theory, models trained on specificities towards one set of peptides may still be able to predict specificity of an orthogonal set of peptides. Pan-specific models elevated the performance of peptide-MHC prediction models [202, 203], and might do the same for TCR-pMHC models.

Machine learning refers to an array of models, however, in this context, the scope is limited to methods that capture non-linearity and higher order correlations of data, including convolutional neural networks (CNNs) [127, 173, 174, 177, 187, 192], variational

autoencoders (VAEs) [176, 177, 204], and recurrent neural networks (RNNs) [174, 176, 204]. These architecture were utilized for TCR-pMHC modeling in a range of recent publications: TcellMatch [174], ERGO [176], ERGO-II [204], NetTCR-2.0 [173], ImRex [187], DeepTCR [177], TCRAI [127], and TITAN [192].

3.3.1 The neural network

Conceptually, the input to neural networks is propagated via the model parameters, or weights, through layers of hidden neurons until the output neuron(s), as illustrated in figure 3.2. Neural networks are constrained to only accept inputs of a pre-specified shape, hence sequences of varying length must be padded to fill out the required dimensions. Each layer may have its own set of rules for how to propagate the signal forward. Here a range of activation functions, such as rectified linear unit (ReLU), hyperbolic tangent (tanh), and the sigmoid function can be implemented to enforce non-linear modelling [205]. The weights are tuned over multiple rounds (epochs) of training to minimize the prediction error in a process known as back-propagation. Again different error functions and optimization algorithms may be used depending on the type of data and the type of network. The CNN, VAE, and RNN are all variations of the simple network architecture, specially designed to capture different kinds of hidden features.

3.3.2 Neural network architectures for specificity modeling

A common architecture for modelling sequence data is the CNN [206], which was originally designed for image classification [207]. In a CNN the neurons are connected by sliding multiple unique kernels across the input, known as convolutions. Each kernel has its own set of weights that are tuned to detect a specific feature. By the concert of multiple kernels a CNN is able to capture the spacial and temporal dependencies in an image (or a peptide encoding). The RNN is an umbrella term of including the popular long-short term memory (LSTM) architecture. These networks were built for inputs of sequential character, like time-series, sentences or proteins, which are tokenized and fed to the network consecutively. The LSTM incorporates the hidden state of the

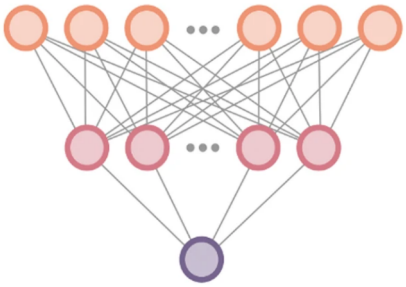


Figure 3.2: Artificial neural network adapted from Montemurro et al., 2021 [173]. The illustration represents a small feed forward neural network. The nodes represent neurons and the edges represent weights. The top layer of orange neurons is denoted the input layer. The second layer of red neurons is a hidden layer, while the last purple node is the output neuron. Each neuron is connected to every neuron of the previous layer by trainable weights, which is how information is propagated through the network.

previous token with the current token continuously until the final token. This process guides the network to capture sequence patterns, which may reflect interaction between residues of the CDR3 loop. Since protein sequences have no reading frame, bi-directional LSTMs have been developed to capture interactions propagated from both ends. The VAE is designed to capture the most important features of data in a latent space, which is a core layer in the network. The architecture consist of an encoder and a decoder. The data is compressed by the encoder into a latent space followed by a decompression by the decoder to reconstruct the input. Information may be lost during the encoding, however this also serves to de-noise data. Typically, the encoder can be used to embed sequences in a dimensionality reduced space and the latent state may be used for unsupervised clustering.

3.3.3 Encoding and embedding

In ML the data must be encoded to comply with a strict format chosen for the model. The type of encoding can impact the model and should be considered with care [208, 209]. Categorical data is often one-hot encoded, assuming that all categories are equally similar. The models including V(D)Jαβ-genes and HLA alleles have utilized one-hot encoding for these features [174, 177, 204].

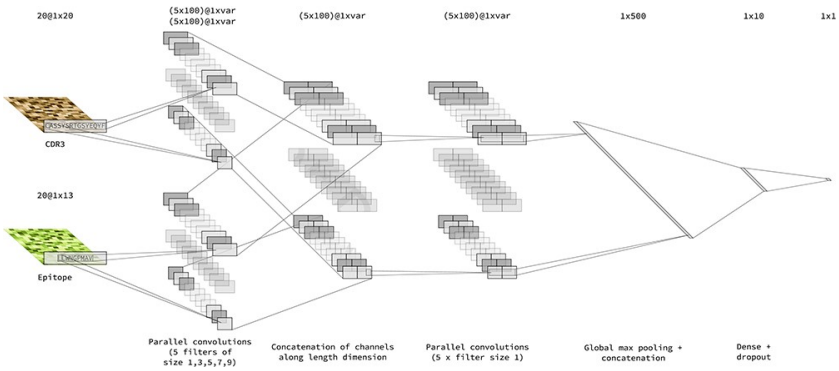


Figure 3.3: Convolutional neural network adapted from Moris et al., 2021 [187]. An image is essentially a 3 dimensional tensor of colored pixels. This concept is mimicked in immunoinformatics by various encoding schemes for sequences into numerical representation. Kernels are used to detect patterns by sliding across the image. Different types of pooling can downsize the network. The last couple of layers consist of fully connected neurons by flattening the CNN tensor. The model in this example solves a binary classification problem where the output may be interpreted as the probability of binding between the given CDR3 and epitope.

Typically CDR3 and epitope sequences are encoded using the evolution-driven BLOSUM score [173, 174, 185, 188, 192], while only ImRex attempts physico-chemical encoding as an interaction map between peptide and CDR3β[187], illustrated in figure 3.4. This forces the model to focus on the features of interaction instead of identifying internal representations of the peptide and CDR3, individually. However ideally, even when peptide and CDR3 sequences are fed separately, a complex network should eventually capture the interaction as a hidden feature.

Yet another alternative of encoding is defining each amino acid as a category and feed the vectorized sequence to an embedding layer as the first layer of the network [127, 176, 177, 192, 204]. The embedding layer consist of trainable weights and therefore learns the optimal projection into an X dimensional space defined by the user.

Other methods, which are actually individual networks themselves, can also be trained to embed a sequence [210, 211]. An example is the VAE modeling of CDR3 sequences, of which the pre-trained

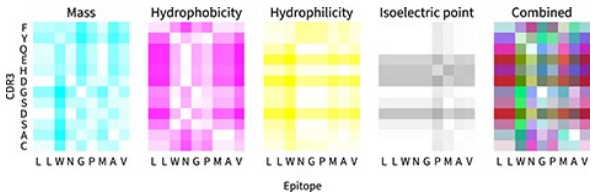


Figure 3.4: The ImRex interaction map adapted from Moris et al., 2021 [187]. The interaction map consist of absolute differences in physico-chemical properties between a CDR3βand an epitope. The rightmost image is a combined representation of the individual physico-chemical interaction maps, resulting in a three-dimensional tensor, here represented with CMYK colour encoding.

encoder may be used as an input embedding [176, 177, 204]. In the non-bioinformatic field of machine learning transformers are praised as the new standard for natural language processing [212–215], and within immunoinformatics few have attempted peptide encoding for MHC ligand predictions [216, 217]. Within TCR-pMHC modeling, no exhaustive benchmark has been carried out to establish the best practise for encoding, however the few tests being reported for individual models did not provide a decisive conclusion [174, 177, 192].

3.3.4 Implications of the data

TCR specificity is really a triad of interactions between TCR, peptide, and MHC. Only NetTCR-2.0 and ERGO-II seem to address this issue, either by exclusively selecting peptides restricted by one HLA allele or by including the allele as input [173, 204]. Most methods do not consider the implications of including peptides across different HLA restrictions, and probably regard the pMHC as an indivisible unit [127, 174, 176, 177, 187, 192].

Another cause of variation is the length of the sequences, or more precisely the amount of padding [218]. If the data incidentally contains a biologically irrelevant bias of for example a CDR3 sequence length within a peptide or within the group of binders, then the model will quickly learn the padding pattern instead of the underlying pattern within the actual sequence. To mitigate this, peptide and CDR3 lengths were restricted to 9mers & 8-18mers in NetTCR and 8-11mers & 10-20mers in ImREX [173, 187]. Al-

though the strict criteria set by the authors of NetTCR-2.0 may have guided the model to learn proper features determining binding, only three peptides were left with sufficient data for robust modeling [173].

Finally, performance may vary per peptide as reliable predictions require at least 150 TCR chains per peptide and probably improve with increasing numbers [173]

3.3.5 Classification or regression?

In the majority of TCR-pMHC models, the epitope is part of the input data, such that the model is tasked with predicting the probability of binding between the given epitope and TCR [173, 176, 187, 192]. Few methods do not provide the epitope, but instead either train several individual epitope-specific models [174, 177] or train a single model with a confined set of epitopes as a multinomial classification problem [127]. The latter two types are more common in models not based on neural networks such as decision trees and random forests [186, 188, 189].

Submitting the TCR together with the cognate epitope target enables capture of hidden features which potentially reflect universal rules of binding essential for developing pan-specific models. However, with the current composition of TCR-pMHC specificities there is not enough information to produce pan-specific models [173, 174, 176, 187, 192, 204]. The authors of TcellMatch specifically investigated how different designs of classification models would affect performance, and their conclusion was that until more data has surfaced, the multinomial classification models appear to provide the best performance [174].

The majority of models are classifiers due to the the binary character of most TCR screening methods (see section 2.3). However, with the publication of the 10x single-cell specificity data, a new type of measurement became available: the pMHC UMI, i.e. the count of pMHC molecules associated with a single T cell. The UMI can be regarded as a proxy for binding affinity, although the measurement may for example be biased by the cellular expression of TCRs [177]. So far, only DeepTCR [177] and TcellMatch

[174] have experimented with a regression model, even though the related field of pMHC predictions has shown great leverage from combining affinity measurements and eluded ligands as respectively continuous and discrete outputs [219].

3.3.6 Performance and overfitting

Since all the network-based models were published almost simultaneously in 2021, an exhaustive benchmark does not exist, and only few of the publications were able to include a single benchmark [127, 173]. Although each of the models report high performances, it is worth noting that the authors of ImREX state that their generalization performance is likely overestimated, because the current data contains a high degree of similarity which does not require true generalization [187]. The authors of NetTCR-2.0 anticipated this and ensured a maximum of 94% Levenshtein similarity across partitions of training sets and external evaluation sets [173]. There is no mentioning of similarity reduction in any of the other publications, and their high performances are therefore also likely to be overestimated [127, 174, 176, 177, 192, 204].

Models built on any kind of neural network consist of large parameter spaces which are tuned to detect the patterns of the data. The more parameters a model contains the more variation it will capture. The prevailing risk is that the model is overfitted to the training data, i.e. adapts even to the inevitable noise, and thereby looses its ability to properly generalize. Overfitting can be avoided using early-stopping which does not allow the model to become better at predicting its training data than a left-out set, or by regularizing the parameters after completed training. In order for early-stopping to properly work, the data must be partitioned to contain as little similarity as possible, to effectively test the generalization capability. This is typically done using k-fold cross-validation, where k is the number of data partitions and $k-1$ is the number of networks in the ensemble. Bias in a model can also be averaged out by combining multiple networks in an ensemble.

3.3.7 Hidden features

Although neural networks are coined ”black box” models, there is still information in the parameters that may elucidate the detected hidden features. This has been exemplified by Montemurro et al. where t-SNE [220] dimensionality reduction of a hidden layer showed improved clustering of TCRs specific to GIL HLA-A*02:01 than when clustering on physico-chemically encoded CDR3 sequences [173], see figure 3.5a+b. Likewise, Zhang et al. showed by UMAP [221] projection how hidden ”fingerprint” layers revealed two distinct clusters of TCRs specific for GIL HLA-A*0201 [127] (figure 3.5c). These clusters were consistent with experimental findings by Song et al. of two major classes with distinct binding modality toward the GIL peptide [64].

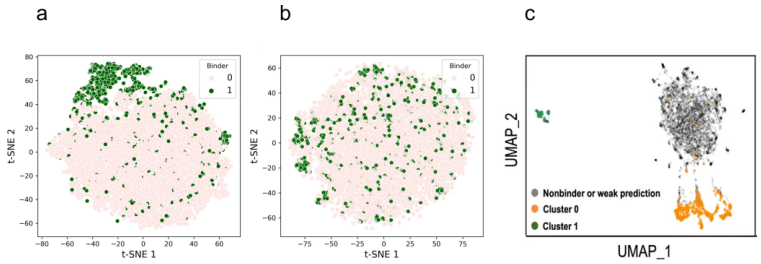


Figure 3.5: Clustering of CDR3 hidden features from NetTCR-2.0 and TCRAI adapted from Monterurro et al., 2021 and Zhang et al., 2021 [127, 173]. (a) t-SNE representation of the CDR3 max-pooled CNN layer of NetTCR-2.0. TCRs positive to GILGFVFTL HLA*A-02:01 are shown in green, and negative TCRs in pink. (b) Same TCRs as in (a) are now instead encoded encoded using a 5-feature physico-chemical scheme. (c) UMAP representation of the fingerprint hidden layer of TCRAI. Two clusters of TCRs positive to GILGFVFTL HLA*A-02:01 are shown, one in orange and the other in green, while negative TCRs are grey.

The investigation of hidden features may guide the field to identify architectures best suited to extract the salient information of TCR-pMHC recognition. As the field is still in its infancy many more types of architectures will be tested moving forward, however, the general conclusion reverberates throughout the field: in reality the missing link to optimal performance is still lack of data - and data diversity.

CHAPTER 4

SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial $CD8^+$ T cell activation in COVID-19 patients

One of the first steps in understanding T cell mediated immunity is the characterization of T cell responses toward selected peptides. The distribution of distinct responses and the sizes thereof are excellent metrics for monitoring disease progression, evaluating response to therapy, or guide rational design of vaccines. The work presented in this chapter maps T cell recognition throughout the proteome of SARS-CoV-2, identifies immunodominant epitopes and investigates the potential of cross-reactive T cells primed for similar common cold corona-viruses.

This project was carried out in a collaboration with Sine Reker Hadrup's group at DTU and Anne Ortved Gang's group at Copenhagen University Hospital Herlev. The experiment was conceived by Anne Ortved Gang, Sine Reker Hadrup, and Sunil Kumar Saini. My primary contribution was development of bioinformatic analyses in close collaboration with Sunil Kumar Saini and Morten Nielsen.

CORONAVIRUS

SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8$^+$ T cell activation in COVID-19 patients

Sunil Kumar Saini[1]*, Ditte Stampe Hersby[2†], Tripti Tamhane[1†], Helle Rus Povlsen[3], Susana Patricia Amaya Hernandez[1], Morten Nielsen[3], Anne Ortved Gang[2], Sine Reker Hadrup[1]*

T cells are important for effective viral clearance, elimination of virus-infected cells, and long-term disease protection. To examine the full spectrum of CD8$^+$ T cell immunity in COVID-19, we experimentally evaluated 3141 major histocompatibility complex (MHC) class I–binding peptides covering the complete SARS-CoV-2 genome. Using DNA-barcoded peptide-MHC complex multimers combined with a T cell phenotype panel, we report a comprehensive list of 122 immunogenic and a subset of immunodominant SARS-CoV-2 T cell epitopes. Substantial CD8$^+$ T cell recognition was observed in patients with COVID-19, with up to 27% of all CD8$^+$ lymphocytes interacting with SARS-CoV-2–derived epitopes. Most immunogenic regions were derived from open reading frame 1 (ORF1) and ORF3, with ORF1 containing most of the immunodominant epitopes. CD8$^+$ T cell recognition of lower affinity was also observed in healthy donors toward SARS-CoV-2–derived epitopes. This preexisting T cell recognition signature was partially overlapping with the epitope landscape observed in patients with COVID-19 and may drive the further expansion of T cell responses to SARS-CoV-2 infection. The phenotype of the SARS-CoV-2–specific CD8$^+$ T cells revealed a strong T cell activation in patients with COVID-19, whereas minimal T cell activation was seen in healthy individuals. We found that patients with severe disease displayed significantly larger SARS-CoV-2–specific T cell populations compared with patients with mild diseases, and these T cells displayed a robust activation profile. These results further our understanding of T cell immunity to SARS-CoV-2 infection and hypothesize that strong antigen-specific T cell responses are associated with different disease outcomes.

INTRODUCTION

The COVID-19 (coronavirus disease 2019) pandemic caused by the highly infectious SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) has challenged public health at an unprecedented scale, causing the death of more than 2 million people worldwide so far (*1*). T cells perform essential functions in the control and elimination of viral infections; CD8$^+$ T cells are critical for efficient clearance of virus-infected cells, whereas CD4$^+$ T cells are important for supporting both the CD8$^+$ T cell response and B cell–mediated production of specific antibodies. Characteristics from the ongoing pandemic suggest that T cell recognition will be critical to mediate long-term protection against SARS-CoV-2 (*2*), because the antibody-mediated response seems to decline in a follow-up evaluation of convalescent patients, although it is not yet understood how this affects the risk of reinfection and what antibody levels are required for disease protection (*3*–*5*). Furthermore, studies of the closely related SARS-CoV infection show persistent memory CD8$^+$ T cell responses even after 11 years in SARS recovered patients without B cell responses (*6*, *7*), emphasizing the potential role of CD8$^+$ memory T cells in long-term protection from coronaviruses.

Several recent studies have reported robust T cell immunity in SARS-CoV-2–infected patients (*8*–*10*), and unexposed healthy individuals also showed functional T cell reactivity restricted to SARS-CoV-2 (*9*, *11*–*15*). The observed T cell cross-reactivity is hypothesized to derive from routine exposure to common cold coronaviruses [human coronavirus (HCoV)] (HCoV-OC43, HCoV-HKU1, HCoV-NL63, and HCoV-229E) that widely circulate, with 90% of the human population being seropositive for these viruses (*16*, *17*) and substantial sequence homology to the SARS-CoV-2 genome (*18*, *19*). However, the influence of such preexisting immunity to the T cell recognition associated with COVID-19 disease is poorly understood.

SARS-CoV-2 infection can result in mild to severe disease (including death), but a large number of asymptomatic infections are also described (*20*–*22*). The presence of preexisting T cell immunity, represented by cross-reactive T cells, could have strong implications for how individuals respond to SARS-CoV-2 infection. However, their biological role upon encounter with SARS-CoV-2 infection remains unclear, and their contribution to disease protection needs to be determined. Furthermore, in severe clinical disease, cytokine release syndrome is reported and might, in some cases, be dampened by immunosuppressive medication or anti–interleukin-6 (IL-6) antibody therapy (*23*, *24*). Such clinical characteristics point to a potential uncontrolled immune response with the involvement of strong T cell activation.

CD8$^+$ T cells are activated by a specific interaction between the T cell receptor (TCR) and the peptide antigen presented by major histocompatibility complex class I (MHC-I) molecules on the surface of virus-infected cells. Although SARS-CoV-2–specific immunity has been reported both in the context of COVID-19 and preexisting T cells, the full spectrum of exact antigens (minimal peptide epitope) within the viral genome, associated with this immunity and their immunodominance in SARS-CoV-2–infected patients, is not fully described. Using our large-scale T cell detection technology based on DNA-barcoded peptide-MHC (pMHC) multimers (*25*), we have mapped T cell recognition throughout the complete SARS-CoV-2

[1]Department of Health Technology, Section of Experimental and Translational Immunology, Technical University of Denmark, Kongens Lyngby, Denmark. [2]Department of Haematology, Herlev Hospital, Copenhagen University Hospital, Herlev, Denmark. [3]Department of Health Technology, Section of Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark.
*Corresponding author. Email: sirha@dtu.dk (S.R.H.); sukusa@dtu.dk (S.K.S.)
†These authors contributed equally to this work.

genome, identified the exact epitopes recognized by SARS-CoV-2–specific CD8$^+$ T cells, and characterized immunodominance of these epitopes in COVID-19 disease. Broad T cell recognition toward SARS-CoV-2–derived peptides was also identified in SARS-CoV-2–unexposed healthy individuals, with a large overlap in the pMHC complexes recognized in the two groups. However, T cell recognition was substantially enhanced in the patient group, with SARS-CoV-2–reactive T cells accounting for up to 27% of all CD8$^+$ T cells. Furthermore, we have evaluated the phenotypic characteristics of SARS-CoV-2–specific T cells and correlated their activation signatures with disease severity.

RESULTS

SARS-CoV-2–specific CD8$^+$ T cells recognize a broad range of epitopes

To reveal the full spectrum of T cell immunity in COVID-19 disease, we used a complete SARS-CoV-2 genome sequence (*26*) to identify immunogenic minimal epitopes recognized by CD8$^+$ T cells. Using NetMHCpan 4.1 (*27*), we selected 2204 potential human leukocyte antigen (HLA)–binding peptides (9 to 11 amino acids) for experimental evaluation. These peptides were predicted to bind one or more of 10 prevalent MHC-I molecules, including HLA-A (A01:01, A02:01, A03:01, and A24:02), HLA-B (B07:02, B08:01, and B15:01), and HLA-C (C06:02, C07:01, and C07:02) loci, leading to a total 3141 pMHC specificities for experimental evaluation (Fig. 1A and table S1). Epitope predictions are covering the full viral genome, with open reading frame 1 (ORF1) being the largest gene region and hence including the highest number of predicted peptides (Fig. 1B). T cell reactivity toward these peptides was analyzed for 18 patients with COVID-19. In this cohort, 11 patients had severe disease requiring hospital care, and 7 patients had mild disease not requiring hospitalization. Blood samples were collected during the active phase of the infection, as close as possible after the first positive SARS-CoV-2 test (table S2). The mean HLA coverage that could be obtained using the 10 selected MHC-I molecules was 3.1 HLA per patient, and patients were evaluated using on average 972 DNA-barcoded pMHC multimers per patient (fig. S1A) (*25*). Briefly, each pMHC complex is multimerized on a PE (phycoerythrin)–labeled or APC (allophycocyanin)–labeled dextran backbone and tagged with a unique DNA barcode. DNA-barcoded pMHC multimers are then pooled to generate an HLA-matching patient-tailored pMHC multimer panel, which is incubated with patient-derived PBMCs (peripheral blood mononuclear cells), and multimers bound to CD8$^+$ T cells are sorted and sequenced to identify T cell recognition toward the probed pMHC complexes. For comparative evaluation, we also included 39 T cell epitopes from common viruses: cytomegalovirus (CMV), Epstein-Barr virus (EBV), and influenza (flu) virus (CEF) (Fig. 1C and table S3).

We found broad and strong SARS-CoV-2–specific CD8$^+$ T cell responses in patients with COVID-19, contributing up to 27% of the total CD8$^+$ T cells (Fig. 1D). A substantial selection of T cells specific to individual immunogenic epitopes measuring up to 14% of the total T cells was detected in several patients (Fig. 1D, fig. S2, and table S4). In total, we identified T cell responses to 142 pMHC complexes corresponding to 122 unique SARS-CoV-2 T cell epitopes across the 10 analyzed HLAs (Fig. 1E) dominated by peptides with high-affinity binding to their corresponding HLA molecule (fig. S1B). We also detected 25 T cell responses to CEF-derived peptides across the 18 patients with COVID-19 (Fig. 1E and table S5). For the SARS-CoV-2–derived peptides, HLA-A01:01, HLA-A02:01, and HLA-B15:01 presentation dominated in terms of the total number of identified epitopes as well as the "immunogenicity score" (i.e., the number of T cell responses normalized to the number of probing pMHC multimers and the number of patients analyzed) (Fig. 1F). HLA-A03:01– and C07:01-specific peptides showed the least T cell reactivity (three epitopes each) despite being analyzed in nine and six patients, respectively (Fig. 1E). Most of the immunogenic epitopes were mapped to the ORF1 protein, followed by S and ORF3 proteins (Fig. 1, G and H, and table S4). Given the size difference of the viral proteins, the immunogenicity score was used to evaluate their relative contribution to T cell recognition. On the basis of such evaluation, we observe that peptides derived from ORF3 displayed the highest relative immunogenicity (in terms of T cell recognition), followed by ORF1 protein (Fig. 1H). The overall frequency of SARS-CoV-2–reactive T cells (the sum of estimated frequencies for all SARS-CoV-2–specific T cells) in individual patients with COVID-19 showed a broad range of T cell involvement and variation in terms of T cell recognition to individual SARS-CoV-2 proteins (Fig. 1I).

In summary, we report SARS-CoV-2–specific CD8$^+$ T cell immunity toward several epitopes and a substantially high presence of SARS-CoV-2–specific T cells in several patients with COVID-19. The ORF1 protein not only contributes the most to T cell recognition of SARS-CoV-2 but is also by far the largest group of proteins. When protein size is considered, ORF3 and ORF1 are the viral regions most frequently recognized by CD8$^+$ T cells.

Strong immunodominance of SARS-CoV-2–derived peptides in patients with COVID-19

Of the 122 epitopes recognized by T cells in the patient cohort, 5 were determined as "immunodominant" based on the presence of T cell recognition in >50% of the tested samples with the given HLA molecule and T cell detection identified in at least two or more patients (Fig. 2A). Unexpectedly, in our patient cohort, none of the immunodominant epitopes were derived from the S protein, despite this being the second largest protein (Fig. 2B). Among the immunodominant epitopes, a very robust HLA-associated immunodominance was observed for two of the epitopes: HLA-A01:01-TTDPSFLGRY–specific (and its variant peptides TTDPSFLGRYM and HLA-A01:01-TDPSFLGRY), with specific T cells detected in all five analyzed patients (estimated frequency reaching up to 25% of total CD8$^+$ T cells), and HLA-B07:02-SPRWYFYYL, with specific T cells observed in four of the five patients evaluated (estimated frequency up to 10%) (Fig. 2A and table S4). To validate the T cell responses identified for the two most immunodominant epitopes (TTDPSFLGRY and SPRWYFYYL), we determined the presence of these T cells using conventional fluorophore-labeled pMHC tetramers in seven patients with COVID-19. For both immunodominant epitopes, the frequency of T cells determined by the individually labeled pMHC tetramers correlated to the frequencies determined based on the DNA barcode–labeled MHC multimer reagents (at a range from 0.01 to 11% of the total CD8$^+$ T cells) (Fig. 2, C and D). Next, we evaluated the cytokine secretion capacity of the SARS-CoV-2–specific T cells by stimulating PBMCs (same time point as used for T cell identification) with respective epitopes. SARS-CoV-2 peptide–induced secretion of interferon-γ (IFN-γ) and tumor necrosis factor–α (TNF-α) was detected in all seven patients, confirming functional activation of T cells raised against dominant and nondominant epitopes (Fig. 2E and table S6).

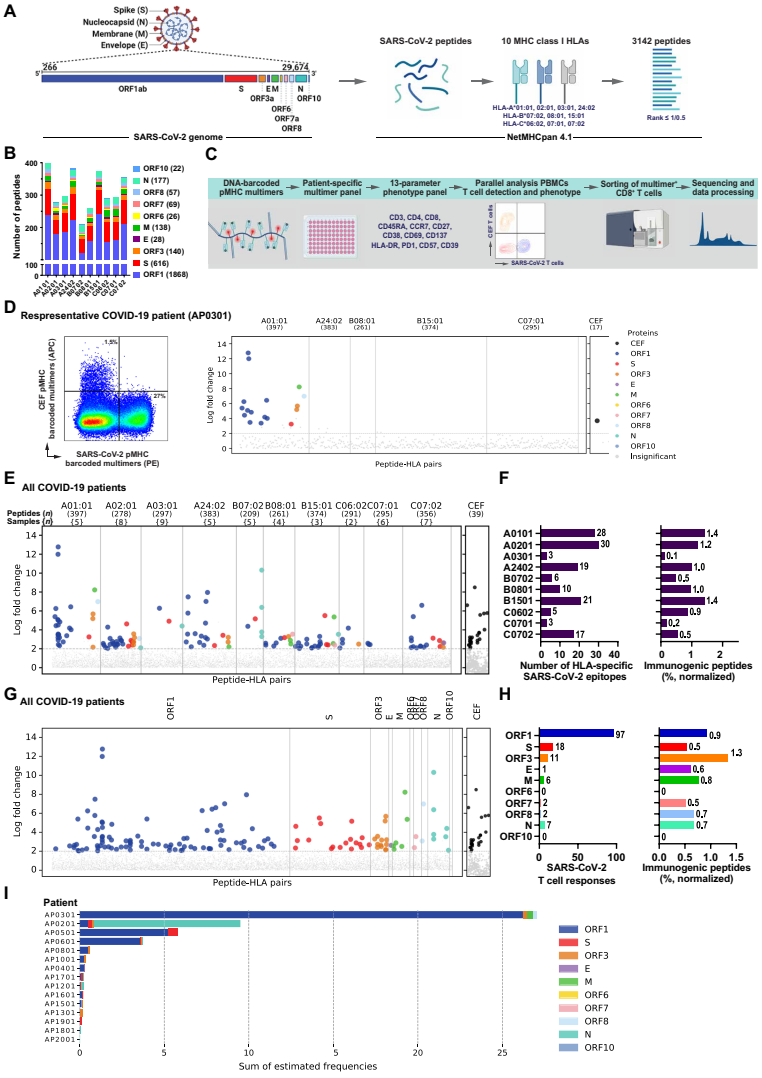


Fig. 1. CD8$^+$ T cell epitope mapping in SARS-CoV-2. (**A**) Schematic representation of the complete SARS-CoV-2 genome used for the identification of 3141 peptides with predicted binding rank (NetMHCpan 4.1) of ≤0.5 (ORF1 protein) and ≤1 (all remaining proteins) for 10 prevalent HLA-A, HLA-B, and HLA-C molecules. (**B**) Bar plot showing the distribution of SARS-CoV-2 peptides related to their HLA-restriction (3141 peptide-HLA pairs) across the viral genome. Total pMHC specificities analyzed for each protein are shown in parentheses next to the respective SARS-CoV-2 protein. (**C**) Experimental pipeline to analyze T cell recognition toward the SARS-CoV-2–derived HLA-binding peptides in PBMCs using pMHC multimers. A 13-antibody panel was used for phenotype analysis of pMHC multimer$^+$ CD8$^+$ T cells. pMHC multimers binding CD8$^+$ T cells were sorted on the basis of PE (SARS-CoV-2–specific) or APC (CEF-specific) signal and sequenced to identify antigen-specific CD8$^+$ T cells. (**D**) Representative analyses for SARS-CoV-2–restricted T cell populations in a patient with COVID-19. Left: Flow cytometry plot of pMHC multimer staining of CD8$^+$ T cells from a patient with COVID-19 stained with pMHC multimer panel showing SARS-CoV-2 (PE) and CEF (APC) multimer$^+$ T cells that were sorted for DNA barcode analysis to identify epitope recognition. Right: CD8$^+$ T cell recognition to individual epitopes was identified on the basis of the enrichment of DNA barcodes associated with each of the tested peptide specificities (LogFc > 2 and $P < 0.001$, using Barracoda). Significant T cell recognition of individual peptide sequences is colored on the basis of their protein of origin and segregated on the basis of their HLA specificity. The black dot shows CD8$^+$ T cells reactive to one of the CEF peptides (here, CMV pp65; YSEHPTFTSQY-HLA-A01:01). All peptides with no significant enrichments are shown as gray dots. (**E**) Summary of all T cell recognition to SARS-CoV-2–derived peptides identified in the 18 analyzed patients with COVID-19. In parentheses, number of peptides tested for each HLA (top row) and the number of patients analyzed for each HLA pool (bottom row). Each dot represents one peptide-HLA combination per patient and is colored according to their origin of protein, similar to that shown in (A). The black dots show CD8$^+$ T cells reactive to the CEF peptides in all analyzed patients. (**F**) Bar plots summarize the number of HLA-specific SARS-CoV-2 epitopes identified and the HLA-restricted immunogenicity (% immunogenic peptides) in the analyzed patient cohort. Immunogenicity represents the fraction of T cell–recognized peptides out of the total number of peptides analyzed for a given HLA restriction across the HLA-matching donors (% normalized). (**G**) Similar to (E), a summary of SARS-CoV-2–specific T cell responses separated based on the protein of origin. (**H**) Bar plots show the number of epitopes derived from each of the SARS-CoV-2 protein and their immunogenicity score (% immunogenic peptides). (**I**) Estimated frequencies (% of total CD8$^+$ T cells) as the sum of all SARS-CoV-2 epitope–reactive T cells identified in individual patients with COVID-19. Bars are colored according to the protein origin of the recognized epitopes.

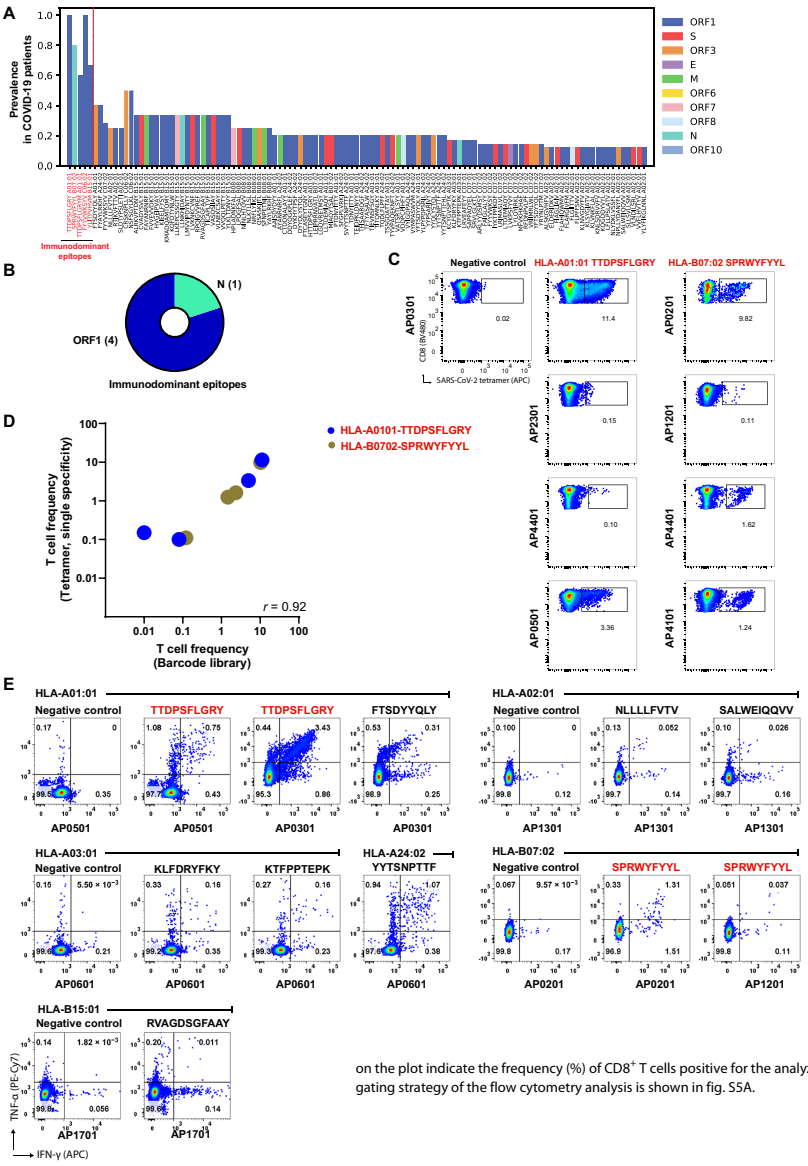


Fig. 2. Strong immunodominance of SARS-CoV-2–derived peptides in patients with COVID-19. (**A**) The prevalence of T cell recognition toward the individual epitopes detected in patients with COVID-19. Dotted line indicates the epitopes determined as immunodominant, based on the presence of T cell recognition in more than 50% analyzed patients (marked in red throughout this figure). Bars are colored according to their protein of origin, similar to that shown in Fig. 1. (**B**) Pie chart of immunodominant epitopes distributed according to their protein of origin. (**C**) APC-labeled pMHC tetramer–based analyses of CD8$^+$ T cells in PBMCs of seven patients with COVID-19 recognizing A01:01/TTDPSFLGRY or B07:02/SPRWYFYYL immunodominant epitopes. Gated population shows the percentage of T cells recognizing pMHC tetramers out of total CD8$^+$ T cells. (**D**) Correlation of estimated T cell frequencies determined using DNA-barcoded multimer estimation (table S4) and using conventional tetramer analysis (C) for the two most immunodominant epitopes in seven patients with COVID-19. Spearman correlation ($P = 0.002$, $r = 0.92$). (**E**) Functional validation of SARS-CoV-2–specific T cell responses identified in COVID-19 patient samples. Flow cytometry plots showing intracellular cytokine staining of PBMCs from patients with COVID-19 pulsed with selected SARS-CoV-2–derived peptides, selected based on the CD8$^+$ T cell responses identified from the DNA-barcoded multimer analysis (table S4). PBMCs from seven patients with COVID-19 were analyzed for functional activation after stimulation with the indicated peptides or with an HLA-matching irrelevant peptide (negative control). The numbers on the plot indicate the frequency (%) of CD8$^+$ T cells positive for the analyzed cytokines IFN-γ and TNF-α. The gating strategy of the flow cytometry analysis is shown in fig. S5A.

Low-avidity recognition toward SARS-CoV-2–derived peptides in healthy individuals

To examine the potential for preexisting SARS-CoV-2–reactive T cells, we next analyzed healthy individuals for T cell recognition against all 3141 SARS-CoV-2–derived peptides. We selected two healthy donor cohorts: The first cohort included SARS-CoV-2–unexposed healthy individuals (HD-1; $n = 18$ individuals, PBMCs collected before the COVID-19 pandemic), and the second cohort included

health care staff at high risk of SARS-CoV-2 exposure but who did not test positive (HD-2; $n = 20$ individuals, PBMCs collected during COVID-19 pandemic). CD8$^+$ T cells from SARS-CoV-2–unexposed healthy individuals showed broad-scale T cell recognition toward SARS-CoV-2–derived peptides across the whole viral genome (Fig. 3A, fig. S3, and table S7). Cumulatively, 214 SARS-CoV-2–derived peptides were recognized by T cells in 16 of the 18 analyzed samples. The high-risk COVID-19 healthy cohort showed similar T cell recognition toward 178 SARS-CoV-2 epitopes (Fig. 3B and table S7) in 15 of the 20 donors. T cell recognition in healthy donors was directed

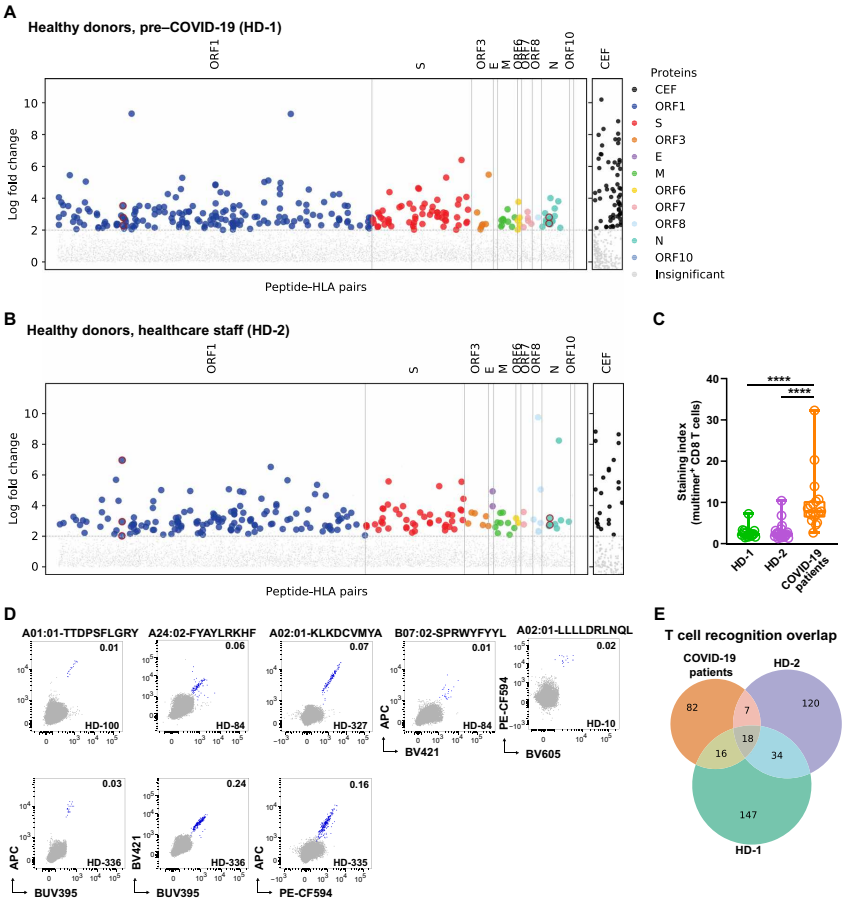


Fig. 3. Broad reactivity toward SARS-CoV-2–derived peptides in healthy individuals. (**A**) CD8$^+$ T cell recognition to individual SARS-CoV-2–derived peptides (table S7) and CEF peptides (table S5) in the pre–COVID-19 healthy donor cohort ($n = 18$ individuals) identified based on the enrichment of DNA barcodes associated with each of the tested peptide specificities (LogFc > 2 and $P < 0.001$, Barracoda). Significant SARS-CoV-2–specific T cell recognition of individual peptide sequences is colored and segregated based on their protein of origin. The black dots show CD8$^+$ T cells reactive to the CEF peptides in all analyzed donors. (**B**) T cell recognition in the high-exposure risk healthy donor cohort (tables S5 and S7) ($n = 20$ individuals). (**C**) Staining index of CD8$^+$ T cells binding SARS-CoV-2–specific pMHC multimers in the three evaluated cohorts. One-way ANOVA (Kruskal-Wallis test) ****$P < 0.0001$ (patient versus HD-1 < 0.0001 and patient versus HD-2 < 0.0001); $n = 18$ (patient), $n = 18$ (HD-1), and $n = 20$ (HD-2). (**D**) Flow cytometry dot plots showing in vitro expanded T cells from healthy donors recognizing SARS-CoV-2–derived epitopes, detected by combinatorial tetramer staining. T cell binding to each pMHC specificity is detected using pMHC tetramers prepared in a two-color combination (blue dots), with gray dots showing tetramer-negative T cells, and the number on the plots shows the frequency (%) of tetramer$^+$ of the CD8$^+$ T cells. Gating strategy used for the flow cytometry analysis is shown in fig. S11A. (**E**) Venn diagram illustrating the overlap of T cell recognition toward SARS-CoV-2–derived peptides in the COVID-19 patient and healthy donor cohorts.

equally toward ORF1 and S proteins, whereas ORF3-derived peptides were recognized less in the healthy donor cohort compared with the COVID-19 patient cohort (fig. S3B). The immunodominant T cell epitopes from ORF1 identified in the patient cohort were not among the most prevalent responses in the healthy donors (fig. S3C).

Despite such broad T cell recognition in both healthy donor cohorts, the presence of SARS-CoV-2–recognizing T cells seems to be of low frequency with limited separation of the $CD8^+$ T cells binding to the pool of DNA-barcoded pMHC multimers (fig. S4A) and measured by a significantly lower staining index of the pMHC multimer binding in healthy donors compared with patients (Fig. 3C). Consequently, a direct estimate of the frequency of the SARS-CoV-2–reactive T cell populations in the individual healthy donors was not feasible. The low frequency and limited separation of these T cells were confirmed by independent analysis using conventional pMHC tetramers for several individual epitopes in healthy donor PBMCs (fig. S4B). Together, these data suggest a lower TCR avidity to the probed pMHC in healthy individuals compared with patients with COVID-19, which could represent potential cross-reactivity from existing T cell populations potentially raised against other coronaviruses (such as common cold viruses HCoV-HKU1, HCoV-229E, HCoV-NL63, and HCoV-OC43) that share some level of sequence homology with SARS-CoV-2, as suggested in recent reports (*13*, *17*, *19*).

To further validate the presence of low-frequency T cells in healthy donors, we expanded T cells in vitro from several COVID-19–unexposed healthy donors and measured T cell binding using conventional pMHC tetramers. On the basis of in vitro peptide-driven expansion, pMHC tetramer binding T cell populations were verified in multiple donors, recognizing SARS-CoV-2–derived peptides, including immunodominant epitopes across four HLAs (A01:01-TTDPSFLRGY, A02:01-LLLLDRLNQL, A02:01-KLKDCVMYA, A24:02-FYAYLRKHF, and B07:02-SPRWYFYYL) (Fig. 3D). Although these T cell responses were of low frequency, a functional cytokine response (measured by IFN-γ and TNF-α production) was observed in in vitro expanded T cell cultures when restimulated with individual peptide epitopes or epitope pools (fig. S5). Forty-one of the COVID-19 immunogenic peptides, including the immunodominant peptides, identified in the patient cohort were also recognized by T cells of healthy donors; this includes the two most frequently observed epitopes of SARS-CoV-2: HLA-A01:01-TTDPSFLRGY and HLA-B07:02-SPRWYFYYL (Fig. 3E and table S7). Together, we show a full spectrum of T cell recognition toward SARS-CoV-2–derived peptides in healthy donors; this is detected at low frequency and shows characteristics of low-avidity interaction based on the staining index of the pMHC multimer interaction.

Enhanced activation profile of SARS-CoV-2–specific T cells associated with COVID-19 disease severity

For phenotypic characterization of SARS-CoV-2–specific $CD8^+$ T cells, we combined pMHC multimer analysis with a 13-parameter antibody panel (table S8) and evaluated the phenotype of the SARS-CoV-2–reactive T cell populations in patients with COVID-19 and healthy donors. This furthermore allowed us to evaluate whether the multimer-specific T cell profile of the high-risk COVID-19 healthy cohort (HD-2) has any distinct features compared with the unexposed cohort (HD-1), despite both cohorts containing presumably unexposed individuals. Dimensional reduction using Uniform Manifold Approximation and Projection (UMAP) showed distinct clustering of SARS-CoV-2 multimer-reactive T cells of the COVID-19 patient cohort compared with the two healthy donor cohorts with higher expression of activation markers CD38, CD69, CD39, HLA-DR, and CD57 and reduced expression of CD8 and CD27 (fig. S6). Compared with both healthy donor cohorts, we observed that more SARS-CoV-2–reactive T cells from patients with COVID-19 expressed the activation markers CD38, CD39, CD69, and HLA-DR and showed a late-differentiated effector memory (EM) profile of reduced CD27 (Fig. 4A). We did not observe activation of SARS-CoV-2–specific $multimer^+$ T cells in the high-risk COVID-19 healthy cohort, except for nonsignificant trends for reduced CD27 and increased CD57 expression (Fig. 4A). SARS-CoV-2–reactive T cells in patients and healthy donor cohorts showed a similar distribution of memory subsets (determined by CCR7 and CD45RA expression); however, higher expression of T cell activation markers (fig. S7) was observed in EM and TEMRA (terminally differentiated EM) subsets in patients. Furthermore, the highly activated and differentiated T cell phenotype in patients with COVID-19 was distinct to SARS-CoV-2–specific T cells and not observed for CEF-specific T cells detected in the same cohort (Fig. 4B). We also observed no difference in CEF-specific $multimer^+$ T cells between the three cohorts in a similar analysis (fig. S8A). In addition, we compared the expression of T cell activation markers in combination with the inflammatory response marker CD38 on $multimer^+$ $CD8^+$ T cells across the three cohorts, which showed significantly enhanced expression of activation molecules (CD39, CD69, and HLA-DR) and PD-1 inhibitory receptor on $CD38^+$ T cells only in the patient cohort (fig. S8, B and C).

We next evaluated the association of SARS-CoV-2–specific $CD8^+$ T cell presence in the patient cohort related to their requirement for hospital care. No overall difference in the total number of recognized SARS-CoV-2–derived epitopes was observed between severely diseased patients requiring hospitalization (n = 11 individuals) and patients with mild symptoms not requiring hospital care (outpatient; n = 7 individuals) (Fig. 4C). For phenotype characterization, 23 additional patient samples (total, n = 41 patients; hospitalized, n = 21; outpatients, n = 20) were analyzed using a patient HLA-matching pMHC multimer library combined with the 13-parameter antibody panel, similar to the initial 18 patients but without resolving individual epitope specificities. On the basis of this extended cohort, a significantly higher frequency of SARS-CoV-2–specific $CD8^+$ T cells was observed in the hospitalized patients compared with outpatient samples (Fig. 4D). Furthermore, a significant increase in the fraction of such cells expressing CD38, CD39, HLA-DR, and PD-1 was observed in the hospitalized patients (Fig. 4E). By measuring the coexpression of immune activation markers—CD38 together with CD39, PD-1, and HLA-DR—a strong elevation in T cells expressing these combinations of activation markers was observed among the hospitalized patients (Fig. 4F). Together, the increased frequency and activation signature suggest a role for SARS-CoV-2–specific $CD8^+$ T cells in severe COVID-19 disease.

We also examined the phenotype of $CD8^+$ T cells specific to the two most immunodominant epitopes TTDPSFLRGY and SPRWYFYYL with respect to disease severity (in eight patients; four hospitalized and four outpatients) using conventional pMHC tetramer–based evaluation of individual T cell specificities. Hospitalized patients displayed increased PD-1 expression compared with the same T cell populations in the outpatients (fig. S9A). Furthermore, a higher frequency of T cells reactive to these two SARS-CoV-2 immunodominant epitopes was observed in the hospitalized patients, but the functional evaluation upon peptide stimulation revealed that only a subfraction

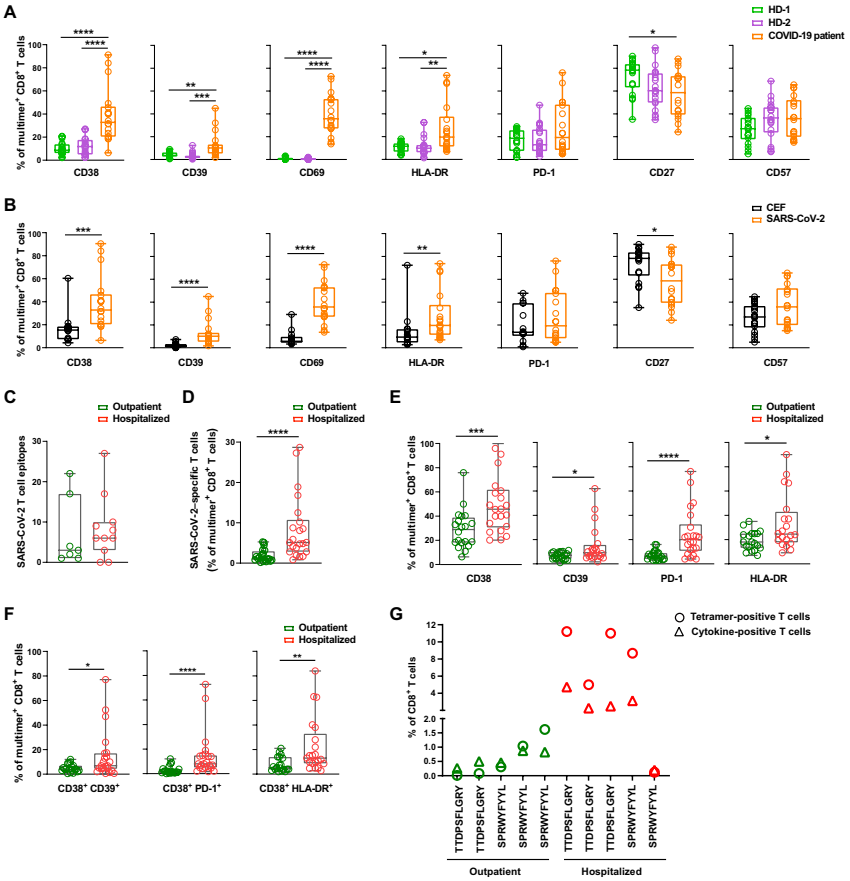


Fig. 4. Enhanced activation profile of SARS-CoV-2–specific T cells correlates with COVID-19 disease severity. (**A**) Box plots comparing percentages of SARS-CoV-2 pMHC multimer binding $CD8^+$ T cells expressing the indicated phenotype surface markers in the COVID-19 patient and the two healthy donor cohorts ($n = 18$ individuals for each cohort). Each dot represents one sample. Frequencies were quantified from flow cytometry data processed using the gating strategy applied in fig. S11. *P* values for one-way ANOVA (Kruskal-Wallis test): CD38 < 0.0001 (HD-1 versus patient < 0.0001 and HD-2 versus patient < 0.0001), CD39 < 0.0001 (HD-1 versus patient = 0.006 and HD-2 versus patient < 0.0001), CD69 < 0.0001 (HD-1 versus patient < 0.0001 and HD-2 versus patient < 0.0001), HLA-DR = 0.002 (HD-1 versus patient = 0.02 and HD-2 versus patient < 0.004), and CD27 = 0.03 (HD-1 versus patient = 0.03). (**B**) Box plots comparing the percentage of SARS-CoV-2 pMHC multimer$^+$ ($n = 18$ patients) and CEF pMHC multimer$^+$ ($n = 14$ patients) $CD8^+$ T cells expressing the indicated surface markers in the COVID-19 patient cohort. Each dot represents one sample. *P* values for hypothesis (Mann-Whitney) test: $P = 0.0002$ (CD38), $P < 0.0001$ (CD39), $P = 0.0001$ (CD69), $P = 0.009$ (HLA-DR), and $P = 0.04$ (CD27). (**C**) Number of SARS-CoV-2 epitopes recognized by T cells in outpatient ($n = 7$) and hospitalized ($n = 11$) patient samples. (**D**) Box plots show frequencies of SARS-CoV-2 multimer$^+$ $CD8^+$ T cells in outpatient ($n = 20$) and hospitalized patients ($n = 21$). *P* value (Mann-Whitney test) of ≤0.0001. (**E**) Box plots showing the percentage of SARS-CoV-2 pMHC multimer$^+$ $CD8^+$ T cells expressing the indicated surface markers in outpatients ($n = 20$) and hospitalized patients ($n = 21$). *P* values for hypothesis (Mann-Whitney) test: $P = 0.001$ (CD38), $P = 0.036$ (CD39), $P < 0.0001$ (PD-1), and $P = 0.027$ (HLA-DR). (**F**) Comparison of the frequency of SARS-CoV-2 pMHC multimer$^+$ $CD8^+$ T cells expressing activation markers (CD39 and HLA-DR) and PD-1 in combination with CD38 [as shown in the representative plots (fig. S8), in hospitalized and outpatient samples]. *P* values for hypothesis (Mann-Whitney) testing: $P = 0.04$ ($CD38^+$ $CD39^+$), $P = 0.005$ ($CD38^+$ $HLA\text{-}DR^+$), and $P < 0.0001$ ($CD38^+$ $PD\text{-}1^+$). (**G**) Comparison of tetramer binding (conventional single-color tetramers) and functional (cytokine-secreting) T cells recognizing the two immunodominant epitopes in 10 patients, grouped according to COVID-19 disease severity. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, and ****$P \leq 0.0001$.

of these high-frequency T cells were responsive to antigen exposure (Fig. 4G and fig. S9B). These data, together with increased PD-1 expression, suggest a functional impairment or selective inhibition of these high-frequency T cell populations, as observed by a recent study (*5*).

A fraction of SARS-CoV-2 epitopes share sequence homology with widely circulating common cold coronaviruses

Preexisting T cell immunity, in the context of SARS-CoV-2–reactive T cells in unexposed healthy individuals, has been reported by several studies (*13*–*15*, *17*, *19*), and it has been hypothesized that this is due to the shared sequence homology between the SARS-CoV-2 genome and other common cold coronaviruses (HCoV-OC43, HCoV-HKU1, HCoV-NL63, and HCoV-229E). Having evaluated the full spectrum of minimal epitopes for T cell recognition, we sought to evaluate the sequence homology at the peptide level and its association with the SARS-CoV-2 T cell reactivity that we observed in healthy donors. First, we searched for immunogenic hotspots across the full SARS-CoV-2 proteome by comparing the number of identified epitopes (in the patient cohort) with the total number of predicted peptides in any given region of the proteins. In general, the epitopes were spread over the full length of the protein sequences while clustering in minor groups throughout all regions of the viral proteome (Fig. 5A). Regions indicated by an asterisk demonstrate significant enrichment of T cell recognition relative to the number of MHC-I–binding peptides in a given region. Both the C- and N-terminal regions of the ORF1 seem to hold fewer T cell epitopes compared with the rest of this protein. When similarly mapping the T cell recognition of SARS-CoV-2–derived peptides observed in healthy donors, we detected a comparable spread of T cell recognition in the healthy donor cohort. Most T cell epitope clusters in the patient cohort coincide with T cell recognition in the healthy donor cohort. The few regions that distinguish the T cell recognition observed in healthy donors from that observed in patients include the C- and N-terminal regions of ORF1, parts of the N, and, in general, a higher level of T cell recognition to S. In these regions, T cell recognition in healthy donors exceeded the observation from patients with COVID-19 (Fig. 5A). When evaluating the prevalence of T cell recognition for the epitopes identified in >25% of the patient (Fig. 2A and table S9) or the healthy donor cohort (fig. S3C and table S9), we observed that most of these T cell responses frequently observed in patients with COVID-19 are also detected in healthy donors, whereas a large fraction of epitopes dominating in healthy individuals were not detected in our patient cohort (Fig. 5B). However, several SARS-CoV-2 reactivities that were identified only in the healthy donors in our study were shown to be present in patients with COVID-19 analyzed by other studies (table S9), which strongly points to a substantial degree of cross-recognition to SARS-CoV-2 from preexisting T cell populations and that such populations might drive the further expansion of T cell responses to SARS-CoV-2 infection.

To further elucidate the potential origin of such a cross-reactive T cell population in the healthy donor cohort, we next evaluated the sequence homology of SARS-CoV-2 MHC-I–binding peptides with the four common cold coronaviruses: HCoV-HKU1, HCoV-NL63, HCoV-OC43, and HCoV-229E. With a variation limit of up to two amino acids in each peptide sequence, 15% of the total predicted peptides showed sequence similarity with one or more HCoV peptide sequence (Fig. 5C, gray pie). Among the T cell–recognized peptides, in both the patient and healthy donor cohorts, this fraction was comparable with 19 and 16%, respectively, of T cell–recognized peptides sharing sequence homology with one or more HCoV (Fig. 5C). As an alternative approach, the similarities were calculated by kernel method for amino acid sequences using BLOSUM62, indicating comparable sequence similarity of the peptides recognized by T cells and those not recognized in reference to HCoV. However, peptides with the lowest similarity to HCoV were not recognized by T cells in the patient cohort (fig. S10).

Because T cell cross-recognition can often be driven by a few key interaction points, predominantly in the "core" of the peptide sequence (i.e., positions 3 to 8) (*28*, *29*), we restricted the sequence similarity to the core of the peptide that would be most likely to interact with the TCR (*30*). On the basis of the protein core only, up to 74% of all the identified epitopes showed sequence homology to HCoV (one or more) (Fig. 5C), suggesting these common cold viruses as a potential source of the observed low-avidity interactions in healthy donors. Furthermore, when evaluating peptides frequently recognized by T cells in both patients with COVID-19 and healthy individuals, we find evidence of substantial homology, as exemplified with the peptide sequences listed in Fig. 5D. However, similar sequence homology is observed for the peptide sequences that are recognized only in the patient cohort (Fig. 5D). Thus, at present, our data point to substantial T cell cross-recognition being involved in shaping the T cell response to SARS-CoV-2 in patients with COVID-19; however, we find no specific enrichment of T cell recognition to peptide sequences with large sequence homology compared with the total peptide library being evaluated, and the responses identified exclusively in the patient samples are not more specific to SARS-CoV-2 compared with those recognized in both cohorts. ORF1 displayed the highest T cell recognition immunogenicity and also the highest sequence identity to HCoV (40%, as opposed to 22 to 34% for all other SARS-CoV-2 proteins, calculated using direct sequence alignment). Future studies seeking to fully understand the role and origin of the underlying T cell cross-recognition will likely require an in-depth evaluation of pre- and postinfection samples.

DISCUSSION

Several studies using overlapping peptide pools spanning different regions of SARS-CoV-2 viral proteins have shown a broad range of T cell activation in convalescent COVID-19 patients (*8*, *9*, *11*, *14*, *15*, *31*–*35*). Our work now provides a detailed characterization of minimal epitopes derived from the complete SARS-CoV-2 genome for their $CD8^+$ T cell immunogenicity, immunodominance, and functional and phenotypical characteristics in patients with COVID-19 and healthy donors. We identified $CD8^+$ T cell responses to 122 epitopes in 18 patients with COVID-19 after screening for T cell recognition based on 3141 peptides derived from the full SARS-CoV-2 genome and selected based on their predicted HLA-binding capacity. Of these, a few immunodominant T cell epitopes were recognized in most of the patients. Both dominant and subdominant T cell epitopes were cross-recognized by low-level preexisting T cell populations in SARS-CoV-2–unexposed healthy individuals. We have observed that the SARS-CoV-2 dominant epitopes mount very strong T cell responses, with up to 27% of all $CD8^+$ lymphocytes recognizing a single epitope (two overlapping peptides with the same peptide core).

Initial analysis of SARS-CoV-2–unexposed individuals revealed substantial presence of $CD4^+$ and $CD8^+$ T cells cross-reactive to SARS-CoV-2 peptides (*11*, *13*, *15*, *17*, *19*, *36*, *37*). Longitudinal analysis

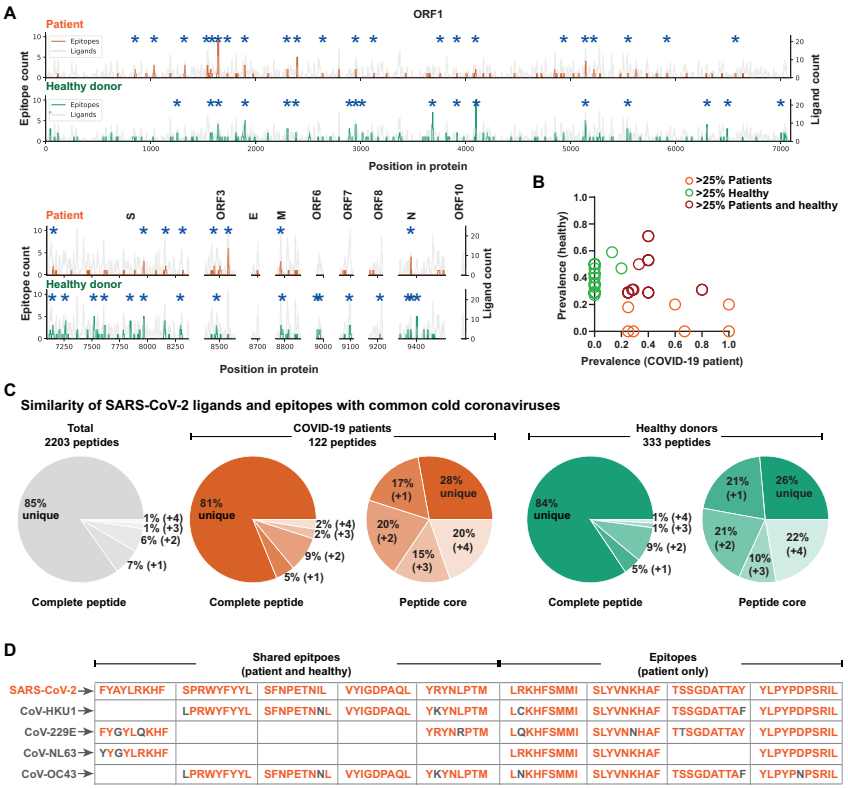


Fig. 5. A fraction of SARS-CoV-2 epitopes share sequence homology with widely circulating common cold coronaviruses. (**A**) SARS-CoV-2 T cell immunogenicity map across the viral proteome comparing the distribution of identified SARS-CoV-2 epitopes (patient cohort, orange line; $n = 16$ patients) with the total analyzed peptides (gray line). The height of a peak indicates the number of ligands (right *y* axis) analyzed in a particular region and the number of identified epitopes (left *y* axis). The bottom panel similarly maps epitopes and ligands from healthy donors (green line, $n = 31$ individuals). Positions significantly enriched ($P < 0.05$) with epitopes compared with the number of tested ligands are marked with an asterisk. (**B**) T cell epitopes selected on the basis of their immunodominant characteristics either in the patient (orange) or healthy donor (green) cohort or represented in both (red) are evaluated for their T cell recognition prevalence in both cohorts. (**C**) Sequence similarity of SARS-CoV-2 peptides with the other four common cold coronaviruses (HCoV) HCoV-HKU1, HCoV-NL63, and HCoV-229E. The gray pie chart indicates the sequence similarity of the total predicted peptides from SARS-CoV-2 with any one (+1), two (+2), three (+3), or all four (+4) HCoV peptides with a variation limit of up to two amino acids within the full-length peptide. The colored pie chart shows a similar analysis for epitopes detected in the patient ($n = 16$) or healthy donor cohort (combined analysis of HD-1 and HD-2, $n = 31$) for full-length peptide and peptide core. (**D**) Examples of sequence homology for shared (between patient and healthy donors) and patient-specific T cell epitopes with one or more HCoV peptide sequence. Nonmatching amino acids are shown in gray.

of cross-reactive and induced CD8$^+$ T cells before and after SARS-CoV-2 infection has been followed in individual cases (*37*), but the role of preexisting T cells in overall immune response and disease outcome is not yet established. Using a genome-wide screen of expanded T cells, a recent study reported cross-reactivity to SARS-CoV epitopes in patients with COVID-19 but not to other commonly circulating coronaviruses (*38*). Our ex vivo evaluation of all 3141 SARS-CoV-2–derived minimal epitopes in two healthy cohorts (COVID-19–unexposed and high risk) shows extensive but low-frequency and low-avidity interaction with CD8$^+$ T cells. Preexisting immunity based on cross-reactive T cells can influence how the immune system reacts upon viral exposure, including through faster expansion of preexisting memory cells upon initial exposure to viral infection. A similar outcome and benefit of preexisting T cell immunity have been shown in the case of the flu pandemic virus H1N1 (*39*, *40*). However, active stimulation of cross-reactive T cells could also lead to exhaustion of rapidly expanded T cells, similar to the higher PD-1 expression and reduced cytokine secretion of the SARS-CoV-2 immunodominant T cells observed by us and others (*5*, *41*, *42*). In addition, hyperactivation of preexisting T cells could contribute to short- and long-term disease

severity via inflammation and autoimmunity, because increased production of IFN-γ by $CD4^+$ and $CD8^+$ T cells has been observed in patients with severe COVID-19 (*43*). Furthermore, it has been reported (*44*) that SARS-CoV-2 infection can be a triggering factor for autoimmune reactions and severe pneumonia with sepsis leading to acute respiratory distress syndrome, bone marrow infection with pancytopenia, and organ-specific autoimmunity (*45*–*47*). Preexisting T cell immunity can influence vaccination outcomes, because they may induce a faster but possibly selective immune response. The ORF1 protein regions are highly conserved within coronaviruses (*48*) and show the highest HCoV identity among SARS-CoV-2 proteins, and most of the immunodominant epitopes that we have identified belong to the ORF1 region. Thus, a detailed evaluation of these T cell epitopes could be of value in vaccine design.

Most vaccine development efforts are currently focusing on mounting antibody responses to the spike protein, with limited focus on T cell immunity. This is due to the receptor binding domain being the main target for neutralizing antibodies produced by B cells (*49*). However, several studies have pointed out relatively low antibody titers in COVID-19 recovered patients (*3*, *50*–*52*). In conditions where antibody titers cannot sufficiently protect against infections, T cell immunity may sustain the antibody responses and provide a direct source of T cells for clearing virus-infected cells. For the involvement of T cell immunity in vaccine development, our data suggest that the inclusion of other virus proteins, such as ORF1 or ORF3, might be highly relevant. For now, the role of antibody- and T cell–mediated immune response after natural infection or after vaccination is not yet resolved and requires extensive longitudinal analysis comparing antibody and T cell kinetics to determine a synergistic or specific effect in long-term disease protection.

T cell recognition of SARS-CoV-2–derived peptides in both patients with COVID-19 and healthy donors has prompted us to understand the role of T cell cross-reactivity in controlling infections. In recent years, technology improvements in TCR characterization have allowed us to interrogate the TCR-pMHC interaction from a structural approach while obtaining experimental information related to the peptide amino acids that are crucial to T cell recognition (*53*–*58*). Such efforts have taught us that T cell cross-recognition is very difficult to predict, without knowing the precise interaction required for the given TCR, because even T cell epitopes with as low as 40% sequence homology can be recognized by a given TCR (*30*). Therefore, the underlining source of T cell cross-reactivity might arise from a larger set of epitopes within the HCoV viruses, including sequences with larger variation than those evaluated here (i.e., maximum of two amino acid variants per peptide sequence/peptide core).

Although T cell recognition itself was largely overlapping in identity between patients and healthy donors, the magnitude of T cell responses and particularly the phenotype of SARS-CoV-2–specific T cells were substantially different. We detected a strong activation profile of SARS-CoV-2–specific T cells only in patients with COVID-19, and this strong "activation signature" (high expression of CD38, CD39, PD-1, and HLA-DR) was further enhanced in patients requiring hospitalization. Such highly activated T cell responses should facilitate viral clearance, and hence, our data further support the notion that some severely affected patients might suffer from hyperactivation of their T cell compartment as a consequence of their primary viral infection, which may even be cleared. Additional signs of functional impairment were observed, and cytokine secretion upon antigen stimulation was incomplete for the high-frequency populations of SARS-CoV-2–specific T cells.

A limitation of the current study relates to the lack of information related to the precise date of infection. This may differ by up to 1 week, because symptoms and hence diagnosis can be delayed. Consequently, differences in T cell mobilization and/or activation may be observed as a function of time, which is not controlled in the present study. However, a measurement of symptoms before the first positive SARS-CoV-2 test indicates that samples were collected at about the same time relative to symptom onset in the two groups of patients, except for three patients from the intensive care unit included later after infection. In addition, although our T cell screening strategy allows for high-throughput epitope mapping, determination of individual responses can only be estimated following the barcode deconvolution strategy, in relation to the pool of pMHC multimer$^+$ T cells upon sorting. For the healthy donor population, the separation was insufficient to precisely determine the frequency of this T cell population, whereas for the patient cohort, both measurements demonstrated strong correlation with measurements of the individual responses using conventional pMHC tetramers.

Together, COVID-19 disease drives substantial T cell activation, with T cell recognition of a large number of SARS-CoV-2–derived peptides. There is also considerable T cell recognition of such peptides in healthy donors, arguing for cross-recognition, potentially from T cells raised against other coronaviruses. The activation profile clearly distinguishes patients from healthy individuals. Patients who required hospitalization for COVID-19 demonstrated a significantly higher frequency of SARS-CoV-2–specific T cells and a more activated phenotype compared with patients with milder disease. The data presented here support a role for T cell recognition in COVID-19 and hypothesize that such T cells are associated with COVID-19 disease severity. Preexisting T cell immunity likely influences the immune response to SARS-CoV-2, which could be leveraged to fight novel infections.

MATERIALS AND METHODS

Study design

This study aimed to identify a full repertoire of $CD8^+$ T cell–mediated immune response to SARS-CoV-2 infection. For a comprehensive evaluation, we determined potential T cell epitopes within the complete SARS-CoV-2 genome and analyzed the resulting 3141 peptides for their T cell recognition, immunodominance, breadth of the T cell response, functional and phenotype of reactive T cells, and contribution in COVID-19 disease severity. We used a DNA barcode–based MHC multimer T cell detection technology in combination with a 13-parameter flow cytometry phenotyping panel for T cell identification in PBMCs in a cohort of 18 patients with COVID-19 (composed of severe and mild disease) and compared with T cell recognition in two healthy donor cohorts (18 COVID-19–unexposed individuals and 20 high-risk health care staff). To understand the association of SARS-CoV-2–specific T cells in disease severity, we included an additional 23 patients for T cell phenotype analysis.

Clinical samples

Approval for the study design and sample collection was obtained from the Committee on Health Research Ethics in the Capital Region of Denmark. All included patients and health care employees gave their informed written consent for inclusion. PBMC samples

from 18 SARS-CoV-2–infected patients were used in this study. Blood samples were collected as close as possible to the first COVID-19–positive test. The patient cohort included samples from individuals with severe symptoms who required hospital care (hospitalized; $n = 11$) and patients with mild symptoms not requiring hospital care (outpatient; $n = 7$). For hospitalized patients, we collected full data from the medical record regarding disease course, age, gender, travel history, performance status, symptoms, comorbidity, medications, laboratory findings, diagnostic imaging, treatment, need of oxygen, need for intensive care, and an overall estimate of disease severity (table S2). For outpatients, we used a questionnaire to collect data on comorbidity, travel history, medications, and performance status.

SARS-CoV-2 infection was diagnosed by one of four platforms as follows: BGI (BGI COVID-19 RT-PCR kit), Panther Fusion (Hologic), Roche Flow (Roche MagNA Pure 96 and Roche LightCycler 480 II real-time PCR), and Qiaflow (QIAsymphony or RotorGene, Qiagen). In the last three platforms, LightMix Modular SARS-CoV (COVID-19) E-gene (#53-0776-96) has been used. The diversity of platforms used was due to supply issues. All platforms were validated using validation kits and panels from the Statens Serum Institute, Denmark. Most patients had more than one positive test for COVID-19. Swabs, sputum, and tracheal secretion were used depending on the setting.

Patients were attempted for inclusion soon after diagnosis. The samples were collected within 2 weeks from COVID-19 diagnosis (except for three patients who were at intensive care after diagnosis). The average number of days with symptoms before sample collection matches closely in the two patient cohorts (10.85 days for the hospitalized group and 10.45 days for the outpatient group) (table S2); however, it was not possible to determine the exact date of infection.

For the pre–COVID-19 healthy donor cohort ($n = 18$), we used samples collected before October 2019 and obtained from the central blood bank, Rigshospitalet, Copenhagen, in an anonymized form. In addition, we included 20 health care employees from Herlev Hospital during the COVID-19 pandemic, who were at high risk of SARS-CoV-2 infection but not detected to be positive, as a cohort to follow immune responses in a potentially exposed population. PBMCs from all three cohorts were isolated immediately after sampling using Ficoll-Paque PLUS (GE Healthcare) density gradient centrifugation and were cryopreserved thereafter at a density of 2×10^6 to 20×10^6 cells/ml.

SARS-CoV-2 peptide selection

Potential HLA class I–binding peptides were predicted from the complete set of 8- to 11-mer peptides contained within the Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1 (GenBank ID: MN908947.3) to a set of 10 prevalent and functionally diverse HLA molecules (HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A24:02, HLA-B07:02, HLA-B08:01, HLA-B15:01, HLA-C06:02, HLA-C07:01, and HLA-C07:02) using a preliminary version of NetMHCpan 4.1 (www.cbs.dtu.dk/services/NetMHCpan/index_v0.php) (PMID: 32406916). For peptides predicted from ORF1 protein, a percentile rank binding threshold of 0.5% was used, and for peptides derived from all other proteins, a threshold of 1% was used. Together, 2203 peptides were selected, binding to one or more HLA molecules, generating 3141 peptide-HLA pairs for experimental evaluation (table S1). All peptides were custom-synthesized by Pepscan Presto BV, Lelystad, The Netherlands. Peptide synthesis was done at a 2-μmol scale with ultraviolet (UV) and mass spectrometry quality control analysis for 5% random peptides with an estimated purity of 70 to 92% by the supplier.

MHC-I monomer production

All 10 MHC-I monomer types were produced using methods previously described (*59*). Briefly, MHC-I heavy chain and human β_2-microglobulin (hβ2m) were expressed in *Escherichia coli* using pET series expression plasmids. Soluble denatured proteins of the heavy chain and hβ2m were harvested using inclusion body preparation. The folding of these molecules was initiated in the presence of UV-labile HLA-specific peptide ligands (*60*). HLA-A02:01 and A24:02 molecules were folded and purified empty, as described previously (*61*). Folded MHC-I molecules were biotinylated using the BirA biotin-protein ligase standard reaction kit (Avidity LLC, Aurora, CO), and MHC-I monomers were purified using size exclusion chromatography (HPLC, Waters Corporation, USA). All MHC-I folded monomers were quality-controlled for their concentration, UV degradation, and biotinylation efficiency and stored at −80°C until further use.

DNA-barcoded multimer library preparation

The DNA-barcoded multimer library was prepared using the method developed by Bentzen *et al.* (*25*). Unique barcodes were generated by combining different A and B oligos, with each barcode representing a 5′ biotinylated unique DNA sequence. These barcodes were attached to PE or APC and streptavidin-conjugated dextran (Fina Biosolutions, Rockville, MD, USA) by incubating them at 4°C for 30 min to generate a DNA barcode dextran library of 1325 unique barcode specificities. SARS-CoV-2 pMHC libraries were generated by incubating 200 μM peptide of each peptide with 100 μg/ml of the respective MHC molecules for 1 hour using UV-mediated peptide exchange (HLA-A01:01, A03:01, B07:02, B08:01, B15:01, C06:02, C07:01, and C07:02) or direct binding to empty MHC molecules (HLA-A02:01 and A24:02). HLA-specific DNA-barcoded multimer libraries were then generated by incubating pMHC monomers to their corresponding DNA barcode–labeled dextrans at 4°C for 30 min, thus providing a DNA barcode–labeled pMHC multimer specifically to probe the respective T cell population. A similar process was followed to generate DNA-barcoded pMHC multimers for CEF epitopes (HLA-A and HLA-B) using APC- and streptavidin-conjugated dextran attached with unique barcodes.

T cell staining with DNA-barcoded pMHC multimers and phenotype panel

All COVID-19 patient and healthy donor samples were HLA-genotyped for HLA-A, HLA-B, and HLA-C loci (next-generation sequencing; IMGM Laboratories GmbH, Germany) (table S10). Patient and healthy donor HLA-matching SARS-CoV-2 and CEF pMHC multimer libraries were pooled [as described previously (*25*)] and incubated with 5×10^6 to 10×10^6 PBMCs [thawed and washed twice in RPMI and 10% fetal calf serum (FCS) and washed once in barcode cytometry buffer] for 15 min at 37°C at a final volume of 60 μl. Cells were then mixed with 40 μl of phenotype panel containing surface marker antibodies (table S8) and a dead cell marker (LIVE/DEAD Fixable Near-IR; Invitrogen, L10119) (final dilution 1/1000) and incubated at 4°C for 30 min. Cells were washed twice with barcode cytometry buffer and fixed in 1% paraformaldehyde.

Cells fixed after staining with pMHC multimers were acquired on a FACSAria flow cytometer instrument (AriaFusion, Becton

Dickinson), and all the PE-positive (SARS-CoV-2 multimer binding) and APC-positive (CEF multimer binding) cells of $CD8^+$ gate were sorted into presaturated tubes (2% bovine serum albumin and 100 μl of barcode cytometry buffer) (fig. S11A). Sorted cells belonging to each sample were then subjected to polymerase chain reaction (PCR) amplification of its associated DNA barcode(s). Cells were centrifuged for 10 min at 5000*g*, and the supernatant was discarded with minimal residual volume. The remaining pellet was used as the PCR template for each of the sorted samples and amplified using the Taq PCR Master Mix Kit (Qiagen, 201443) and the sample-specific forward primer (serving as sample identifier) A-key36. PCR-amplified DNA barcodes were purified using the QIAquick PCR Purification kit (Qiagen, 28104) and sequenced at PrimBio (USA) using the Ion Torrent PGM 314 or 316 chip (Life Technologies).

DNA barcode sequence analysis and identification of pMHC specificities

To process the sequencing data and automatically identify the barcode sequences, we designed a specific software package, "Barracoda" (https://services.healthtech.dtu.dk/service.php?Barracoda-1.8). This software tool identifies the barcodes used in a given experiment, assigns sample ID and pMHC specificity to each barcode, and calculates the total number of reads and clonally reduced reads for each pMHC-associated DNA barcode. Furthermore, it includes statistical processing of the data. Details are given in (*25*). The analysis of barcode enrichment was based on methods designed for the analysis of RNA sequencing data and was implemented in the R package edgeR. Fold changes in read counts mapped to a given sample relative to mean read counts mapped to triplicate baseline samples were estimated using normalization factors determined by the trimmed mean of *M* values. *P* values were calculated by comparing each experiment individually to the mean baseline sample reads using a negative binomial distribution with a fixed dispersion parameter set to 0.1 (*25*). False discovery rates (FDRs) were estimated using the Benjamini-Hochberg method. Specific barcodes with FDR < 0.1% were defined as significant, determining T cell recognition in the given sample. At least 1 per 1000 reads associated with a given DNA barcode relative to the total number of DNA barcode reads in that given sample was set as the threshold to avoid false-positive detection of T cell populations due to the low number of reads in the baseline samples. T cell frequency associated with each significantly enriched barcode was measured on the basis of the percentage read count of the associated barcode out of the total percentage $multimer^+$ $CD8^+$ T cell population in patient samples. In healthy donors, T cell recognition was identified on the basis of barcode enrichment analysis, the same as in patient samples; however, a frequency estimate of the corresponding T cell populations was not determined for significant responses identified in healthy donors because of insufficient separation of $multimer^+$ cells. To exclude potential pMHC elements binding to T cells in a nonspecific fashion, non–HLA-matching healthy donor material was included as a negative control. Any T cell recognition determined in these samples was subtracted from the full dataset.

T cell expansion and combinatorial tetramer staining

PBMCs from healthy donors were expanded in vitro using pMHC-dextran complexes conjugated with SARS-CoV-2–derived peptides and cytokines (IL-2 and IL-21) for 2 weeks either with single pMHC specificity or with a pool of up to 10 pMHC specificities. PBMCs were expanded for 2 weeks in X-VIVO Media (Lonza, BE02-060Q) supplemented with 5% human serum (Gibco, 1027-106). Expanded cells were used to measure peptide-specific T cell activation or stained using pMHC tetramers to detect T cells recognizing SARS-CoV-2 epitopes.

In vitro expanded healthy donor PBMCs were examined for SARS-CoV-2–reactive T cells using combinatorial tetramer staining (*62*). Individual HLA-restricted pMHC complexes were generated using direct peptide loading (HLA-A02:01 and A24:02) or UV-mediated peptide exchange (all other HLAs) as described above and conjugated with fluorophore-labeled streptavidin molecules. For 100 μl of pMHC monomers, 9.02 μl [0.2 mg/ml of stock; SA-PE-CF594 (streptavidin-PE/CF594; BD Biosciences, 562318) and SA-APC (BioLegend, 405207)] or 18.04 μl [0.1 mg/ml of stock; SA-BUV395 (Brilliant Ultraviolet 395; BD Biosciences, 564176), SA-BV421 (Brilliant Violet 421; BD Biosciences, 563259), and SA-BV605 (Brilliant Violet 605; BD Biosciences, 563260)] of streptavidin conjugates was added and incubated for 30 min at 4°C, followed by the addition of D-biotin (Sigma-Aldrich) at 25 μM final concentration to block any free binding site. pMHC tetramers for each specificity were generated in two colors by incubating pMHC monomers and mixed in a 1:1 ratio before staining the cells. Expanded cells were stained with 1 μl of pooled pMHC multimers per specificity (in combinatorial encoding) by incubating 1×10^6 to 5×10^6 cells for 15 min at 37°C in 80 μl of total volume. Cells were then mixed with 20 μl of antibody staining solution CD8-BV480 (BD Biosciences, B566121) (final dilution 1/50), dump channel antibodies [CD4-FITC (BD Biosciences, 345768) (final dilution 1/80), CD14-FITC (BD Biosciences, 345784) (final dilution 1/32), CD19-FITC (BD Biosciences, 345776) (final dilution 1/16), CD40-FITC (Serotech, MCA1590F) (final dilution 1/40), and CD16-FITC (BD Biosciences, 335035) (final dilution 1/64)], and a dead cell marker (LIVE/DEAD Fixable Near-IR; Invitrogen, L10119) (final dilution 1/1000) and incubated for 30 min at 4°C. Cells were then washed twice in fluorescence-activated cell sorting buffer (phosphate-buffered saline and 2% FCS) and acquired on a flow cytometer (Fortessa, Becton Dickinson). Data were analyzed using FlowJo analysis software.

T cell functional analysis

For functional evaluation of T cells from PBMCs of patients with COVID-19 or PBMCs expanded from healthy donors, 1×10^6 to 2×10^6 cells were incubated with 1 μM SARS-CoV-2 minimal epitope or pool of up to 10 epitopes (1 μM per peptide) for 9 hours at 37°C in the presence of protein transport inhibitor (final dilution 1/1000; GolgiPlug; BD Biosciences, 555029). Functional activation of T cells was measured using intracellular cytokines IFN-γ (final dilution 1/20; BD Biosciences, 341117) and TNF-α (final dilution 1/20; BioLegend, 502930). Cells incubated with Leukocyte Activation Cocktail (final dilution 1/500; BD Biosciences, 550583) were used as a positive control, and HLA-specific irrelevant peptides were used as negative controls. Surface marker antibodies CD3-FITC (final dilution 1/20; BD Biosciences, 345764), CD4-BUV395 (final dilution 1/300; BD Biosciences, 742738), and CD8-BV480 (final dilution 1/50; BD Biosciences, B566121) and dead cell marker (final dilution 1/1000; LIVE/DEAD Fixable Near-IR; Invitrogen, L10119) were used to identify $CD8^+$ T cells producing intracellular cytokines (gating strategy; fig. S5A).

Flow cytometry analysis

For phenotype analysis, all samples were analyzed using FlowJo data analysis software (FlowJo LLC). Frequencies of specific cell populations were calculated according to the gating strategy shown in fig. S11B. For combinatorial tetramer staining, T cell binding to specific pMHC tetramers was identified using the gating plan described in the original study (*63*). For UMAP analysis (*64*), FCS (Flow Cytometry Standard) files of samples from the patient and healthy cohorts were concatenated (160,000 total cells), downsampled (FlowJo plugin), and visualized using UMAP (version 2.2, FlowJo plugin) analysis based on the following selected markers: CD3, CD4, CD8, CD38, CD39, CD69, CD137, HLA-DR, PD-1, CCR7, CD45RA, CD27, and CD57.

Sequence homology analyses

To evaluate the homology between SARS-CoV-2 and HCoV, both epitopes (peptides recognized by T cells) and ligands (peptide not recognized by T cells) were mapped to their respective source protein from the SARS-CoV-2 proteome. Enrichment analysis of the epitopes in the given region of the proteins was based on testing whether the count of observed epitopes exceeded what we expected from the number of ligands tested at each position. Epitopes were considered successes, and the count of ligands was regarded as the number of trials in a binomial test. The probability of success was derived from the average ratio of epitope to ligand per position across each protein. The test was "one-sided" with a significance level at 0.05.

The similarity of SARS-CoV-2 ligands and epitopes from both patient and healthy donor cohorts to a set of human common cold corona viruses (HCoV-HKU1, HCoV-229E, HCoV-NL63, and HCoV-OC43) was tested using two methods. The first approach used a kernel method for amino acid sequences using BLOSUM62 (*65*). The second approach was a simple string search allowing up to two mismatches. On the basis of the second approach, each epitope was categorized by how many, if any, of the common cold viruses it would match with. Both methods were applied to the full peptide length and to the peptide core.

Data processing and statistics

T cell recognition was determined on the basis of the DNA-barcoded pMHC multimer analysis and evaluated through Barracoda (see above). The data were plotted using Python 3.7.4. For all plots, peptide sequences with no significant enrichments are shown as gray dots, and all peptides with a negative enrichment are set to LogFc equal zero (Figs. 1, D, E, and G, and 3, A and B, and fig. S2). Box plots for data quantification and visualization were generated, and their related statistical analyses were performed using GraphPad Prism (GraphPad Software Inc.) (Figs. 3C and 4, A to F, and figs. S1, A and B, S7B, S8, B and C, and S9A) or R studio (fig. S10). For unpaired comparisons, Mann-Whitney test was applied, and to compare more than two groups, one-way analysis of variance (ANOVA) (Kruskal-Wallis) test was performed using GraphPad Prism. All *P* values are indicated in the figure legends. Flow cytometry data were analyzed using FlowJo (version 10). Immunogenicity scores (Fig. 1, F and H, and fig. S3) were calculated (as percentage) by dividing the total identified T cell reactivity associated with an HLA or protein with the total number of specificities analyzed in a given cohort (number of peptides multiplied by the number of patient with a given HLA). Staining index (Fig. 3C) was calculated as follows: [mean fluorescence intensity (MFI) of multimer$^+$ cells – MFI of multimer$^-$ cells]/(2 × SD of multimer$^-$ cells). MFI of multimer$^+$ and multimer$^-$ CD8$^+$ T cells and the SD of the multimer$^-$ CD8$^+$ T cells are from FlowJo analysis for patient and healthy donor samples.

SUPPLEMENTARY MATERIALS

immunology.sciencemag.org/cgi/content/full/6/58/eabf7550/DC1
Figs. S1 to S11
Tables S1 to S11

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. Coronavirus disease (COVID-19) (2021); www.who.int/emergencies/diseases/novel-coronavirus-2019.
2. P. F. Cañete, C. G. Vinuesa, COVID-19 makes B cells forget, but T cells remember. *Cell* **183**, 13–15 (2020).
3. N. Vabret, Antibody responses to SARS-CoV-2 short-lived. *Nat. Rev. Immunol.* **20**, 519 (2020).
4. J. Seow, C. Graham, B. Merrick, S. Acors, S. Pickering, K. J. A. Steel, O. Hemmings, A. O'Byrne, N. Kouphou, R. P. Galao, G. Betancor, H. D. Wilson, A. W. Signell, H. Winstone, C. Kerridge, I. Huettner, J. M. Jimenez-Guardeño, M. J. Lista, N. Temperton, L. B. Snell, K. Bisnauthsing, A. Moore, A. Green, L. Martinez, B. Stokes, J. Honey, A. Izquierdo-Barras, G. Arbane, A. Patel, M. K. I. Tan, L. O'Connell, G. O'Hara, E. MacMahon, S. Douthwaite, G. Nebbia, R. Batra, R. Martinez-Nunez, M. Shankar-Hari, J. D. Edgeworth, S. J. D. Neil, M. H. Malim, K. J. Doores, Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat. Microbiol.* **5**, 1598–1607 (2020).
5. A. Bonifacius, S. Tischer-Zimmermann, A. C. Dragon, D. Gussarow, A. Vogel, U. Krettek, N. Gödecke, M. Yilmaz, A. R. M. Kraft, M. M. Hoeper, I. Pink, J. J. Schmidt, Y. Li, T. Welte, B. Maecker-Kolhoff, J. Martens, M. M. Berger, C. Lobenwein, M. V. Stankov, M. Cornberg, S. David, G. M. N. Behrens, O. Witzke, R. Blaszczyk, B. Eiz-Vesper, COVID-19 immune signatures reveal stable antiviral T cell function despite declining humoral responses. *Immunity* **54**, 340–354.e6 (2021).
6. H. Peng, L.-t. Yang, L.-y. Wang, J. Li, J. Huang, Z.-q. Lu, R. A. Koup, R. T. Bailer, C.-y. Wu, Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein in SARS-recovered patients. *Virology* **351**, 466–475 (2006).
7. F. Tang, Y. Quan, Z.-T. Xin, J. Wrammert, M.-J. Ma, H. Lv, T.-B. Wang, H. Yang, J. H. Richardus, W. Liu, W.-C. Cao, Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: A six-year follow-up study. *J. Immunol.* **186**, 7264–7268 (2011).
8. D. Weiskopf, K. S. Schmitz, M. P. Raadsen, A. Grifoni, N. M. A. Okba, H. Endeman, J. P. C. van den Akker, R. Molenkamp, M. P. G. Koopmans, E. C. M. van Gorp, B. L. Haagmans, R. L. de Swart, A. Sette, R. D. de Vries, Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. *Sci. Immunol.* **5**, eabd2071 (2020).
9. L. Ni, F. Ye, M. L. Cheng, Y. Feng, Y. Q. Deng, H. Zhao, P. Wei, J. Ge, M. Gou, X. Li, L. Sun, T. Cao, P. Wang, C. Zhou, R. Zhang, P. Liang, H. Guo, X. Wang, C. F. Qin, F. Chen, C. Dong, Detection of SARS-CoV-2-specific humoral and cellular immunity in COVID-19 convalescent individuals. *Immunity* **52**, 971–977.e3 (2020).
10. Y. Peng, A. J. Mentzer, G. Liu, X. Yao, Z. Yin, D. Dong, W. Dejnirattisai, T. Rostron, P. Supasa, C. Liu, C. López-Camacho, J. Slon-Campos, Y. Zhao, D. I. Stuart, G. C. Paesen, J. M. Grimes, A. A. Antson, O. W. Bayfield, D. E. D. P. Hawkins, D. S. Ker, B. Wang, L. Turtle, K. Subramaniam, P. Thomson, P. Zhang, C. Dold, J. Ratcliff, P. Simmonds, T. de Silva, P. Sopp, D. Wellington, U. Rajapaksa, Y. L. Chen, M. Salio, G. Napolitani, W. Paes, P. Borrow, B. M. Kessler, J. W. Fry, N. F. Schwabe, M. G. Semple, J. K. Baillie, S. C. Moore, P. J. M. Openshaw, M. A. Ansari, S. Dunachie, E. Barnes, J. Frater, G. Kerr, P. Goulder, T. Lockett, R. Levin, Y. Zhang, R. Jing, L. P. Ho, E. Barnes, D. Dong, T. Dong, S. Dunachie, J. Frater, P. Goulder, G. Kerr, P. Klenerman, G. Liu, A. McMichael, G. Napolitani, G. Ogg, Y. Peng, M. Salio, X. Yao, Z. Yin, J. Kenneth Baillie, P. Klenerman, A. J. Mentzer, S. C. Moore, P. J. M. Openshaw, M. G. Semple, D. I. Stuart, L. Turtle, R. J. Cornall, C. P. Conlon, P. Klenerman, G. R. Screaton, J. Mongkolsapaya, A. McMichael, J. C. Knight, G. Ogg, T. Dong, Broad and strong memory CD4$^+$ and CD8$^+$ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. *Nat. Immunol.* **21**, 1336–1345 (2020).
11. T. Sekine, A. Perez-Potti, O. Rivera-Ballesteros, K. Strålin, J.-B. Gorin, A. Olsson, S. Llewellyn-Lacey, H. Kamal, G. Bogdanovic, S. Muschiol, D. J. Wullimann, T. Kammann, J. Emgård, T. Parrot, E. Folkesson, O. Rooyackers, L. I. Eriksson, J.-I. Henter, A. Sönnerborg, T. Allander, J. Albert, M. Nielsen, J. Klingström, S. Gredmark-Russ, N. K. Björkström, J. K. Sandberg, D. A. Price, H.-G. Ljunggren, S. Aleman, M. Buggert, Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *Cell* **183**, 158–168.e14 (2020).

12. B. J. Meckiff, C. Ramírez-Suástegui, V. Fajardo, S. J. Chee, A. Kusnadi, H. Simon, A. Grifoni, E. Pelosi, D. Weiskopf, A. Sette, F. Ay, G. Seumois, C. H. Ottensmeier, P. Vijayanand, Single-cell transcriptomic analysis of SARS-CoV-2 reactive CD4$^+$ T cells. *bioRxiv* 2020.06.12.148916, (2020).
13. N. Le Bert, A. T. Tan, K. Kunasegaran, C. Y. L. Tham, M. Hafezi, A. Chia, M. H. Y. Chng, M. Lin, N. Tan, M. Linster, W. N. Chia, M. I. C. Chen, L. F. Wang, E. E. Ooi, S. Kalimuddin, P. A. Tambyah, J. G. H. Low, Y. J. Tan, A. Bertoletti, SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* **584**, 457–462 (2020).
14. A. Grifoni, D. Weiskopf, S. I. Ramirez, J. Mateus, J. M. Dan, C. R. Moderbacher, S. A. Rawlings, A. Sutherland, L. Premkumar, R. S. Jadi, D. Marrama, A. M. de Silva, A. Frazier, A. F. Carlin, J. A. Greenbaum, B. Peters, F. Krammer, D. M. Smith, S. Crotty, A. Sette, Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* **181**, 1489–1501.e15 (2020).
15. A. Nelde, T. Bilich, J. S. Heitmann, Y. Maringer, H. R. Salih, M. Roerden, M. Lübke, J. Bauer, J. Rieth, M. Wacker, A. Peter, S. Hörber, B. Traenkle, P. D. Kaiser, U. Rothbauer, M. Becker, D. Junker, G. Krause, M. Strengert, N. Schneiderhan-Marra, M. F. Templin, T. O. Joos, D. J. Kowalewski, V. Stos-Zweifel, M. Fehr, A. Rabsteyn, V. Mirakaj, J. Karbach, E. Jäger, M. Graf, L.-C. Gruber, D. Rachfalski, B. Preuß, I. Hagelstein, M. Märklin, T. Bakchoul, C. Gouttefangeas, O. Kohlbacher, R. Klein, S. Stevanović, H.-G. Rammensee, J. S. Walz, SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat. Immunol.* **22**, 74–85 (2021).
16. S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**, 860–868 (2020).
17. J. Braun, L. Loyal, M. Frentsch, D. Wendisch, P. Georg, F. Kurth, S. Hippenstiel, M. Dingeldey, B. Kruse, F. Fauchere, E. Baysal, M. Mangold, L. Henze, R. Lauster, M. A. Mall, K. Beyer, J. Röhmel, S. Voigt, J. Schmitz, S. Miltenyi, I. Demuth, M. A. Müller, A. Hocke, M. Witzenrath, N. Suttorp, F. Kern, U. Reimer, H. Wenschuh, C. Drosten, V. M. Corman, C. Giesecke-Thiel, L. E. Sander, A. Thiel, SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* , 587, 270–274 (2020).
18. U. Sterbo, S. Rahmann, T. Roch, T. H. Westhof, N. Babel, SARS-CoV-2 reactive T cells in uninfected individuals are likely expanded by beta-coronaviruses. *bioRxiv* 2020.07.01.182741, (2020).
19. J. Mateus, A. Grifoni, A. Tarke, J. Sidney, S. I. Ramirez, J. M. Dan, Z. C. Burger, S. A. Rawlings, D. M. Smith, E. Phillips, S. Mallal, M. Lammers, P. Rubiro, L. Quiambao, A. Sutherland, E. D. Yu, R. da Silva Antunes, J. Greenbaum, A. Frazier, A. J. Markmann, L. Premkumar, A. de Silva, B. Peters, S. Crotty, A. Sette, D. Weiskopf, Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* **370**, 89–94 (2020).
20. F. P. Havers, C. Reed, T. Lim, J. M. Montgomery, J. D. Klena, A. J. Hall, A. M. Fry, D. L. Cannon, C. F. Chiang, A. Gibbons, I. Krapiunaya, M. Morales-Betoulle, K. Roguski, M. A. U. Rasheed, B. Freeman, S. Lester, L. Mills, D. S. Carroll, S. M. Owen, J. A. Johnson, V. Semenova, C. Blackmore, D. Blog, S. J. Chai, A. Dunn, J. Hand, S. Jain, S. Lindquist, R. Lynfield, S. Pritchard, T. Sokol, L. Sosa, G. Turabelidze, S. M. Watkins, J. Wiesman, R. W. Williams, S. Yendell, J. Schiffer, N. J. Thornburg, Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23-May 12, 2020. *JAMA Intern. Med.* **180**, 1576–1586 (2020).
21. D. P. Oran, E. J. Topol, Prevalence of asymptomatic SARS-CoV-2 infection. *Ann. Intern. Med.* **173**, 362–367 (2020).
22. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
23. J. B. Moore, C. H. June, Cytokine release syndrome in severe COVID-19. *Science* **368**, 473–474 (2020).
24. C. Zhang, Z. Wu, J. W. Li, H. Zhao, G. Q. Wang, Cytokine release syndrome in severe COVID-19: Interleukin-6 receptor antagonist tocilizumab may be the key to reduce mortality. *Int. J. Antimicrob. Agents* **55**, 105954 (2020).
25. A. K. Bentzen, A. M. Marquard, R. Lyngaa, S. K. Saini, S. Ramskov, M. Donia, L. Such, A. J. S. Furness, N. McGranahan, R. Rosenthal, P. T. Straten, Z. Szallasi, I. M. Svane, C. Swanton, S. A. Quezada, S. N. Jakobsen, A. C. Eklund, S. R. Hadrup, Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* **34**, 1037–1045 (2016).
26. F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, Y. Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
27. B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
28. J. J. A. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Keşmir, B. Peters, Properties of MHC class I presented peptides that enhance immunogenicity. *PLOS Comput. Biol.* **9**, e1003266 (2013).
29. S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, C. Kesmir, Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLOS ONE* **3**, e1831 (2008).
30. A. K. Bentzen, L. Such, K. K. Jensen, A. M. Marquard, L. E. Jessen, N. J. Miller, C. D. Church, R. Lyngaa, D. M. Koelle, J. C. Becker, C. Linnemann, T. N. M. Schumacher, P. Marcatili, P. Nghiem, M. Nielsen, S. R. Hadrup, T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nat. Biotechno.* 10.1038/nbt.4303 , (2018).
31. A. S. Shomuradova, M. S. Vagida, S. A. Sheetikov, K. V. Zornikova, D. Kiryukhin, A. Titov, I. O. Peshkova, A. Khmelevskaya, D. V. Dianov, M. Malasheva, A. Shmelev, Y. Serdyuk, D. V. Bagaev, A. Pivnyuk, D. S. Shcherbinin, A. V. Maleeva, N. T. Shakirova, A. Pilunov, D. B. Malko, E. G. Khamaganova, B. Biderman, A. Ivanov, M. Shugay, G. A. Efimov, SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T cell receptors. *Immunity* **53**, 1245–1257.e5 (2020).
32. A. Alison Tarke, J. Sidney, C. K. Kidd, D. Weiskopf, A. Grifoni, A. Sette, A. Tarke, J. M. Dan, S. I. Ramirez, E. Dawen Yu, J. Mateus, R. da Silva Antunes, E. Moore, P. Rubiro, N. Methot, E. Phillips, S. Mallal, A. Frazier, S. A. Rawlings, J. A. Greenbaum, B. Peters, D. M. Smith, S. Crotty, D. Weiskopf, A. Grifoni, A. Sette, Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Rep. Med.* **2**, 100204 (2021).
33. H. Kared, A. D. Redd, E. M. Bloch, T. S. Bonny, H. Sumatoh, F. Kairi, D. Carbajo, B. Abel, E. W. Newell, M. P. Bettinotti, S. E. Benner, E. U. Patel, K. Littlefield, O. Laeyendecker, S. Shoham, D. Sullivan, A. Casadevall, A. Pekosz, A. Nardin, M. Fehlings, A. A. R. Tobian, T. C. Quinn, SARS-CoV-2-specific CD8$^+$ T cell responses in convalescent COVID-19 individuals. *J. Clin. Invest.* **131**, e145476 (2021).
34. A. Poran, D. Harjanto, M. Malloy, C. M. Arieta, D. A. Rothenberg, D. Lenkala, M. M. Van Buuren, T. A. Addona, M. S. Rooney, L. Srinivasan, R. B. Gaynor, Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. *Genome Med.* **12**, 70 (2020).
35. M.-S. Rha, H. W. Jeong, J.-H. Ko, Y. Choi, K. R. Peck, E.-C. Shin, S. J. Choi, I.-H. Seo, J. S. Lee, M. Sa, A. R. Kim, E.-J. Joo, J. Y. Ahn, J. H. Kim, K.-H. Song, E. S. Kim, D. H. Oh, M. Y. Ahn, H. K. Choi, J. H. Jeon, J.-P. Choi, H. Bin Kim, Y. K. Kim, S.-H. Park, W. S. Choi, J. Y. Choi, K. R. Peck, E.-C. Shin, PD-1-expressing SARS-CoV-2-specific CD8$^+$ T cells are not exhausted, but functional in patients with COVID-19. *Immunity* **54**, 44–52.e3 (2021).
36. K. W. Ng, N. Faulkner, G. H. Cornish, A. Rosa, R. Harvey, S. Hussain, R. Ulferts, C. Earl, A. G. Wrobel, D. J. Benton, C. Roustan, W. Bolland, R. Thompson, A. Agua-Doce, P. Hobson, J. Heaney, H. Rickman, S. Paraskevopoulou, C. F. Houlihan, K. Thomson, E. Sanchez, G. Y. Shin, M. J. Spyer, D. Joshi, N. O'Reilly, P. A. Walker, S. Kjaer, A. Riddell, C. Moore, B. R. Jebson, M. Wilkinson, L. R. Marshall, E. C. Rosser, A. Radziszewska, H. Peckham, C. Ciurtin, L. R. Wedderburn, R. Beale, C. Swanton, S. Gandhi, B. Stockinger, J. McCauley, S. J. Gamblin, L. E. McCoy, P. Cherepanov, E. Nastouli, G. Kassiotis, Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science* **370**, 1339–1343 (2020).
37. I. Schulien, J. Kemming, V. Oberhardt, K. Wild, L. M. Seidel, S. Killmer, Sagar, F. Daul, M. S. Lago, A. Decker, H. Luxenburger, B. Binder, D. Bettinger, O. Sogukpinar, S. Rieg, M. Panning, D. Huzly, M. Schwemmle, G. Kochs, C. F. Waller, A. Nieters, D. Duerschmied, F. Emmerich, H. E. Mei, A. R. Schulz, S. Llewellyn-Lacey, D. A. Price, T. Boettler, B. Bengsch, R. Thimme, M. Hofmann, C. Neumann-Haefelin, Characterization of pre-existing and induced SARS-CoV-2-specific CD8$^+$ T cells. *Nat. Med.* **27**, 78–85 (2021).
38. A. P. Ferretti, T. Kula, Y. Wang, S. A. Bertino, S. Chattopadhyay, G. MacBeath, Unbiased screens show CD8$^+$ T cells of COVID-19 patients recognize shared epitopes in SARS-CoV-2 that largely reside outside the spike protein. *Immunity* **53**, 1095–1107.e3 (2020).
39. T. M. Wilkinson, C. K. F. Li, C. S. C. Chui, A. K. Y. Huang, M. Perkins, J. C. Liebner, R. Lambkin-Williams, A. Gilbert, J. Oxford, B. Nicholas, K. J. Staples, T. Dong, D. C. Douek, A. J. McMichael, X.-N. Xu, Preexisting influenza-specific CD4$^+$ T cells correlate with disease protection against influenza challenge in humans. *Nat. Med.* **18**, 274–280 (2012).
40. S. Sridhar, S. Begom, A. Bermingham, K. Hoschler, W. Adamson, W. Carman, T. Bean, W. Barclay, J. J. Deeks, A. Lalvani, Cellular immune correlates of protection against symptomatic pandemic influenza. *Nat. Med.* **19**, 1305–1312 (2013).
41. A. Kusnadi, C. Ramírez-Suástegui, V. Fajardo, S. J. Chee, B. J. Meckiff, H. Simon, E. Pelosi, G. Seumois, F. Ay, P. Vijayanand, C. H. Ottensmeier, Severely ill COVID-19 patients display impaired exhaustion features in SARS-CoV-2-reactive CD8$^+$ T cells. *Sci. Immunol.* **6**, eabe4782 (2021).
42. J. R. Habel, T. H. O. Nguyen, C. E. van de Sandt, J. A. Juno, P. Chaurasia, K. Wragg, M. Koutsakos, L. Hensen, X. Jia, B. Chua, W. Zhang, H. X. Tan, K. L. Flanagan, D. L. Doolan, J. Torresi, W. Chen, L. M. Wakim, A. C. Cheng, P. C. Doherty, J. Petersen, J. Rossjohn, A. K. Wheatley, S. J. Kent, L. C. Rowntree, K. Kedzierska, Suboptimal SARS-CoV-2–specific CD8$^+$ T cell response associated with the prominent HLA-A*02:01 phenotype. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24384–24391 (2020).

43. F. Wang, H. Hou, Y. Luo, G. Tang, S. Wu, M. Huang, W. Liu, Y. Zhu, Q. Lin, L. Mao, M. Fang, H. Zhang, Z. Sun, The laboratory tests and host immunity of COVID-19 patients with different severity of illness. *JCI Insight* **5**, e137799 (2020).
44. M. Ehrenfeld, A. Tincani, L. Andreoli, M. Cattalini, A. Greenbaum, D. Kanduc, J. Alijotas-Reig, V. Zinserling, N. Semenova, H. Amital, Y. Shoenfeld, Covid-19 and autoimmunity. *Autoimmun. Rev.* **19**, 102597 (2020).
45. D. Gagiannis, J. Steinestel, C. Hackenbroch, M. Hannemann, V. G. Umathum, N. Gebauer, M. Stahl, H. M. Witte, K. Steinestel, COVID-19-induced acute respiratory failure – an exacerbation of organ-specific autoimmunity? *medRxiv* 2020.04.27.20077180, (2020).
46. L. A. Henderson, S. W. Canna, G. S. Schulert, S. Volpi, P. Y. Lee, K. F. Kernan, R. Caricchio, S. Mahmud, M. M. Hazen, O. Halyabar, K. J. Hoyt, J. Han, A. A. Grom, M. Gattorno, A. Ravelli, F. Benedetti, E. M. Behrens, R. Q. Cron, P. A. Nigrovic, On the alert for cytokine storm: Immunopathology in COVID-19. *Arthritis Rheumatol.* **72**, 1059–1063 (2020).
47. D. S. Hersby, T. H. Do, A. O. Gang, T. H. Nielsen, COVID-19-associated pancytopenia can be self-limiting and does not necessarily warrant bone marrow biopsy for the purposes of Sars-CoV-2 diagnostics. *Ann. Oncol.* **32**, 121–123 (2021).
48. J. Cui, F. Li, Z. L. Shi, Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
49. A. Sette, S. Crotty, Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell* **184**, 861–880 (2021).
50. J. M. Dan, J. Mateus, Y. Kato, K. M. Hastie, E. Dawen Yu, C. E. Faliti, A. Grifoni, S. I. Ramirez, S. Haupt, A. Frazier, C. Nakao, V. Rayaprolu, S. A. Rawlings, B. Peters, F. Krammer, V. Simon, E. Ollmann Saphire, D. M. Smith, D. Weiskopf, A. Sette, S. Crotty, Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**, eabf4063 (2021).
51. D. F. Robbiani, C. Gaebler, F. Muecksch, J. C. C. Lorenzi, Z. Wang, A. Cho, M. Agudelo, C. O. Barnes, A. Gazumyan, S. Finkin, T. Hägglöf, T. Y. Oliveira, C. Viant, A. Hurley, H. H. Hoffmann, K. G. Millard, R. G. Kost, M. Cipolla, K. Gordon, F. Bianchini, S. T. Chen, V. Ramos, R. Patel, J. Dizon, I. Shimeliovich, P. Mendoza, H. Hartweger, L. Nogueira, M. Pack, J. Horowitz, F. Schmidt, Y. Weisblum, E. Michailidis, A. W. Ashbrook, E. Waltari, J. E. Pak, K. E. Huey-Tubman, N. Koranda, P. R. Hoffman, A. P. West, C. M. Rice, T. Hatziioannou, P. J. Bjorkman, P. D. Bieniasz, M. Caskey, M. C. Nussenzweig, Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020).
52. A. Wajnberg, F. Amanat, A. Firpo, D. R. Altman, M. J. Bailey, M. Mansour, M. Mcmahon, P. Meade, D. R. Mendu, K. Muellers, D. Stadlbauer, K. Stone, S. Strohmeier, V. Simon, J. Aberg, D. L. Reich, F. Krammer, C. Cordon-Cardo, Robust neutralizing antibodies to SARS-CoV-2 infection persist for months. *Science* **370**, 1227–1230 (2020).
53. K. C. Garcia, M. Degano, R. L. Stanfield, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton, I. A. Wilson, An αβ T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science* **274**, 209–219 (1996).
54. G. P. Linette, E. A. Stadtmauer, M. V. Maus, A. P. Rapoport, B. L. Levine, L. Emery, L. Litzky, A. Bagg, B. M. Carreno, P. J. Cimino, G. K. Binder-Scholl, D. P. Smethurst, A. B. Gerry, N. J. Pumphrey, A. D. Bennett, J. E. Brewer, J. Dukes, J. Harper, H. K. Tayton-Martin, B. K. Jakobsen, N. J. Hassan, M. Kalos, C. H. June, Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863–871 (2013).
55. A. K. Bentzen, S. R. Hadrup, T-cell-receptor cross-recognition and strategies to select safe T-cell receptors for clinical translation. *Immuno-Oncol. Technol.* **2**, 1–10 (2019).
56. J. J. Adams, S. Narayanan, M. E. Birnbaum, S. S. Sidhu, S. J. Blevins, M. H. Gee, L. V. Sibener, B. M. Baker, D. M. Kranz, K. C. Garcia, Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nat. Immunol.* **17**, 87–94 (2016).
57. M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Özkan, M. M. Davis, K. W. Wucherpfennig, K. C. Garcia, Deconstructing the peptide-MHC specificity of t cell recognition. *Cell* **157**, 1073–1087 (2014).
58. K. C. Garcia, M. Degano, L. R. Pease, M. Huang, P. A. Peterson, L. Teyton, I. A. Wilson, Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science* **279**, 1166–1172 (1998).
59. S. K. Saini, E. T. Abualrous, A. S. Tigan, K. Covella, U. Wellbrock, S. Springer, Not all empty MHC class I molecules are molten globules: Tryptophan fluorescence reveals a two-step mechanism of thermal denaturation. *Mol. Immunol.* **54**, 386–396 (2013).
60. S. R. Hadrup, M. Toebes, B. Rodenko, A. H. Bakker, D. A. Egan, H. Ovaa, T. N. M. Schumacher, High-throughput T-cell epitope discovery through MHC peptide exchange. *Methods Mol. Biol.* **524**, 383–405 (2009).
61. S. K. Saini, T. Tamhane, R. Anjanappa, A. Saikia, S. Ramskov, M. Donia, I. M. Svane, S. N. Jakobsen, M. Garcia-Alai, M. Zacharias, R. Meijers, S. Springer, S. R. Hadrup, Empty peptide-receptive MHC class I molecules for efficient detection of antigen-specific T cells. *Sci. Immunol.* **4**, eaau9039 (2019).
62. R. Sick Andersen, P. Kvistborg, T. M. Frøsig, N. W. Pedersen, R. Lyngaa, A. H. Bakker, C. J. Shu, P. t. Straten, T. N. Schumacher, S. R. Hadrup, Parallel detection of antigen-specific T cell responses by combinatorial encoding of MHC multimers. *Nat. Protoc.* **7**, 891–902 (2012).
63. S. R. Hadrup, A. H. Bakker, C. J. Shu, R. S. Andersen, J. Van Veluw, P. Hombrink, E. Castermans, T. Straten, C. Blank, J. B. Haanen, M. H. Heemskerk, T. N. Schumacher, Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat. Methods* **6**, 520–526 (2009).
64. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018);http://arxiv.org/abs/1802.03426.
65. W.-J. Shen, H.-S. Wong, Q.-W. Xiao, X. Guo, S. Smale, Towards a Mathematical Foundation of Immunology and Amino Acid Chains (2012);http://arxiv.org/abs/1205.6031.

Acknowledgments: We thank all patients and healthy donors for participating and contributing the analyzed samples; B. Rotbøl, A. G. Burkal, A. F. Løye, and P. T. Petersen for excellent technical support; the clinical research unit for patient inclusion—K. F. Kokholm, M. L. Sieg, and J. Kock; the Department of Microbiology, Herlev Hospital, L. Nielsen, and all the collaborators for active participation to this work. **Funding:** This work is supported by the Independent Research Fund Denmark (grant no. 0214-00053B, 2020) to S.R.H. **Author contributions:** S.K.S. conceived the idea, designed and performed experiments, analyzed data, prepared figures, and wrote the manuscript. D.S.H. designed the research, facilitated patient samples, and wrote the manuscript. T.T. designed the research, performed experiments, and analyzed data. H.R.P. analyzed data, performed bioinformatic analysis, prepared figures, and wrote the manuscript. S.P.A.H. performed the experiments. M.N. supervised and performed the research and wrote the manuscript. A.O.G. conceived the idea; supervised the clinical study, patient participation, and data and sample collection; and wrote the manuscript. S.R.H. conceived the idea, designed the research, wrote the manuscript, and supervised the research. **Competing interests:** S.R.H. and S.K.S. are cofounders of Tetramer Shop, and S.R.H. is a cofounder of PokeAcell and Immumap. S.R.H. is a co-inventor on patents related to the DNA barcode–labeled MHC multimer technology, which is licensed to Immudex. These activities pose no competing interests related to the data reported here. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. The raw sequencing data can be accessed from the corresponding author upon reasonable request.

Submitted 19 November 2020
Accepted 8 April 2021
Published First Release 14 April 2021
Final published 12 July 2021
10.1126/sciimmunol.abf7550

Citation: S. K. Saini, D. S. Hersby, T. Tamhane, H. R. Povlsen, S. P. A. Hernandez, M. Nielsen, A. O. Gang, S. R. Hadrup, SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial $CD8^+$ T cell activation in COVID-19 patients. *Sci. Immunol.* **6**, eabf7550 (2021).

Science Immunology

SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8 T cell activation in COVID-19 patients

Sunil Kumar SainiDitte Stampe HersbyTripti TamhaneHelle Rus PovlsenSusana Patricia Amaya HernandezMorten NielsenAnne Ortved GangSine Reker Hadrup

Sci. Immunol., 6 (58), eabf7550. • DOI: 10.1126/sciimmunol.abf7550

Mapping SARS-CoV-2 T cell recognition
Cellular immunity mediated by cytotoxic CD8 T cells contributes to protection against viral infection, but the full spectrum of SARS-CoV-2 T cell recognition and role of preexisting T cell immunity remain incompletely understood. Saini *et al*. used DNA-barcoded peptide–MHC-I multimers to scan the SARS-CoV-2 genome for CD8 T cell recognition in patients with COVID-19. Across 10 analyzed HLA molecules, 122 unique SARS-CoV-2 CD8 T cell epitopes were detected, including 5 immunodominant epitopes primarily concentrated within ORF1. Healthy donors displayed broad T cell recognition of lower affinity and shared epitopes could be partially attributed to homology with seasonal human coronaviruses. The frequency and activation of SARS-CoV-2–specific CD8 T cells were increased during severe compared with mild disease, highlighting differences in T cell responses associated with disease progression.

Science Immunology (ISSN 2470-9468) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Immunology* is a registered trademark of AAAS.

CHAPTER 5

ATRAP - Accurate T cell Receptor Antigen Pairing through data-driven filtering of sequencing information from single-cells

The key advancement in interrogating T cell specificity is the development of immune profiling platforms for single cell sequencing. Although the platform has been available since 2019 no TCR-pMHC specificity data sets have surfaced beyond the showcase example from 10x Genomics. Even this flagship data contains a high degree of ambiguous specificity annotations which has likely limited its application. We set out to define a process for cleaning such data to retrieve reliable large-scale TCR-pMHC pairs. We developed ATRAP, which is a data-driven filtering approach based on our own generation of data. Since no golden standard exist we evaluate our method internally, on metrics designed for the purpose, and externally, by comparing single-cell specificities to responses detected by fluorescent-labeled pMHC multimer staining. This paper provides a detailed protocol for the filtering steps of ATRAP, illustrations of its application, and a description of the advantages and disadvantages of applying the single cell immune profiling framework.

ATRAP - Accurate T cell Receptor Antigen Pairing through data-driven filtering of sequencing information from single-cells

Helle Rus Povlsen*, Amalie Kai Bentzen*, Mohammad Kadivar, Leon Eyrich Jessen, Sine Reker Hadrup*, Morten Nielsen*

Abstract

Novel single-cell based technologies hold the promise of matching T cell receptor (TCR) sequences with their cognate peptide-MHC recognition motif in a high-throughput manner. Parallel capture of TCR transcripts and peptide-MHC is enabled through the use of reagents labeled with DNA barcodes. However, analysis and annotation of such single-cell sequencing (SCseq) data is challenged by dropout, random noise, and other technical artifacts that must be carefully handled in the downstream processing steps.

We here propose a rational, data-driven method termed ATRAP (Accurate T cell Receptor Antigen Paring) to deal with these challenges, filtering away likely artifacts, and enable the generation of large sets of TCR-pMHC sequence data with a high degree of specificity and sensitivity, thus outputting the most likely pMHC target per T cell. We have validated this approach across 10 different virus-specific T cell responses in 16 healthy donors. Across these samples we have identified up to 1494 high-confident TCR-pMHC pairs derived from 4135 single-cells.

Introduction

T cells are essential for immune protection and play a critical role in the immune response to pathogens or cancer, where they directly kill infected or malignant host cells or orchestrate the response of other immune cells. Recognition is mediated through the heterodimeric T-cell receptor (TCR) expressed on the surface of T cells, which engages specifically with a peptide antigen (p) displayed in the MHC. Accurate specificity and broad coverage of antigen recognition is obtained through somatic recombination of the genetic loci, V(D)J, that encodes the α (VJ) and β (VDJ) chains of TCR. The process creates an extensively variable and dynamic repertoire, with an

estimated 10^7 distinct αβTCRs in an individual (Arstila et al., 1999; Davis & Bjorkman, 1988).

Due to this diversity, the individual TCR transcripts can be used as endogenous cellular barcodes inherited by the T cell progeny. This has been utilized for providing quantitative insight into TCR diversity (Robins et al., 2009), to trace lineage decisions of T cells (Gerlach et al., 2013) and to monitor the dynamics of T cells across immune-related diseases, such as infectious disease (Dziubianau et al., 2013; Hou et al., 2016), cancer (Kirsch et al., 2015; Sherwood, 2013; S. Q. Zhang et al., 2018) and autoimmunity (Acha-Orbea et al., 1988; Madi et al., 2014). Most of such TCR repertoire studies have been confined to bulk experiments, tracing the TCR β chain because of its greater diversity (compared to the alpha chain) and because it is less ambiguous due to allelic exclusion (Bergman, 1999). However, accurate pairing of the variable TCR α and β regions is valuable for uncovering the biological function of a T cell and is generally lost in bulk experiments since the transcripts are separately encoded. Moreover, we and others have earlier demonstrated such approaches are suboptimal for the characterization of TCR specificity, and that this characterization is dependent on both the a and b chains (Montemurro et al., 2021).

To accurately obtain TCR αβ-sequence-pair single-cell sequencing platforms can be applied to simultaneously capture both TCR chains, while retaining cell origin information. To further assign specificity information to such TCRs, T cells can be stained with barcode-labeled pMHC multimers to simultaneous identify pMHC specificity and TCR sequence of individual cells (Bentzen et al., 2016; S. Q. Zhang et al., 2018). Moreover, via DNA barcoded antibodies, the platform facilitates screening of surface proteins to distinguish cellular subtypes and enables cell hashing to trace origin of a given cell to e.g., a given donor, sample, or time-point, which is highly valuable in patient-studies.

We deployed the droplet-based single-cell platform from 10x Genomics. Ideally a droplet contains a single cell with all its analytes and a gel-bead in emulsion (GEM). The gel-bead contains barcoded primers which ensures tracing of transcripts back to the cell-of-origin, referred to as GEMs. While the platform is highly promising, the sequence deconvolution is associated with substantial noise, and challenging to discriminate true from false signals. Common confounding factors include stochastic gene expression, cell cycle variations, apoptosis, and technical artifacts such as multiplet capture, contamination, dropout, and batch effects. Dropout and stochastic gene expression both result in zero-inflated gene counts and are typically insensitive to low expression levels (Buettner et al., 2015; Kharchenko, Silberstein, & Scadden, 2014; Yamawaki et al., 2021). Multiplet capture is the event of capturing two or more cells in a single GEM and it is proportional to the capture rate of cells introduced to the system (Bloom, 2018; Zheng et al., 2017). The capture rate is determined by the rate of pulsing cells relative to the rate of gel-beads. Thus, to include even low frequency cell populations, the capture rate must be high at the expense of

introducing more multiplets. Contamination is particularly an issue when including analytes such as pMHC multimers which may be dissolved in cell suspension (Gaublomme et al., 2019). The platform has no means of controlling how ambient analytes and their barcodes are partitioned with gel-beads in emulsion (GEMs) which leads to GEMs that appear like multiplets or consist of ambiguous annotations from multiple analyte barcodes. The reverse issue arises from the risk that analytes may dissociate from the cell before capture. The listed confounders may result in both false positive and false negative discoveries

The main concerns when screening for TCR specificity are nonspecific binding of pMHC and/or cell hashing analytes, incomplete TCR annotation, and T cell multiplets. Nonspecific binding and T cell multiplets may completely dilute the signal from actual interactions, while incomplete TCRs which are missing the annotation for either α- or β-chain render the single-cell setup superfluous. To ensure that a screening is fully exploited and interpreted correctly, we set out to develop a data driven algorithm that facilitates a consistent and reproducible TCR categorization (clonotyping), peptide-MHC (pMHC) annotation, and antibody-based cell hashing referencing of the donors and their HLA profile.

We applied this algorithm to two datasets, each derived from screening PBMCs from 16 healthy donors for T cell recognition against common viruses. In total, we evaluated TCR recognition against 10 different pMHC multimers, each labeled with their unique barcode. We demonstrate that following the filtering steps described here we can obtain a confident pairing of pMHC specificity and TCR sequence. This strategy will open novel opportunities to evaluate the structural interplay and the sequence-driven signatures of pMHC recognition at large scale.

Results

Parallel capture of TCRαβ sequences, peptide-MHC specificity and sample origin from single-cells

To obtain single-cell-derived triad information on TCR sequence, pMHC specificity, and sample origin; we stained peripheral blood mononuclear cells (PBMC) from a total of 16 different healthy donors (Table 1). All samples were stained with the same panel composed of 10 different viral-derived peptide-MHC (pMHC) multimers, each labeled with a unique barcode for that specificity and a common fluorescent label (allophycocyanin (APC)) (Fig. 1) (Table 2). To serve as an experimental control for the purity of the isolated T cells, we moreover stained the cells with three additional viral-derived pMHC multimers bearing a different fluorochrome (phycoerythrin (PE)) and labeled with their own unique DNA barcode (Supplementary Table 2). We sorted

only the APC-labeled pMHC multimer binding T cells (and hence deselected the PE-labeled T cells) and included these in the down-stream single-cell processing.

Prior to sorting, each sample was stained with a distinct hashing antibody to provide a sample identification barcode associated with the GEMs of the resultant single-cell data set. This is done to enable mixing of cells from different samples, while retaining the information of sample origin, and utilizing the capacity of capturing 6,000-10,000 cells per lane in the 10xGenomics workflow. This is essential when capturing T cells based on their specificity since the MHC multimer positive population is generally of low frequency (<1% of CD8 T cells). When several samples are mixed in the process of running the single-cell analysis, all mRNA and DNA barcodes (derived from hashing antibodies or the MHC multimers) associated with a given cell will be encoded with the same 10x-barcode, proving the GEM association (Fig 1) (Supplementary Table 1).

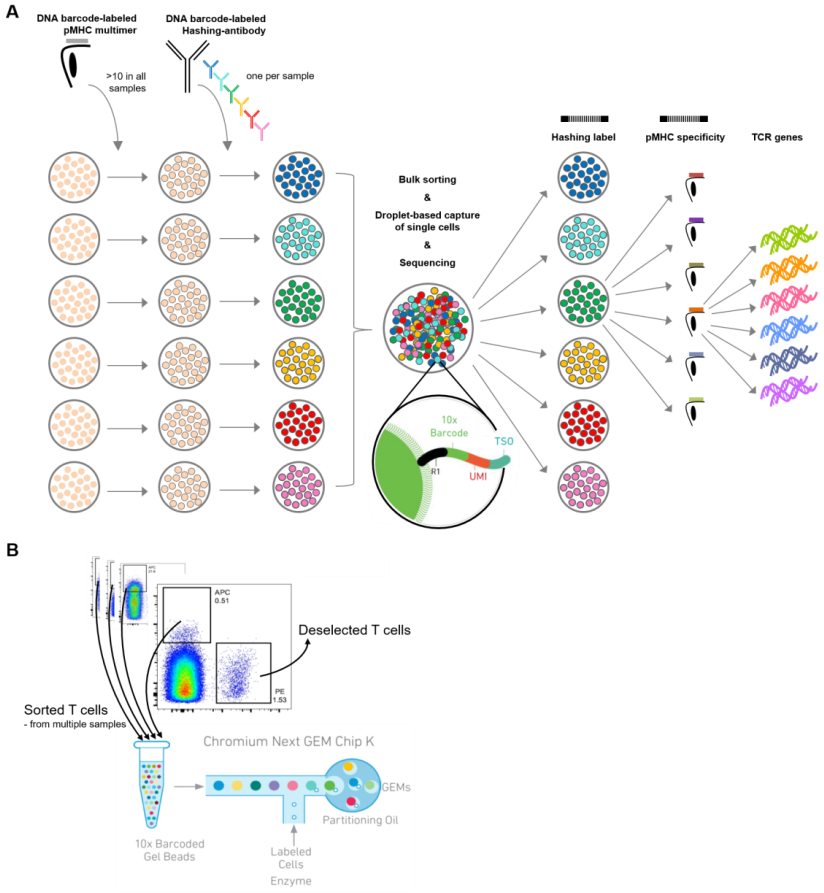


Figure 1. a) Schematic of the experimental strategy. All samples are incubated with the same library of barcode-labeled pMHC multimers and subsequently with a sample-specific barcode-labeled hashing antibody to individually label cells derived from a given sample. Multimer-binding cells from all samples are sorted in bulk and processed through the 10x Chromium workflow. The sequencing output simultaneously captures the sample barcode, the pMHC barcode and the TCR sequences, which are all matched to a single cell based on the 10x-barcode. This also provides the means of retrospectively assigning each cell to their sample of origin, via the sample specific hashing barcode. b) Example showing how the APC labeled pMHC multimers are sorted collectively from all samples into one tube that is further carried into the 10x workflow. The PE labeled pMHC multimers are not sorted and hence deselected. A total of 1800 APC labeled cells are sorted from each donor. Here showing BC126 (large dotplot) and BC341 (small dotplot).

Total data from simultaneous capture of cell, TCR, pMHC and SampleID

The single-cell data is annotated using 10x Chromium Cellranger multi v6.1. This results in each GEM being quantified by a count of unique molecular identifiers (UMIs) (Kivioja et al., 2011) for the three components (TCR, pMHC and sample hashing) based on transcripts of TCR α- and β-chains, barcodes co-attached to pMHC multimers and barcodes co-attached to cell hashing antibodies (Supplementary Table 2).

To obtain the data presented here, a total of 1800 pMHC multimer positive cells were sorted per donor irrespective of the frequency or the number of different antigen-specific T cell responses in a given sample, accumulating to a total of 28,800 cells sorted. An estimated 45% of the sorted cells are lost in the process of loading on the Chromium, hence approximately 15,700 pMHC multimer labeled cells were included in the 10x experimental workflow. Initially, each GEM was annotated based on the most abundant transcripts from TCRαβ, pMHC, and cell hashing. However, this can lead to erroneous annotations, as the noise level can differ substantially for the different reagents, resulting in different levels of UMIs.

Based on raw, unfiltered data, we found 6,073 GEMs which contained all three components i.e., TCR, pMHC and sample hashing, corresponding to 40% of the loaded cells (Fig. 2a). 716,069 GEMs only contained one or two of the components, with the majority containing only the cell hashing barcode (n=677,502) and the second largest share containing cell hashing as well as pMHC barcodes (n=37,277). This number vastly exceeds the number of cells in the assay (15,700 cells loaded) and indicates contamination from ambient barcodes in suspension. This is further supported by the observation that the sample hashing UMI count was significantly higher ($p < 0.0005$, Mann-Whitney U) in the 6,073 GEMs containing a TCR compared to the GEMs void of TCR (Fig. 2b). 43,455 GEMs captured a DNA barcode associated with the pMHC library and only 14% of these were completed with TCR transcripts and sample hashing barcodes. In the GEMs containing a TCR, 84% were completed with all three components i.e., included hashing and pMHC barcodes, while less than 0.05% of these GEMs were void of both sample hashing and pMHC barcodes. In the following, we will only consider the 6,073 GEMs containing all three components, while taking into account that the high degree of noise also affects these seemingly completely mapped GEMs.

The GEMs are distributed across three categories of TCR and two categories of pMHC observations: GEMs either missing a TCR chain, contain multiple TCR chains, or contain a unique TCRαβ-pair and GEMs containing either a single or multiple pMHC barcodes (Fig. 2c). Sample hashing multiplets constitue 100% of GEMs containing sample hashing barcodes, and there is both a large proportion of pMHC multiplets (65%) and GEMs missing either α- or β TCR-chain (39%), hence,

multiplets of pMHC and sample hashing is the predominant issue. Few GEMs were detected with multiple TCR α- or β-chains (6%). This may be caused partly by naturally occurring multiplets of α-chain (4%), due to the incomplete gene restriction of the thymocyte during negative selection (Elliott & Altmann, 1995; Petrie et al., 1993), or due to experimental features of the 10x platform causing an expected 6.9% of multiplets based on the number of cells loaded in our experiment.

Without further filtering, the pMHC-TCR pairing is subjected to extensive noise (Fig.2d) and we capture all the 10 DNA barcodes associated with the APC-labeled pMHCs in a varying number of GEMs. Importantly, the three negative control responses (GIL A0201, GLC A0201, and NLV A0201), which were present in the donors but not sorted, are only captured in a few GEMs. This indicates that the cell isolation via sorting is effective in terms of capturing only the desired cells and relevant pMHC-associated barcode-labels. The most frequently detected pMHC across all GEMs is RVR A0301, which is present with high UMI counts across all GEMs. Only RPH(10-mer) B0702-associated UMIs was consistently detected at low numbers per GEM. It was also evaluated whether the HLA allele of the pMHC matches the HLA haplotype of the donor(s) given via cell hashing (Fig. 2d). Typically, the mismatches are found in GEMs where the most abundant pMHC is detected at low UMI counts while the matches consist of GEMs with higher pMHC UMI counts. Of the 65% GEMs containing pMHC multiplets (Fig. 2c), 13% contained two or more pMHCs at the exact same UMI level (Supplementary Table 3), which may either represent noise or true cross-binding events.

The detected specificities in our data have been cross-referenced with the IEDB (Vita et al., 2019) and VDJ (Bagaev et al., 2020) databases (Fig. 2d). Based on the unfiltered data we found five TCR-pMHC matches (across 9 GEMs) and one TCR (1 GEM), which was annotated with a different pMHC (Fig. 2d). This latter is a case of a GEM with multiple pMHCs present with almost equal number of UMIs, where the most abundant pMHC is RVR A0301 (11 UMIs) and the second most abundant pMHC is GLC A0201 (9 UMIs), which is the peptide registered as target in IEDB and VDJdb.

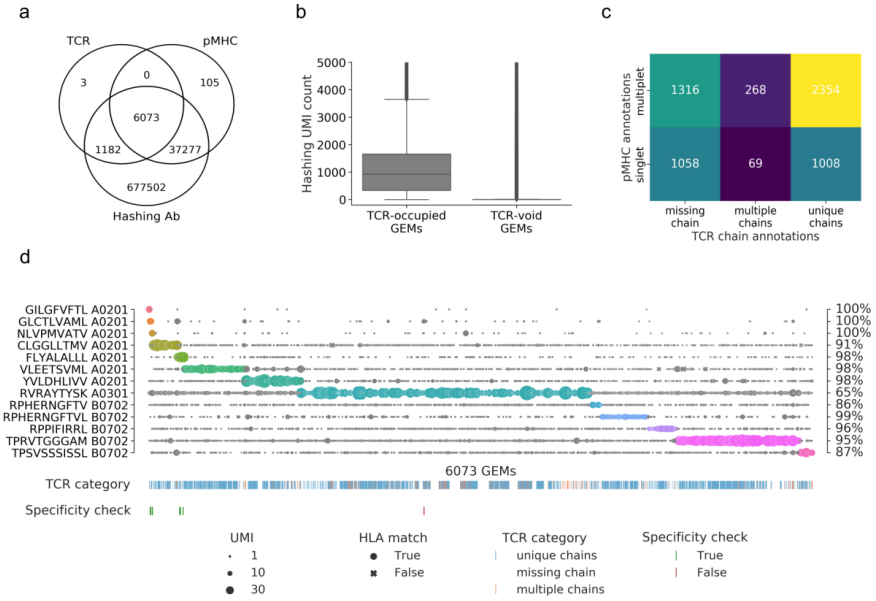


Figure 2: a) Venn-diagram of the content of all GEMs from 10x Chromium drop-seq. Each GEM is expected to contain three components: transcripts of TCR and DNA barcodes from the target pMHC multimer as well as the sample hashing antibody. The Venn-diagram illustrates the extent of GEMs with complete capture (capture of all three components) in contrast to the GEMs with incomplete capture (capture of a subset of components). b) Comparison of distributions of UMI counts of sample hashing barcode between GEMs that contain TCR transcripts (TCR-occupied GEMs) and GEMs that do not contain TCR transcripts (TCR-void GEMs) ($p < 0.0005$, Mann-Whitney U). c) Matrix of the distribution of pMHC singlets and multiplets across GEMs with TCRs either missing a chain, detected with multiple chains, or with a single, unique αβ-pair. The counts are given for each field and illustrated by a color. The lighter color represents higher counts. d) Scatterplot of all detected pMHC barcodes (y-axis) within each of the 6073 GEMs (x-axis) that contain all three components: TCR, pMHC and sample hashing. In each GEM the most abundant pMHC is marked by a color, while the remaining pMHCs in the GEM are gray. The marker size reports the UMI count of the given pMHC and the shape recounts whether the HLA allele of the pMHC matches the HLA haplotype of the donor, which is deduced from sample hashing. The fraction of HLA matches within the GEMs displaying a given specificity is annotated to the right of the plot. The first colorbar indicates the type of TCR chain annotation; whether the TCR has a unique αβ-pair, is missing a chain or consists of multiple chains. The second colorbar is a specificity check against the specificity databases IEDB and VDjdb. Colors highlight the GEMs where the CDR3αβ sequences are contained in the databases. The green color represents a match between the database pMHC and the detected pMHC, while red indicates a mismatch.

The data in Fig. 2d suggests that most of the captured T cells interact with several of the screened pMHCs to a degree that exceeds the level expected from natural

cross-recognition. Therefore, it is reasonable to assume that a large proportion of these multiplets are formed as a result of ambient pMHC leaking into GEMs.

A data-driven filtering approach

From these observations, it is clear that a substantial part of the data consists of noise i.e., GEMs with multiplets of pMHC and sample hashing, and that the data must be filtered for proper interpretation.

Clonotype annotation

The definition of T-cell clones (clonotypes) is fundamental for pairing a given TCR clonotype to its respective pMHC recognition. Initial clonotypes were called using 10x Genomics Cellranger which defines a clonotype as a set of cells that share identical receptor sequences at the nucleotide level, spanning the entirety of the V(D)J-C genes as well as the junction segments. Assuming reliable gene and CDR3 sequence calls by 10x Cellranger, we redefine clonotypes based on TCR annotation. Subsequently, GEMs with no clonotype annotation from 10x were annotated to existing clonotypes conditioned on matching VJαβ-genes and CDR3αβ sequences or as novel clonotypes. Similarly, clonotypes with identical VJ-CDR3αβ were merged to form larger groups of theoretically identical TCRs (Supplementary. fig 1). Merging GEMs of the same TCR is essential to make statistical inference based on those groupings e.g., determine expected pMHC target per clonotype. The outcome was a set of 2,441 TCR clonotypes across the 6,073 GEMs containing both TCR and pMHC. For the 337 GEMs containing TCR chain multiplets, the most abundant chain was for the subsequent analyses selected to represent the true TCR.

Defining pMHC recognition for selected TCR clonotypes

As we have seen earlier, not all GEMs within a given clonotype support the same pMHC target, and defining the pMHC target of a TCR based on individual GEMs thus results in contradicting annotations. The key to identify the expected target for a clonotype is therefore to determine which pMHC identity represents the majority of UMIs across all GEMs within a given clonotype. Fig. 3 illustrates an example from a pilot study which accentuates the importance of studying GEMs in ensemble rather than individually. Most GEMs are annotated with multiplets of pMHCs and across all GEMs the most abundant pMHC varies. While all pMHCs are found most abundant in at least one GEM, three pMHCs (TPR B0702, VTE A0101, and RAK B0801) are more often found most abundant (Fig. 3a). Although TPR B0702 is detected in fewer GEMs (136) than VTE A0101 (260) and RAK B0801 (186), TPR B0702 is present at generally higher UMI counts (Fig. 3b). It is evident that there is a difference in UMI distributions between the different pMHC within the GEMs of a given clonotype, and that TPR B0701 is the significantly most abundant pMHC across the ensemble of GEMs even though this pMHC is only present in a minority proportion of the GEM (Fig. 3b). Based on these observations, we argue that the significantly most

abundant pMHC should be annotated as the expected binder for the given clonotype rather than annotating based on the majority.

Having annotated the “true” pMHC of a given clonotype, one can next go back to the individual GEMs, and label GEMs where the most abundant pMHC corresponds to the expected binder, as “true”, and all others as “false”, and use these annotations to quantify the accuracy of the GEM annotations. Within each clonotype, one can compute a specificity concordance i.e., the fraction of GEMs detected with a certain specificity (defined by most abundant pMHC i.e., highest pMHC UMI per GEM) (Fig. 3c). In many cases across the full data set, the expected specificity for a clonotype coincides with the specificity, defined on a per-GEM level, resulting in high concordance. However, for some clonotypes e.g., clonotype 1, GEMs have diverging annotations and therefore lower concordance dispersed across multiple specificities (Fig 3). The clonotype visualized in Fig. 3 is specifically chosen to exemplify how this lower concordance can affect the analysis. For clonotype 1 the fraction of GEMs that support VTE A0101 (0.33) is higher than the fraction of GEMs that supports TPR B0702 (0.26). This results in an overall low concordance, and only by considering the complete ensemble of clonotype 1 GEMs, can the correct pMHC target be identified (Fig. 3b).

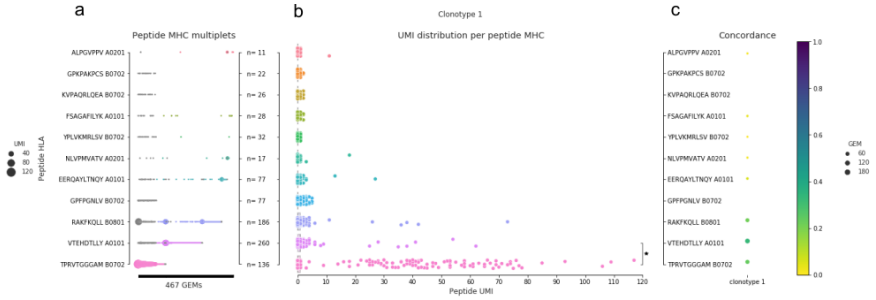


Figure 3: An example of pMHC concordance in clonotype 1 (example from pilot study). a) All detected pMHC (y-axis) in each GEM (x-axis, n=467) of clonotype 1. The marker size shows the UMI count for the particular pMHC in a given GEM, and the color indicates the pMHC with the highest UMI count, similar to what is shown in Fig. 1d. If two pMHCs are equally abundant in a GEM they are both colored. No marker means no detection of that pMHC in that given GEM. b) The compiled distribution of UMI counts for each peptide (assigning 0 UMI when the pMHC is not detected in a GEM). The asterisk marks that a Wilcoxon test showed that the UMI counts of TPR B0702 were on average higher than for VTE A0101 UMI counts. c) The specificity concordance across the GEMs of clonotype 1. Concordance is shown by a color gradient i.e., the larger the fraction of GEMs supporting a given specificity the darker the color.

Improving concordance between GEM and clonotype annotation based on grid search on UMI features

To rationally filter data, an accuracy metric was defined, and optimized through the filtering process. For all specificities belonging to clonotypes with an assigned expected target, we calculated the overall accuracy as the proportion of GEMs where highest abundance pMHC annotation corresponds to the expected target of the clonotype. The raw unfiltered data yielded accuracy and average concordance scores of 69.6% and 83.8%, respectively. Next, we set out to investigate how different data driven UMI filters could improve these performance values, removing noise and artifacts from the data. This removal would also reduce the number of included observations, hence the performance of different thresholds for filtering the data was evaluated based on a tradeoff between increased accuracy and discarded number of GEMs.

We tested various thresholds on UMI count and UMI ratios i.e., the ratio between the most abundant and second most abundant UMI feature, for pMHC and TCRαβ respectively. The optimal thresholds were chosen to maximize the weighted average between accuracy and fraction of retained GEMs to favor increase in accuracy above losing some GEMs. This filtering analysis resulted in optimal thresholds of 2 pMHC UMI counts and a ratio pMHC UMI counts between top one and two >1.The latter results in removal of GEMs where two pMHC were equally abundant for low UMI counts. The search did not result in thresholds imposing restrictions on neither TCR UMI counts nor TCR UMI ratio, which underpins that the TCRs with a missing chain as well as multiple chains also contribute to good performance. Imposing this filter yielded 5,061 GEMs (83% of total), 2,233 clonotypes (91% of total), and resulted in 96.4% accuracy, and a mean concordance of 93.6%.

Additional filters

Additional filters can be added to further clean the data. We compared how two filters, integrated in the 10x Genomics software, Cellranger, performed in removing potential noise from our data set (Supplementary Fig 2). The purpose of these filters is to evaluate, with high confidence, whether a GEM has captured a cell:
“is cell” is defined based on the TCR transcript level in a given GEM and “is cell (GEX)” is defined based on the full transcript level (10xGenomics, n.d. a). Alternatively, viable cells are identified from the transcript data, independently of Cellranger, based on mitochondrial load and a minimum and maximum gene count per GEM. All three filterings are comparable (Supplementary Fig 2), and taken into account in the further evaluations. It is worth noting that, while the filterings based on the full transcript data might remove slightly more noise, the economic costs associated could propose that this should only be applied when the transcript data is required for additional purposes.

Cell hashing enables filtering based on sample demultiplexing methods such as Seurat hashtag oligo (HTO) demultiplexing to identify hashing singlets (Stoeckius et al., 2018) (Supplementary Fig 3 and Supplementary note). In this setup, cell hashing also enables filtering based on matching HLA between the donor haplotypes and the HLA of the detected pMHC. Additionally, depending on the subsequent use of the data, retaining only complete TCRs containing both α and β may be desirable. Including only GEMs where the TCR-pMHC pair is observed more than once i.e., specificity multiplets, reduces the uncertainty described above. Below we investigate the impact of imposing such filters.

Impacts of filtering

Evaluating filters by comparing TCR similarity across specificity

To objectively evaluate the performance impact of the presented filters, we define a quantitative evaluation based on the hypothesis that T cells binding the same pMHC (intra specificity) will share a higher sequence similarity compared to TCRs of different specificities (inter specificity) (Fig. 4). Thus, filtering away artifacts should increase intra-similarity while decreasing the inter-similarity. Here, the similarity score between two TCRs was calculated from the summed score of the pairwise α- and β-chain similarities calculated using a kernel method described in (Shen, Wong, Xiao, Guo, & Smale, 2012) and applied in (Chronister et al., 2021).

Based on this kernel similarity metric, the filters were tested individually and cumulatively i.e., each filter was added to the previous set of filters. The general trend is that TCRs with the same specificity are more similar to each other than to TCRs of different specificities, when computing the intra and inter similarities per pMHC before and after filtering on the optimized UMI thresholds (Fig. 4a-b). Before filtering, nine out of 13 pMHCs displayed a higher mean intra-similarity than inter-similarity scores, whereas this number was 11 out 13 pMHCs when applying the UMI thresholds. The outliers before filtering were GIL A0201, VLE A0201, CLG A0201, and RPP B0702, while the outliers were reduced to VLE and RPP after filtering. Generally, the similarity scores often have a wide, overlapping range between the intra and inter categories. The three pMHCs that were deselected during sorting, GIL A0201, GLC A0201, and NLV A0201, are only detected in a few TCR binding events. To enhance the power of comparison, the intra and inter scores were pooled respectively across the individual pMHCs (Fig. 4c-d). The results demonstrate that intra-similarity is significantly higher than inter-similarity at each filtering step, both individually and combined. Moreover, we observe that the differences between intra- and inter-similarity appear to increase as filters are cumulatively added and fewer observations are left (Fig. 4d). Particularly, the median inter-similarity score is lowered, suggesting that the filtering steps predominantly removes false-positives.

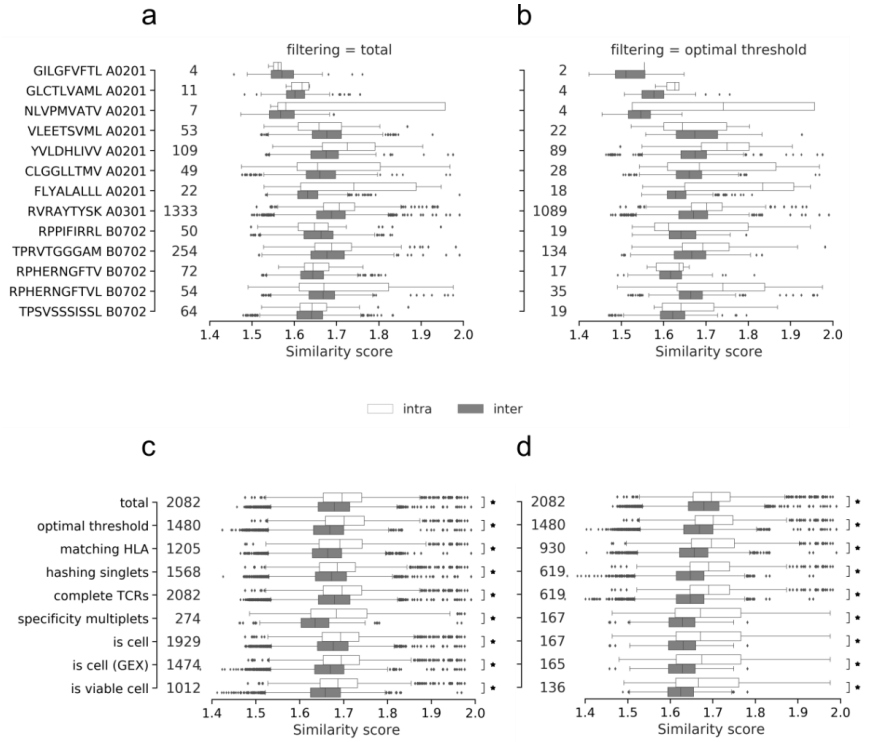


Figure 4: Intra- and inter TCR-similarity scores per peptide of the a) total (unfiltered) dataset, b) the data filtered by the optimized threshold. The similarity per peptide plots a) and b) illustrate the distribution of paired similarity scores for each clonotype (containing both α- and β-chain). For each pMHC each clonotype is compared to the remaining clonotypes of the same specificity (intra) and across specificities (inter). The count of compared clonotypes is listed just to the right of the y-axis in both a) and b). c) Displays the pooled intra- and inter-scores across all peptides for each of the filtering methods: total (no filtering), optimal threshold, matching HLA, hashing singlets, complete TCRs, specificity multiplets, “is cell” by cell-flagging, “is cell” by cell-flagging when including GEX data, and viable cell from analyzing GEX data. An asterisk marks filters where intra-similarity is significantly larger than inter-similarity (Wilcoxon, α=0.05). d) Displays the pooled intra- and inter-scores across all peptides for each of the filtering methods where each filtering is added cumulatively to the previously listed above it. An asterisk marks filters where intra-similarity is significantly larger than inter-similarity (Wilcoxon, α=0.05). The count of compared clonotypes is listed just to the right of the y-axis in both c) and d).

Evaluating filters across selected performance metrics

To compare the effect of the filters, the similarity scores were converted to the performance metric: AUC (area under the receiver operating characteristic (ROC)

curve). Here, intra specificity comparisons are regarded as true positive observations and inter specificity comparisons as true negatives. Based on these performance metric definitions, we quantify the effect of each filtering step (Fig. 5), and find that the highest accuracy and highest average concordance is obtained by filtering on the optimal threshold (95.3% and 90.6%), while the highest AUC is obtained from filtering on specificity singlets (70.5%) (Fig. 5a). Expectantly, the accuracy and average concordance increases when the filters are imposed cumulatively (Fig. 5b). The accumulation of filters also results in drastic reduction of the GEMs, and it is evident that one must carefully weigh out the need for specificity over sensitivity when selecting the desired set of filters.

We conclude that the minimal filtering must include optimal threshold and matching HLA between pMHC and donor haplotype. Filtering on specificity multiplets would inherently result in more reliable observations, risking the removal of rare, low-avidity binding events. Generally, we did not find that including GEX data improved performance considerably. Finally, filtering on incomplete TCRs yields the second highest accuracy and average concordance. Unfortunately, the filter almost halves the number of GEMs. Hence, this filtering should be considered depending on future use of the data.

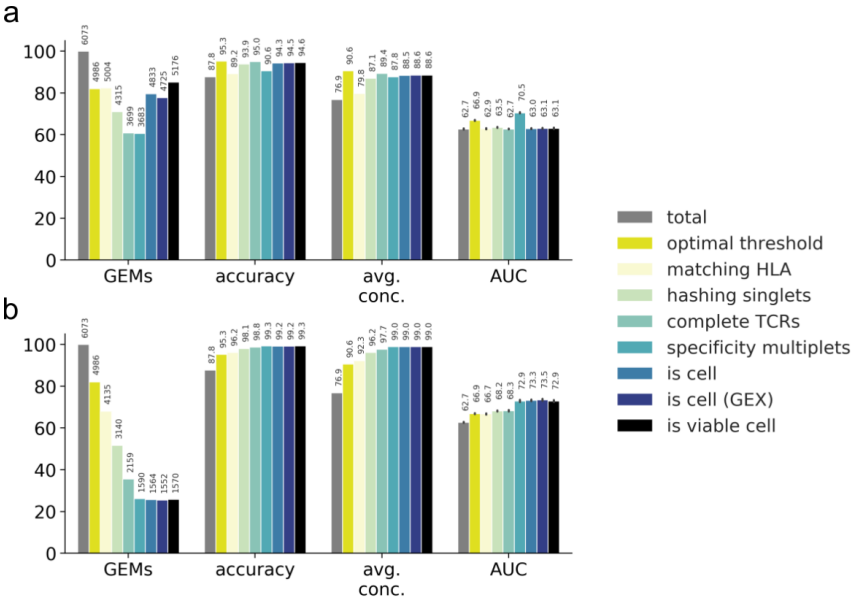


Figure 5: Performance metrics for evaluating the filtering steps. Performance is measured by number and ratio of retained GEMs (GEMs), accuracy defined by proportion of GEMs where most abundant pMHC matches the expected binder (accuracy), average binding concordance (avg. conc.) and AUC of similarity scores (AUC). The filtering steps consist of total (raw, unfiltered data), optimal threshold obtained from grid search, matching HLA,

hashing singlets identified from Seurat HTO demultiplexing, complete TCRs with a unique set of α- and β-chain, specificity multiplets such that each TCR-pMHC pair must be observed in two or more GEMs, is cell defined by 10x Genomics Cellranger, is cell (GEX) defined by Cellranger where GEX data is included, and is viable cell defined by mitochondrial load and gene counts. a) Presentation of the individual effect of each filter. b) Presentation of the accumulated effects of the listed filters.

Inspecting the filtered data

To determine the impact of the filtering steps, we have compiled the binding concordance for all clonotypes and applied three selected filtering steps: a) the raw, unfiltered data, b) filtering on optimal UMI thresholds and matching HLA, and c) additionally filtering on complete TCRs (Fig. 6). The raw, unfiltered data displays many clonotypes where the most abundant pMHC in GEMs of a given clonotype are dispersed across multiple of the screened pMHCs (Fig. 6a). When imposing the recommended set of filters, optimal threshold and HLA match, the outliers are greatly reduced, although not all low-concordance GEMs are removed (Fig. 6b). By additionally filtering on complete TCRs even fewer outliers are left (Fig. 6c). Note again that we have purposely deselected T cells specific for GIL A0201, GLC A0201, and NLV A0201, explaining the few observations for these otherwise frequently recognized epitopes.

Many of the remaining low concordance GEMs still suggest the improbable event of cross-binding across HLA restriction. We suspect that these are artifacts that we have not successfully removed. When the most strict filtering is imposed (Fig. 6c) there are 56 GEMs (out of 2833) with a binding concordance of 0.5 or lower, which will be referred to as outliers. 50 of those GEMs contain pMHC multiplets. 94% of the multiplet outliers actually do contain the pMHC which defines the high-concordance GEMs, however, at a lower UMI count. In the GEMs with multiple pMHC annotations, the HLA is conserved across the pMHCs in 14% of the cases. In 68% of the cases the HLAs are different, but still match the HLA haplotype of the donor given by the cell hashing. Of the 56 outliers, the most dominant pMHCs are RVR A0301 (41%) and TPR B0702 (27%). Prior to filtering the data, six clonotypes were identified which were already registered in IEDB and VDJdb, five with matching pMHC and one with a different annotation than in our observation (Fig. 2d). The five matching clonotypes (9 GEMs) were successfully retained, while the mismatching clonotype (1 GEM) was filtered away.

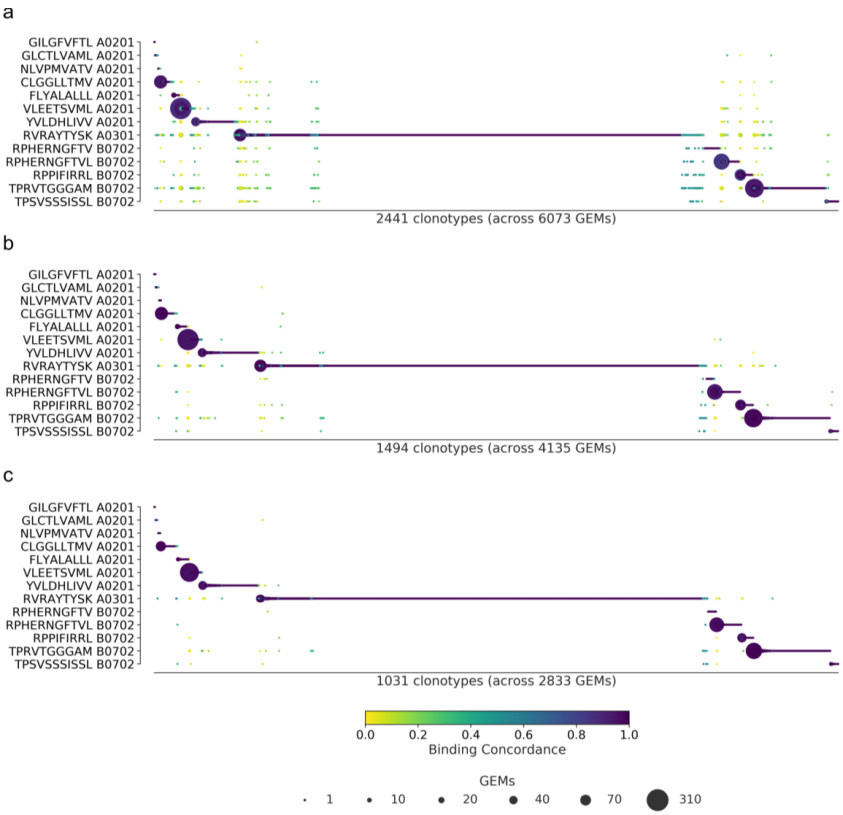


Figure 6: Specificity per clonotype. The library peptides are listed on the y-axis and each clonotype is represented on the x-axis. Below the x-axis is annotated the total number of clonotypes and GEMs in the presented data. The marker size shows the number of GEMs supporting a given specificity. The color indicates the binding concordance which is calculated as the fraction of GEMs within a clonotype that support a given pMHC. The higher the concordance, the larger the fraction of supporting GEMs. The three plots illustrate the impact of three filtering criteria. a) Presents raw data with no filtering applied. b) Presents data filtered on optimal threshold and HLA matches. c) Presents data filtered as in b) with the additionional requirement of only complete TCRs.

Comparing single-cell data with fluorescent-based pMHC multimer screening

Investigating dominant clones

Beyond mapping the landscape of known TCR-pMHC interactions, single-cell screening enables investigation of T cell repertoire diversity. The high resolution both reveals the specificity and the TCR clonality within the individual T cell populations,

which is not possible to recover in classical stainings using fluorescent labeled pMHC multimers (fluorescent multimers). The T cell diversity in the nine donors towards the set of analyzed pMHCs reveals a clear hierarchy with dominant responses in fluorescent multimer staining (Fig. 7a), however the clonality of each specificity is only available via single-cell data (Fig. 7b). Here ATRAP represents data filtered by optimal UMI thresholds and matching HLA between pMHC and donor haplotype (given via cell hashing). Single-cell screening further enables comparison of the clonal distribution and the total clonal size per specificity. In this respect, the samples BC328 and BC62 are strikingly similar in their distribution of expanded clones. They both display a large and broad response towards RVR A0301 and two smaller responses towards RPP B0702 and TPS B0702 which are both dominated by a single clonotype. Further, most peptides elicit diverse relative responses between samples. For example, RPP B0702 is the dominant response in samples BC328 and BC62, but the minority response in sample BC314. Sample BC300 contains primarily small clones i.e., fewer cells in each clonotype, however, this sample is generally represented with low amounts of total data (46 GEMs). Of note, small clones might be a result of suboptimal single-cell capture, or because high-frequency responses can potentially mask any lower frequency responses present in a given donor (Supplementary Table 4) when only 1800 cells are sorted from each sample. Samples represented with many GEMs are expected to be fully covered and therefore may contain more different expanded clonotypes, as sample BC360.

Evaluating ATRAP by “ground truth” of fluorescent-based pMHC multimer screening

The net-overlap of identified T cell responses between the two screenings (Fig.7a+b) is estimated to 0.63 by Matthew’s Correlation Coefficient (MCC). Most of the T cell populations detected by fluorescent multimers are also captured in the single-cell screening, reflected by a recall of 0.95. However, the single-cell capture of small T cell clones (Fig 7b) that were not detected using fluorescent multimers (Fig 7a), negatively impacts the precision, yielding a score of 0.71. These “false positives” could result from low affinity T cell clones where the fluorescent signal would not be distinguishable from the background. Importantly, most of these responses were only represented by 1 GEM per clonotype, (demonstrated by a light gray outer circle in Fig 7b). In only two cases T cell populations were detected with fluorescent multimers but not captured in single cells: BC316/CLG A0201 and BC62/ RPH(10mer) B0702 (Fig. 7a). The large T cell population of BC316/CLG A0201 was likely a technical artifact related to the barcode-labeled pMHC multimers.

We calculated the number of antigen specific T cells sorted per donor, based on the total number of sorted cells/donor (n=1800) and the frequency of each T cell population (Fig 7c and Supplementary Table 4). This number of sorted cells for a given specificity was strongly correlated with the numbers of single-cell GEMs assigned to the same specificity (Pearson correlation coefficient, PCC=0.73,

p<0.0005). We also fitted a linear regression for T cell populations sorted and assigned with at least one adjusted cell count or GEM in the log-log space (R2=0.56). The regression indicates that ~10% of sorted cells will be captured in a single-cell screening with TCR-pMHC information yielded by ATRAP.

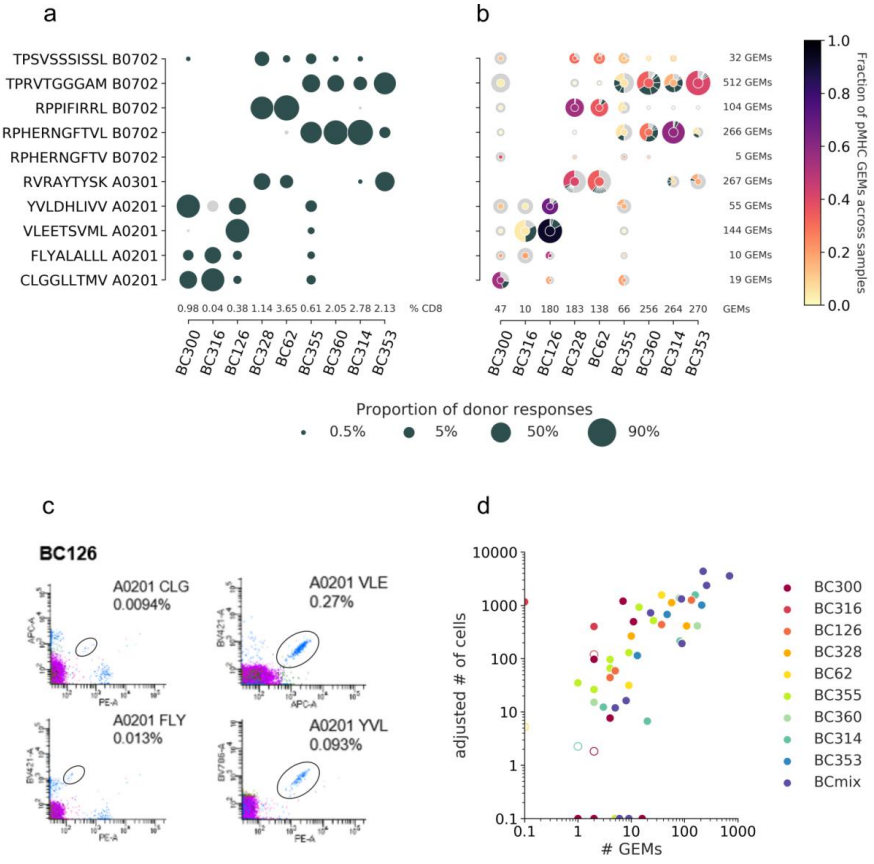


Figure 7: T cell diversity per peptide across the individual samples. The nine samples, PBMCs from nine individual donors are represented on the x-axis. The marker size defines the distribution of T cells recognizing a given peptide, normalized per sample. a) The T cell frequencies are visualized as the proportion of a given multimer positive response within a donor. The black markers represent responses detected above the threshold i.e., ≥10 cells and ≥ 0.002% of total CD8 T cells, or ≤10 cells but ≥0.01% of total CD8 T cells. The gray dots represent detected specificities below threshold but represented by ≥2 cells. Summed frequencies of detected responses within a donor are given as % of total CD8 T cells and listed just above the x-axis. b) The T cell frequencies are based on GEM counts normalized per sample from the single-cell data. Absolute GEM counts per sample are listed above the x-axis. The marker is colored by the fraction of GEMs within a specificity that originate from a

given sample. Absolute GEM counts per peptide are listed to the right of the plot. The marker contains a donut diagram illustrating the distribution of clonotypes specific for the given peptide in the given sample. The wedge that represents the dominant clone is colored according to the center of the donut. Remaining clones (>1 GEM) are anthracite gray and all clonotypes only supported by one GEM only are pooled and represented by a single light gray wedge. Comparing the sizes of the T cell populations for each specificity per donor between the two screening methods in a) and b) yielded the following Spearman correlations: BC126 (1.00, $p<0.0005$), BC328 (0.90, $p=0.006$), BC355 (0.74, $p=0.02$), BC360 (0.89, $p=0.04$), BC314 (0.90, $p=0.04$), and BC353 (1.00, $p<0.0005$). c) Representative example showing the four different responses detected with fluorescent-labeled pMHC multimers in donor BC126. d) Correlation between T cell responses detected by fluorescent labeled MHC multimers (y-axis) and single-cell capturing (x-axis). Correlation is given by Pearson correlation coefficient 0.73 ($p<0.0005$). The responses from fluorescent-based screening is given as an adjusted number of cells based on the detected response frequency out of 1800 cells (see calculations in supplementary table 4). The hollow markers represent responses below detection threshold as described in a). The responses are colored by the donor-of-origin. BC mix corresponds to BC311, BC11, BC83, BC88, BC341, BC342, and BC76.

Discussion

Here, we have described and validated, ATRAP; a data-driven approach for Accurate Pairing of T cell Receptor and Antigen. We have successfully filtered single-cell 10x Genomics data to identify reliable TCR-pMHC interactions of up to 1494 clonotypes. The method can be adapted to any single-cell immune profiling data set and is highly transparent in the steps taken, allowing the user to choose appropriate stringency of filtering.

Our recommended approach of cleaning data with minimal elimination of GEMs is obtained by two sets of filters: 1) the optimized data-driven UMI thresholds combined with 2) information on matching HLA specificity (as obtained from donor-specific hashing). Increasing filtering is naturally at the expense of the number of GEMs which might reflect the trade-off between specificity and sensitivity of the assay. However, any benchmarking or validation is made difficult without a golden standard. Our best attempt at quantifying the impact of filtering is based on three metrics: annotation accuracy, binding concordance, and AUC of clonotype similarity for which ATRAP yielded the scores 96.2, 92.3, and 66.7, respectively. Evaluation of ATRAP with responses from fluorescent pMHC multimer staining revealed strong correlation (PCC=0.73, MCC=0.63) between the number of sorted T cells and the number of detected GEMs across all specificities.

Accuracy of pMHC annotation was based on selected clonotypes where the expected target was statistically distinct and UMI thresholds were set to optimize the annotation accuracy. Rare clonotypes are not considered in this metric and clones are not expected to display cross-reactivity amongst the included pMHC multimers.

The optimal UMI thresholds are intended to remove observations deviating from the expected target. The identified UMI thresholds are data specific and cannot be universally applied, but must be fitted for individual experiments. The thresholds are based on the assumption that contamination will predominantly exist at lower UMI counts than actual binding events. This limits the sensitivity of the method in cases of low-affinity low-frequency interactions which otherwise might be of great scientific and clinical interest.

The binding concordance is a metric that highlights cross-reactive clonotypes. In assays where cross-reactivity is not an expected outcome, binding concordance can be useful to evaluate the clonotypes where an expected target could not be identified. On the contrary, for data where T cell cross-recognition is of particular interest, the binding concordance can be used to establish the relative TCR binding contribution of each of the attributed pMHC targets. Growing evidence point to the relevance of T cell cross-recognition in both infectious disease (Dowell et al., 2021) and cancer (Fluckiger et al., 2020). Hence, novel tools to interrogate this phenomena on a single T cell level is highly warranted.

The last evaluation metric, AUC of clonotype similarity, is based on the assumption that T cells sharing specificity have more similar TCR sequences than T cells of different specificities (Chronister et al., 2021). This approach showed increasing separation of intra specificities and inter specificities as filters were cumulatively added, indicating that non-specific binders were effectively removed. To further increase the AUC, discarding clonotype singlets (i.e. TCR clonotypes represented by only 1 GEM) was the best single filtering step to improve the AUC of similarity scores (AUC=70.5, Fig 5a). This likely reflects that a fraction of such clonotype singlets represents non-specific binding events. However, removing these as a standard procedure of ATRAP, results in a substantial loss of TCR capture, represented by all T cell specificities with a light gray outer circle in Fig 7b. Thus, when aiming for capture of very low-frequency T cell specificities, a balance should be made between including this more stringent filtering step, or including such events, as demonstrated here.

To the best of our knowledge only one other method (ICON) has been proposed to clean TCR-pMHC single-cell data (W. Zhang et al., 2021). ICON was developed based on the public 10x Genomics data which includes six negative control pMHCs (Boutet et al., 2019), and 44 pMHC for positive selection of T cell populations. Comparing our method with ICON suggests that we present a more flexible and customizable approach. Where ICON retained ~30% of their original data (W. Zhang et al., 2021), the ATRAP method presented here allows varying yields, depending on the level of filtering applied. The optimal filtering combination of UMI thresholds and HLA matching retained ~70% of the data, while the combination of all presented filters retained ~26%. As ICON does not consider the donor haplotype information, ~15% of their specificities contained HLA mismatches, and a number of the T cell

annotations includes TCR chains of diverting clonotype definitions given by 10x Genomics. For both methods, a particular awareness should be assigned to properly handle the range of avidity displayed by different clonotypes. One clonotype may display natural low avidity towards its cognate target which might appear like noise in the comparison to other high avidity clonotypes. This diversity in signals is challenging to handle in a one-fit-all filtering process, and for projects with specific interest in low-avidity cell interactions, a specific focus should be addressed not to lose such information.

Effective pairing of TCR and pMHC will open new avenues to interrogate T cell recognition and the role of different T cell populations in pathogenic processes. Intensive efforts have been made to identify antigen specificity based on the TCR sequence (Gielis et al., 2019; Montemurro et al., 2021; Moris et al., 2021; Sidhom, Larman, Pardoll, & Baras, 2021; Weber, Born, & Rodriguez Martínez, 2021; W. Zhang et al., 2021), and access to both TCR α- and β-chain is important to improve such prediction strategies (Montemurro et al., 2021).
The coveted data is ensured via the ATRAP framework for single-cell data of TCRs and associated barcodes. The perspectives of further exploiting the transcriptomic information, allowing in-depth tracking of specific T cell subsets based on the clonotypes, suggests that we are on the verge of achieving substantial novel insight to T cell involvement and behavior in health and disease.

Acknowledgments
We would like to thank all healthy donors contributing material to this study. This research was funded in part through the Independent Research Fund Denmark (DFF 7014-00055 to S.R.H. and M.N.), the Lundbeck Foundation (R322-2019-2445 and R324-2019-1671 to A.K.B, and R190-2014-4178 to S.R.H), the European Research Council, StG 677268 NextDART to S.R.H., and National Institute of Allergy and Infectious Diseases (NIAID), under award number 75N93019C00001 to H.R.P and M.N.

AUTHOR CONTRIBUTIONS
H.R.P and A.K.B. conceived the idea, designed experiments, analyzed data, made figures and wrote the manuscript; A.K.B. and M.K performed experiments, H.R.P. and L.E.J. developed bioinformatic processing; S.R.H. and M.N. conceived the idea, supervised the study, designed experiments and analyzed data; S.R.H., M.N., M.K. and L.E.J. revised the manuscript.

COMPETING INTERESTS
A.K.B. and S.R.H. are co-inventors on a patent covering the use of DNA barcode labeled MHC multimers (WO2015185067 and WO2015188839), which is licensed to Immudex.

Methods

Ethical approval

All healthy donor material was collected under approval by the Scientific Ethics Committee of the Capital Region, Denmark, and written informed consent was obtained according to the Declaration of Helsinki.

Cell samples

Peripheral blood mononuclear cells (PBMCs) from healthy donors were isolated from whole blood by density centrifugation on Lymphoprep (Axis-Shield PoC) and cryopreserved at −150 °C in FCS (Gibco) + 10% DMSO.

DNA barcodes and dextran conjugation

Oligonucleotides modified with a 5′ biotin tag were purchased from LCG Biosearch Technologies (Denmark). Read from 5' to 3', the oligonucleotides were designed with the 10x equivalent Read2N sequence, a 10 nt unique molecular identifier (UMI), a distinct 15mer nucleotide sequences (extracted from (Xu et al., 2009), a 9 nt UMI and ending in a 13 nt capture sequence complementary to the TSO of the 10x 5' capture oligo (sequences are listed in Supplementary Table 1). Barcodes were dissolved to 100 µM in RNAse free water and stored at −20 °C. For a working solution the DNA barcodes were further diluted in PBS + 0.5% BSA + 1 mg/mL herring DNA + 2 mM EDTA to 2.17 µM and attached to PE- or APC- and streptavidin-conjugated dextran from FINA Biosolutions LCC (USA). The amount of DNA barcode attached to each new lot of dextran was titrated as described in Bentzen et al., 2016. DNA barcodes were attached by mixing with dextran-conjugate, followed by incubation, 30 min at 4 °C. DNA barcode-assembled dextran-conjugates were stored for up to 24 hours at 4 °C.

Peptides and MHC monomer production

Peptides were purchased from Pepscan (Pepscan Presto) and dissolved to 10 mM in DMSO. UV-sensitive ligands were synthesized as previously described (Bakker et al., 2008; Rodenko et al., 2006; Toebes et al., 2006). Recombinant HLA-A*0201, HLA-A*0301 and HLA-B*0702, heavy chains and human β2 microglobulin light chain were produced in Escherichia coli. HLA heavy and light chains were refolded with UV-sensitive ligands and purified as described in (Hadrup et al., 2009). Specific peptide-MHC complexes were generated by UV-mediated peptide MHC exchange (Chang et al., 2013; Frøsig et al., 2015; Rodenko et al., 2006; Toebes et al., 2006).

Generation of DNA barcode-labeled MHC multimer libraries

Unoccupied SA-binding sites on the DNA barcode-assembled dextran conjugates were used for the co-attachment of biotinylated pMHC molecules. 0.8 pmol pMHC

monomer was mixed with 160 × 10^{-15} mol DNA-barcoded dextran-conjugate and incubated 30 min at RT. MHC multimers were diluted in PBS with 5.2 µM d-biotin (Avidity, Bio200) to 909 nM and incubated 20 min on ice. DNA-barcoded MHC multimers were stored for up 1 week at −20 °C (PBS supplemented with glycerol and BSA, final concentrations 5% and 0.5%, respectively). Immediately before staining barcode-labeled MHC multimers were thawed at 4 °C, centrifuged (5 min at 3,300g), and pooled (0.8 pmol of each pMHC/sample) to enable the detection of multiple T-cell responses in parallel. The pooled MHC multimers were centrifuged once more; 5 min at 3,300g, to sediment aggregates before the volume of the reagent pool was reduced by ultrafiltration to obtain a final volume of ~80 µL of MHC multimers as described in Bentzen et al., 2016. Any aggregates in the MHC multimer reagent pool were sedimented by centrifugation, 5 min at 3,300g before addition to the cell sample.

MHC multimer staining

Cryopreserved PBMCs were thawed and washed by sedimentation, 5 min, 390*g*, 4 °C, in RPMI + 10% FCS. Cells were further washed in a barcode-cytometry buffer (PBS + 0.5% BSA). 5 × 10^6 cells were incubated, 60 min, 4 °C, with pooled DNA-barcoded multimers in a total volume of 100 µL (final concentration of each distinct pMHC, 8 nM), and washed three times by sedimentation, 5 min, 390*g*, 4 °C. 5 µl of Human TruStain FcX™ Fc Blocking reagent was added to a total of 50 µl cell suspension, and incubated 10 min, 4 °C. Hashing antibodies (Biolegend, TotalSeq™-C0251 anti-human Hashtag 1-10 Antibodies) were centrifuged 10 min, 14,000 x *g*, 4 °C, and 0.5 µl were added to each a distinct sample (Supplementary table 2), and incubated 15 min, 4 °C. Next a 5× antibody mix composed of CD8-BV480 (BD 566121, clone RPA-T8) (final dilution 1/50), dump channel antibodies: CD4-FITC (BD 345768) (final dilution 1/80), CD14-FITC (BD 345784) (final dilution 1/32), CD19-FITC (BD 345776) (final dilution 1/16), CD40-FITC (Serotech MCA1590F) (final dilution 1/40), CD16-FITC (BD 335035) (final dilution 1/64) and a dead cell marker (LIVE/DEAD Fixable Near-IR; Invitrogen L10119) (final dilution 1/1000) was mixed for each sample. The antibody mix was added to cell samples and incubated 30 min, 4 °C. Cells were washed three times in barcode-cytometry buffer and kept on ice until acquisition.

Cell sorting

Cells were sorted on a FACS Melody (BD) into tubes containing 100 µL of PBS + 0.5% BSA (tubes were saturated with PBS + 2% BSA in advance). Using FACS Chorus software, we gated on single, live, CD8-positive and 'dump' (CD4, 14, 16, 19 and 40)-negative lymphocytes and sorted only APC-positive (PE-negative) cells within this population (Supplementary Fig 4 for gating strategy). Cells sorted from individual samples were collected into the same tube (Fig 1b). The sorted cells were centrifuged for 10 min at 390*g* and the buffer was removed.

DNA barcode-labeled MHC multimer stained cells on 10x platform

We utilize the 10x 5’ v2 chemistry that allows the cell barcode to be appended at the 5’-end of transcripts, which is essential for capturing the CDR3 region of the V(D)J transcripts. In the 5’ chemistry, the template switch oligo (TSO) is encoded with a cell barcode i.e., one unique 10x barcode for every Gel Bead-in-emulsion (GEM). The TSO thus comprises the capture oligo, whereas the poly-dT primer is added in free suspension. Reverse transcription is initiated from binding of the poly-dT primer at the 3’-end, and mRNA is captured when the reverse transcriptase enzyme switches at the 5’-end of the transcript to the TSO. All DNA barcodes, partially complementary to the 10x Genomics 5’ TSO, are captured directly onto the GEMs. Annealing and extension during the reverse-transcription reaction associates the cell barcode and unique molecular identifier (UMI) from the gel bead oligo with the pMHC and hashing antibody tags in parallel with the mRNAs in the same droplet.
Downstream processing of mRNA and DNA barcodes are performed according to manufacturer's instructions (Chromium Next GEM Single Cell 5' Reagent Kits v2 (Dual Index), with the Feature Barcode technology for Cell Surface Protein & Immune Receptor Mapping) (10x Genomics, USA). ~15,700 cells were loaded (based on 55% recovery from 28,800 sorted cells) to yield a maximum of 9,000 cells with an intermediate/high doublet rate (6,9%). Targeted amplification was performed for 13 cycles and the products were separated according to size into <400 bp (DNA barcode-tags) and >400 bp (the TCR sequences) using 0.6x SPRIselect beads (Beckman Coulter, B23318). Separate processing of the >400 bp bead-bound TCR sequences and the <400 bp in solution DNA barcodes was conducted according to manufacturer's instruction and the TCR amplification products were sequenced on a NovaSeq running a 150 paired-end program. DNA barcodes, TCR sequences and mRNA was sequenced to a depth of 13,332, 12,503, and 18,398 mean reads per cell, respectively.

Bioinformatics

Processing of 10x single-cell data

Hashing barcode reads, peptide-MHC barcode reads, and T cell gene expression reads, were provided in fastq format and were processed using 10x Genomics Cellranger multi v6.1.0 (10xGenomics, n.d. b). The relevant outputs were the unfiltered count matrices of DNA barcodes and gene expression as well as clonotype annotations of each sequencing contig containing CDR3α/β sequences, V(D)J-C genes and unique molecular identifier (UMI) counts.

Postprocessing 10x Cellranger clonotyping

The raw contig annotations from Cellranger were selected for downstream analysis with filtering on incomplete and unproductive receptor transcripts. Incomplete contigs

are not full length i.e., do not span the V-gene start codon until the J-gene stop codon. Unproductive contigs contain a frameshift which either induces an early stop codon or completely removes the stop codon.

Clonotypes defined by 10x were merged when consisting of identical VJ-CDR3αβ, thus reducing functional duplicates.

Cellranger flags rare nucleotide transcripts as likely artifacts, meaning the GEMs are flagged as unlikely to contain a cell and are therefore not assigned a clonotype (10xGenomics, n.d. a). Therefore, GEMs that were not annotated with a clonotype were imputed by searching the duplicate-reduced clonotype set. If no match, a new clonotype ID was annotated to the GEM.

Filtering based on gene expression

Filtering on gene expression data was performed as described in (W. Zhang et al., 2021). Low-quality GEMs such as doublets may be removed by excluding GEMs with more than 2500 genes. Dead cells may be removed by excluding GEMs with fewer than 200 genes and a ratio of mitochondrial gene expression to the total gene expression above 0.2.

Demultiplexing samples via cell hashing

Cell Hashing uses oligo-tagged antibodies against ubiquitously expressed surface proteins to place a sample barcode on each single cell, enabling different samples to be multiplexed together and run in a single experiment. To demultiplex the samples the method presented by Stoeckius et. al was implemented (Stoeckius et al., 2018). The method clusters the normalized count matrix using k-medoid clustering into k clusters, $k = n_{samples} + 1$. For each barcode a negative binomial distribution is fitted to the pool of all clusters except the cluster with the highest average expression for the given barcode. Each GEM is classified as positive if the barcode value exceeds a 0.99 quantile threshold for the negative distribution, and otherwise classified as negative. If GEMs contain multiple barcodes which pass the threshold, the GEM is annotated as a doublet (Stoeckius et al., 2018).

Defining the expected binder

The pMHC and cell hashing barcode annotations were merged with the T cell annotations on the GEMs which contained both TCR and pMHC attributes. Each clonotype is expected to have a preferred target within the pMHC library, thus each clonotype is evaluated to find the pMHC which is most likely to be that target. Each clonotype is evaluated to identify the expected target within the pMHC library. The pMHCs that are detected within the GEMs annotated to a given clonotype are compared by their UMI count distribution. The two pMHCs that have the highest mean UMI count are compared, testing the hypothesis that the expected binder will have a significantly higher mean UMI count than the other pMHC (Wilcoxon,

α=0.05). Clonotypes of less than 10 GEMs were not tested. The clonotypes where the mean UMI of the top two pMHCs was significantly different were collected as a training set. The pMHC which had significantly higher mean UMI was annotated as the expected target and specificity annotations of the GEMs were individually evaluated. The GEMs in the training set where the most abundant pMHC matched the expected target were labeled as true interactions, and the rest were labeled as false interactions.

Defining specificity concordance

Concordance is an indirect measure of how cross-reactive a certain clonotype is. Specificity concordance is defined as the ratio of GEMs of a single clonotype which are annotated to bind a particular pMHC. The more GEMs in a clonotype annotated to the same pMHC the larger concordance. If a clonotype is only detected with one pMHC the specificity concordance is 1.

Grid search on UMI features

Based on the labels of the training set a performance metric, o, was defined to evaluate the accuracy at increasing thresholds for UMI count and UMI ratio of pMHC, α-chain, and β-chain. The UMI ratio measures multiplets and is defined as the ratio between the highest UMI count and the second highest UMI count in a GEM:

$$UMI_{ratio} = \frac{UMI_{max}}{UMI_{sec} + 0.25}$$

A small number (0.25) was added in the denominator to avoid division by zero.

The performance metric, o, is a weighted average of accuracy and fraction of retained GEMs, given by the following equation:

$$o = \frac{2 \cdot acc + f_{retained\ GEMs}}{3}$$

The accuracy metric is defined by the ratio of training set GEMs that were labeled as true interactions over the total number of GEMs in the training set. The performance metric, o, was maximized by finding the set of filters that increase the accuracy without excluding too much data.
The thresholds for filtering were selected from a complete grid search. Each feature was tested in the range of 0 to the median value, determined ad hoc from the experience that thresholds never approached the median value.

Comparing TCR similarity

Effects of filtering were also evaluated through a comparison of TCR similarity. The similarity score is based on the kernel similarity score underlying the TCRmatch method between CDR3 sequences (Chronister et al., 2021). This score can be calculated for CDR3s of variable length, and takes a value between 0-1, with the value of 1 representing identical pairs. As both the α- and β-chain partake in the pMHC interaction, TCRs will be compared based on the summed similarity between the α- and β-chains, and GEMs missing a chain will be excluded to avoid bias. Two similarity scores are computed for each clonotype: an intra score and an inter score. The intra score is based on the maximum similarity of the given clonotype to all other clonotypes sharing its pMHC target (intra specificity). The inter score is based on the maximum similarity of the given clonotype to an equal sized set of clonotypes specific to other pMHC targets (inter specificity). The computation is done peptide-wise, such that clonotypes with maximum concordance for a given peptide will, for that peptide, be included in an intra similarity score, but for another peptide be included in an inter similarity score. Clonotypes consisting of GEMs causing diverging specificities were limited to the expected target pMHC or, if non-existing, to the specificity of highest concordance to avoid potential overlaps from “cross-reactive” clonotypes in the computation.

The similarity difference between intra and inter specificity clonotypes was tested for the hypothesis that intra similarity is greater than inter similarity (Wilcoxon, $\alpha=0.05$). The similarity test was performed on all filtering methods described in the paper.

Validating single-cell capture against fluorescent multimer staining responses

The 16 donors were known to respond to the panel of peptides used in the screening. Response proportions of sorted CD8+ T cells were detected by fluorescent multimer staining, as described previously. 1800 cells were selected from each donor and, based on the detected response proportions, an adjusted count of cells could be computed. Cells were selected based on two criteria: ≥10 cells and ≥ 0.002% of total CD8 T cells, or ≤10 cells but ≥0.01% of total CD8 T cells. The multimer responses were compared to GEMs filtered on UMI thresholds and matching HLA. To visually compare the two screening methods, the responses were normalized within each sample and plotted side-by-side. The methods were also quantitatively compared, both in absolute counts of responses and as binary classes with multimer responses as true labels and single-cell responses as query labels. The following evaluation metrics were computed based on binary classification of responses:

Matthew’s correlation coefficient (MCC)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Recall (sensitivity or true positive rate)

$$recall = \frac{TP}{TP + FN}$$

Precision (positive predictive value)

$$precision = \frac{TP}{TP + FP}$$

The following metrics were computed based on numerical values of responses, i.e. the adjusted number of cells and the count of GEMs.

Pearson correlation coefficient (PCC), where n is the number of screened peptides times the number of samples which have been screened. x and y represent the size of responses in single-cell screening and multimer staining, respectively.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \underline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \underline{y})^2}}$$

The correlation was also fitted via linear regression on log10 transformed data, resulting in the following equation.

$$log_{10}(y) = 0.86 \cdot log_{10}(x) + 1.18, R^2 = 0.56$$

The equation was used to estimate the yield of single-cell captured cells relative to multimer screening. Three examples were computed to estimate an approximate 10% yield.

$$log_{10}(10) = 0.86 \cdot log_{10}(0.6) + 1.18$$
$$log_{10}(100) = 0.86 \cdot log_{10}(9.0) + 1.18$$
$$log_{10}(1000) = 0.86 \cdot log_{10}(131.2) + 1.18$$

References

10xGenomics. (n.d.-a). Cell Ranger Installation -Software -Single Cell Immune Profiling - Official 10x Genomics Support. Retrieved July 12, 2022, from https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/installation

10xGenomics. (n.d.-b). V(D)J Cell Calling Algorithm -Software -Single Cell Immune Profiling -Official 10x Genomics Support. Retrieved July 12, 2022, from https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/algorithms/cell-calling

Acha-Orbea, H., Mitchell, D. J., Timmermann, L., Wraith, D. C., Tausch, G. S., Waldor, M. K., … Steinman, L. (1988). Limited heterogeneity of T cell receptors from lymphocytes mediating autoimmune encephalomyelitis allows specific immune intervention. *Cell*, *54*(2), 263–273. https://doi.org/10.1016/0092-8674(88)90558-2

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., & Kourilsky, P. (1999). A direct estimate of the human alphabeta T cell receptor diversity. *Science (New York, N.Y.)*, *286*(5441), 958–961. https://doi.org/10.1126/SCIENCE.286.5441.958

Bagaev, D. V., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., … Shugay, M. (2020). VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, *48*(D1), D1057–D1062. https://doi.org/10.1093/NAR/GKZ874

Bakker, A. H., Hoppes, R., Linnemann, C., Toebes, M., Rodenko, B., Berkers, C. R., … Schumacher, T. N. M. (2008). Conditional MHC class I ligands and peptide exchange technology for the human MHC gene products HLA-A1, -A3, -A11, and -B7. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(10), 3825–3830. https://doi.org/10.1073/PNAS.0709717105/SUPPL_FILE/09717FIG7.JPG

Bentzen, A. K., Marquard, A. M., Lyngaa, R., Saini, S. K., Ramskov, S., Donia, M., … Hadrup, S. R. (2016). Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nature Biotechnology 2016 34:10*, *34*(10), 1037–1045. https://doi.org/10.1038/NBT.3662

Bergman, R. (1999). How useful are T-cell receptor gene rearrangement studies as an adjunct to the histopathologic diagnosis of mycosis fungoides? *The American Journal of Dermatopathology*, *21*(5), 498–502. https://doi.org/10.1097/00000372-199910000-00019

Bloom, J. D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, *2018*(9). https://doi.org/10.7717/PEERJ.5578/SUPP-4

Boutet, S. C., Walter, D., Stubbington, M. J. T., Pfeiffer, K. A., Lee, J. Y., Taylor, S. E. B., … Mikkelsen, T. S. (2019). Scalable and comprehensive characterization of antigen-specific CD8 T cells using multi-omics single cell analysis. *The Journal of Immunology*, *202*(1 Supplement).

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology 2014 33:2*, *33*(2), 155–160. https://doi.org/10.1038/nbt.3102

Chang, C. X. L., Tan, A. T., Or, M. Y., Toh, K. Y., Lim, P. Y., Chia, A. S. E., … Grotenbreg, G. M. (2013). Conditional ligands for Asian HLA variants facilitate the definition of CD8+ T-cell responses in acute and chronic viral diseases. *European Journal of Immunology*, *43*(4), 1109–1120. https://doi.org/10.1002/EJI.201243088

Chronister, W. D., Crinklaw, A., Mahajan, S., Vita, R., Koşaloğlu-Yalçın, Z., Yan, Z., … Peters, B. (2021). TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. *Frontiers in Immunology*, *12*, 673. https://doi.org/10.3389/FIMMU.2021.640725/BIBTEX

Davis, M. M., & Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition.

Nature, *334*(6181), 395–402. https://doi.org/10.1038/334395A0

Dowell, A. C., Butler, M. S., Jinks, E., Tut, G., Lancaster, T., Sylla, P., … Ladhani, S. (2021). Children develop robust and sustained cross-reactive spike-specific immune responses to SARS-CoV-2 infection. *Nature Immunology 2021 23:1*, *23*(1), 40–49. https://doi.org/10.1038/s41590-021-01089-8

Dziubianau, M., Hecht, J., Kuchenbecker, L., Sattler, A., Stervbo, U., Rödelsperger, C., … Babel, N. (2013). TCR Repertoire Analysis by Next Generation Sequencing Allows Complex Differential Diagnosis of T Cell–Related Pathology. *American Journal of Transplantation*, *13*(11), 2842–2854. https://doi.org/10.1111/AJT.12431

Elliott, J. I., & Altmann, D. M. (1995). Dual T cell receptor alpha chain T cells in autoimmunity. *The Journal of Experimental Medicine*, *182*(4), 953. https://doi.org/10.1084/JEM.182.4.953

Fluckiger, A., Daillère, R., Sassi, M., Sixt, B. S., Liu, P., Loos, F., … Zitvogel, L. (2020). Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. *Science*, *369*(6506), 936–942. https://doi.org/10.1126/SCIENCE.AAX0701/SUPPL_FILE/AAX0701_FLUCKIGER_SM.PDF

Frøsig, T. M., Yap, J., Seremet, T., Lyngaa, R., Svane, I. M., Thor Straten, P., … Hadrup, S. R. (2015). Design and validation of conditional ligands for HLA-B*08:01, HLA-B*15:01, HLA-B*35:01, and HLA-B*44:05. *Cytometry Part A*, *87*(10), 967–975. https://doi.org/10.1002/CYTO.A.22689

Gaublomme, J. T., Li, B., McCabe, C., Knecht, A., Yang, Y., Drokhlyansky, E., … Regev, A. (2019). Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nature Communications 2019 10:1*, *10*(1), 1–8. https://doi.org/10.1038/s41467-019-10756-2

Gerlach, C., Rohr, J. C., Perié, L., Van Rooij, N., Van Heijst, J. W. J., Velds, A., … Schumacher, T. N. M. (2013). Heterogeneous differentiation patterns of individual CD8+ T cells. *Science (New York, N.Y.)*, *340*(6132), 635–639. https://doi.org/10.1126/SCIENCE.1235487

Gielis, S., Moris, P., Bittremieux, W., De Neuter, N., Ogunjimi, B., Laukens, K., & Meysman, P. (2019). Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology*, *10*, 2820. https://doi.org/10.3389/FIMMU.2019.02820/BIBTEX

Hadrup, S. R., Toebes, M., Rodenko, B., Bakker, A. H., Egan, D. A., Ovaa, H., & Schumacher, T. N. M. (2009). High-throughput T-cell epitope discovery through MHC peptide exchange. *Methods in Molecular Biology (Clifton, N.J.)*, *524*, 383–405. https://doi.org/10.1007/978-1-59745-450-6_28

Hou, X., Wang, M., Lu, C., Xie, Q., Cui, G., Chen, J., … Diao, H. (2016). Analysis of the Repertoire Features of TCR Beta Chain CDR3 in Human by High-Throughput Sequencing. *Cellular Physiology and Biochemistry : International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, *39*(2), 651–667. https://doi.org/10.1159/000445656

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods 2014 11:7*, *11*(7), 740–742. https://doi.org/10.1038/nmeth.2967

Kirsch, I. R., Watanabe, R., O'Malley, J. T., Williamson, D. W., Scott, L. L., Elco, C. P., … Clark, R. A. (2015). TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Science Translational Medicine*, *7*(308). https://doi.org/10.1126/SCITRANSLMED.AAA9122

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. https://doi.org/10.1038/NMETH.1778

Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., … Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Research*, *24*(10),

1603–1612. https://doi.org/10.1101/GR.170753.113
Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., … Nielsen, M. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology*, *4*(1), 1–13. https://doi.org/10.1038/s42003-021-02610-3
Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., … Meysman, P. (2021). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/BIB/BBAA318
Petrie, H. T., Livak, F., Schatz, D. G., Strasser, A., Crispe, I. N., & Shortman, K. (1993). Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *The Journal of Experimental Medicine*, *178*(2), 615. https://doi.org/10.1084/JEM.178.2.615
Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., … Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, *114*(19), 4099–4107. https://doi.org/10.1182/BLOOD-2009-04-217604
Rodenko, B., Toebes, M., Hadrup, S. R., van Esch, W. J. E., Molenaar, A. M., Schumacher, T. N. M., & Ovaa, H. (2006). Generation of peptide–MHC class I complexes through UV-mediated ligand exchange. *Nature Protocols 2006 1:3*, *1*(3), 1120–1132. https://doi.org/10.1038/nprot.2006.121
Shen, W.-J., Wong, H.-S., Xiao, Q.-W., Guo, X., & Smale, S. (2012). *Towards a Mathematical Foundation of Immunology and Amino Acid Chains*. https://doi.org/10.48550/arxiv.1205.6031
Sherwood, J. (2013). Colonisation - it's bad for your health: the context of Aboriginal health. *Contemporary Nurse*, *46*(1), 28–40. https://doi.org/10.5172/CONU.2013.46.1.28
Sidhom, J. W., Larman, H. B., Pardoll, D. M., & Baras, A. S. (2021). DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications 2021 12:1*, *12*(1), 1–12. https://doi.org/10.1038/s41467-021-21879-w
Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., … Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/S13059-018-1603-1/FIGURES/3
Toebes, M., Coccoris, M., Bins, A., Rodenko, B., Gomez, R., Nieuwkoop, N. J., … Schumacher, T. N. M. (2006). Design and use of conditional MHC class I ligands. *Nature Medicine*, *12*(2), 246–251. https://doi.org/10.1038/NM1360
Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., … Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/NAR/GKY1006
Weber, A., Born, J., & Rodriguez Martínez, M. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, *37*(Suppl 1), i237. https://doi.org/10.1093/BIOINFORMATICS/BTAB294
Yamawaki, T. M., Lu, D. R., Ellwanger, D. C., Bhatt, D., Manzanillo, P., Arias, V., … Li, C. M. (2021). Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics*, *22*(1), 1–18. https://doi.org/10.1186/S12864-020-07358-4/FIGURES/8
Zhang, S. Q., Ma, K. Y., Schonnesen, A. A., Zhang, M., He, C., Sun, E., … Jiang, N. (2018). High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nature Biotechnology 2018 36:12*, *36*(12), 1156–1159. https://doi.org/10.1038/nbt.4282
Zhang, W., Hawkins, P. G., He, J., Gupta, N. T., Liu, J., Choonoo, G., … Atwal, G. S. (2021). A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances*, *7*(20). https://doi.org/10.1126/SCIADV.ABF5835
Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J.

H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8:1*, *8*(1), 1–12. https://doi.org/10.1038/ncomms14049

CHAPTER 6

Benchmarking data-driven filtering approaches for single-cell screening of T cell specificity

To truly benefit from the costly singe-cell assay immunoinformatic methods are crucial to balance a satisfying ratio of signal-to-noise with proper yield. The field of de-noising single-cell data composed of T cell specificities is new and inexperienced. Currently, only two frameworks have been developed. Both methods are data-driven, but rely on different aspects of the data to filter away presumed techical artifacts. This chapter presents a benchmark of the two methods and highlights advantages and disadvantages of each approach.

Benchmarking data-driven filtering approaches for single-cell screening of T cell specificity

Helle Rus Povlsen, Morten Nielsen
Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract

Pairing of T cell receptor (TCR) with its cognate peptide-MHC (pMHC) is a cornerstone in T cell mediated immunity. Recently, single-cell sequencing coupled with DNA-barcoded multimer staining has made high-throughput study of T cell specificity available. However, the immense variability of the TCR-pMHC interaction combined with the technology's low ratio of signal-to-noise is complicating the study. Here we present a benchmark of two computational frameworks, ICON and ATRAP, for de-noising single-cell specificity data. As no golden standard exist, the methods are evaluated on the publicly available immune profiling data provided by 10x Genomics by metrics developed for the purpose. The key difference between the two frameworks is the balance of specificity and sensitivity.

Introduction

The specificity of T cells form the hallmark of cellular immunity. T cell specificity is determined by a triad of interactions between the T cell receptor (TCR), a peptide (p), and its restricting major histocompatibility complex (MHC). The TCR is a heterodimeric protein, typically composed of an α- and β-chain, which are formed during T cell development as a result of stochastic V(D)J gene recombination [222–226]. As a result of the somatic recombination, highly variable joining segments are introduced, facilitating a diverse TCR repertoire which ensures protection from a broad and ever-changing range of pathogens or cancerous mutations [3, 227, 228]. The joining segments are contained in a region known as

the complementarity determining region 3 (CDR3). CDR1 and CDR2 reside in highly polymorphic regions of the V-gene. The three CDRs form flexible loops of the TCR which engage with the peptide-MHC (pMHC) complex and thereby determine the specificity of the T cell [229–232].

Recent studies have elucidated common TCR sequence features of T cells that share specificity, and for selected pMHCs it has been possible to predict the binding probability to TCRs novel to the trained model [127, 173, 177, 178, 184, 187, 191, 192, 204]. The current primary limitation is lack of both quantity and diversity of training data generated by traditional specificity assays such as multimer sorting and reexposure assays, followed by bulk sequencing of typically the TCRβ-chain. However, the advent of single-cell sequencing platforms promises high-throughput data which in addition intrinsically provides information of false binding pairs as well as true pairs [179]. This type of data is expected accelerate the understanding of TCR specificity.

10x Genomics has specifically developed an immune profiling platform that couples TCR sequencing of both α- and β-chains with DNA barcoded peptide-MHC (pMHC) multimers, DNA barcoded surface marker antibodies, and DNA barcoded cell hashing antibodies. The platform is designed to capture a single cell together with a gel-bead in emulsion (GEM) [142, 143]. Each GEM contains GEM-specific barcoded primers which ensures back-tracing of transcripts to the cell-of-origin. As the platform promises single-cell capture, the contents of a GEM should reflect a single cell and its associated barcoded analytes, hence GEM and cell may be used interchangeably. The GEM primers also contain a unique molecular identifier (UMI) which ensures quantification of transcripts unbiased by PCR amplification [146]. Thus, single-cell screening of TCR-pMHC interactions yield the TCRαβ sequence and the expression level of both chains as well as count of each unique pMHC binding which might be interpreted as T cell avidity [179].

In 2019, 10x Genomics released a large, state-of-the-art data set [179] which spurred activity within the TCR-pMHC modeling community [127, 173, 174, 177, 187, 192]. However, the new data presented new challenges. The single-cell platform is generally as-

sociated with a poor signal-to-noise ratio, which also affects this specificity data. The challenge was handled in various ways. In NetTCR-2.0, the data was utilized to define negative TCR-pMHC pairs, i.e. pairs that were not detected to bind any of the investigated pMHC complexes and thereby avoided handling the noise within the detected binders [173]. Since the true TCR-pMHC pairs are a point of contention, the authors of ImRex purposefully omitted the 10x data [187], while the authors of TcellMatch and DeepTCR relied on the network to extract the salient pMHC specific features of the TCRs [174, 177]. The authors of TCRAI were the first to develop a computational method, named ICON (Integrative COntext-specific Normalization), to discriminate true TCR-pMHC binding signal from nonspecific background noise [127]. Recently, we have proposed an alternative framework ATRAP (Accurate T cell and Antigen Pairing) [233]. These methods might pave the way for improved specificity models.

ICON was developed based on the public 10x Genomics data which contains T cell specificities from four healthy donors screened against a panel of 50 pMHCs which includes 44 pMHCs for positive selection and six negative control pMHCs [179]. ICON utilizes the negative controls to empirically estimate the background binding noise per donor. The UMI counts of pMHCs were then corrected by subtracting the donor-specific estimated background noise. UMI counts were further corrected by penalizing pMHCs multiplets i.e., GEMs containing multiple DNA barcodes corresponding to two or more different pMHCs. The final step of ICON is normalization of UMI counts across pMHCs and GEMs to make them directly comparable. As a result, ICON identified a total of 53,062 T cells belonging to to 5,722 unique clonotypes.

ATRAP takes a different approach. The framework was developed and tested on in-house single-cell data generated using the 10x Genomics platform similar to the public 10x Genomics data. The ATRAP framework consists of a series of filtering approaches to obtain increasingly accurate TCR-pMHC pairing. The first filtering step was based on identifying expected targets by comparing the UMI distributions of all pMHCs detected within a clonotype consisting of 10 or more GEMs. The key is to study GEMs in en-

semble rather than individually, because deviations are averaged out. If a pMHC was distributed with a significantly higher mean UMI in the ensemble, we expected this pMHC to reflect the true target of the clonotype, collectively providing a golden standard. Based on the labeling of true and false targets, we could compute an accuracy score. Thresholds were set on UMI counts to maximize the accuracy. By globally applying the optimal threshold, the remaining clonotypes should ideally represent the same level of accuracy in their pMHC annotations. Another key step of ATRAP filtering is ensuring HLA correspondence between pMHC and the HLA haplotype of the T cell donor. In immune profiling assays the option to hash cells by donor-of-origin enables assignment of HLA haplotype restriction to each cell. Correspondence between allele of pMHC and donor haplotype can be used to verify assignment of the pMHC, assuming that a T cell is absolutely restricted to the allele for which it was selected during the thymocyte maturation process. In the public 10x data, the cells are not hashed, however, the experiment was run in parallel for each donor, enabling *in silico* hashing of the individual single-cell runs.

In this study, we will report a benchmark of the two frameworks to recommend future application of single-cell specificity data. Both methods will be evaluated on the 10x Genomics data since this is the only data set containing negative controls as is required by ICON. As no external golden-standard exist, the methods will be evaluated on metrics presented by Povlsen et al., [233]: GEM retention, accuracy, average binding concordance, and AUC of similarity scores. The fraction of retained GEMs simply quantifies how many observations are removed by a filter. Accuracy measures the proportion of GEMs where highest abundance pMHC annotation corresponds to the expected target of large clonotypes (>10 GEMs). Average binding concordance is a measure of target dispersal within a clonotype, the more different pMHCs detected as highest abundance pMHC per clonotype, the smaller average concordance. In this metric clonotype singlets are omitted since they inherently have a binding concordance of 1, but with low certainty of whether the specificity is reliable. These clonotype singlets may however be included in the AUC of similarity scores if they contain both α- and β-chain. The AUC metric discloses how

well intra-specificity TCRs can be discerned from inter-specificity TCRs by similarity scores.

Results

Summary of the public 10x data

In the public data set made available by 10x Genomics, a total of 181,913 GEMs were detected containing at least one TCR-pMHC pair. The data set is the result of screening $CD8^+$ T cells from four healthy donors against a panel of 50 pMHC DNA barcode labeled multimers. Donors were selected by HLA haplotype to ensure overlap with the HLA alleles of the pMHC panel. 44 of the multimers contain antigenic peptides derived from CMV, EBV, influenza, HTLV, HPV, HIV and known cancer antigens. Note, that the donors were all seronegative for HIV, HBV, and HBC. The remaining six multimers contained negative control peptides restricted by five HLAs, selected without further elaboration or reasoning. The specificities from each of the four donors were screened in parallel i.e., of four different experimental runs. Therefore, unique GEM-specific 10x barcodes (GEM barcode) were in some cases observed in replicas across runs. In order to distinguish these evidently distinct GEMs, an extra suffix was added denoting the donor (sample ID). The unfiltered output is portrayed in figure 6.1, which clearly demonstrates the issue of noise, as very GEM contains multiple pMHCs. Most GEMs contain TCRs annotated with a unique α- and β-chain, however, 10% are annotated with multiple α- or β-chains, which further challenges the investigation of specificity.

Alignment of ICON- and 10x-assigned GEMs revealing inconsistent annotations

In order to compare the two filtering frameworks, the outputs from each were aligned based on the GEM barcode, consisting on 16 nucleotides, a suffix pertaining to the sequencing well, and a sample ID suffix. ICON report retention of 53,062 GEMs out of the total set of 181,913 GEMs. However, ICON only contains 5031 GEMs that matches the original data based on the full GEM

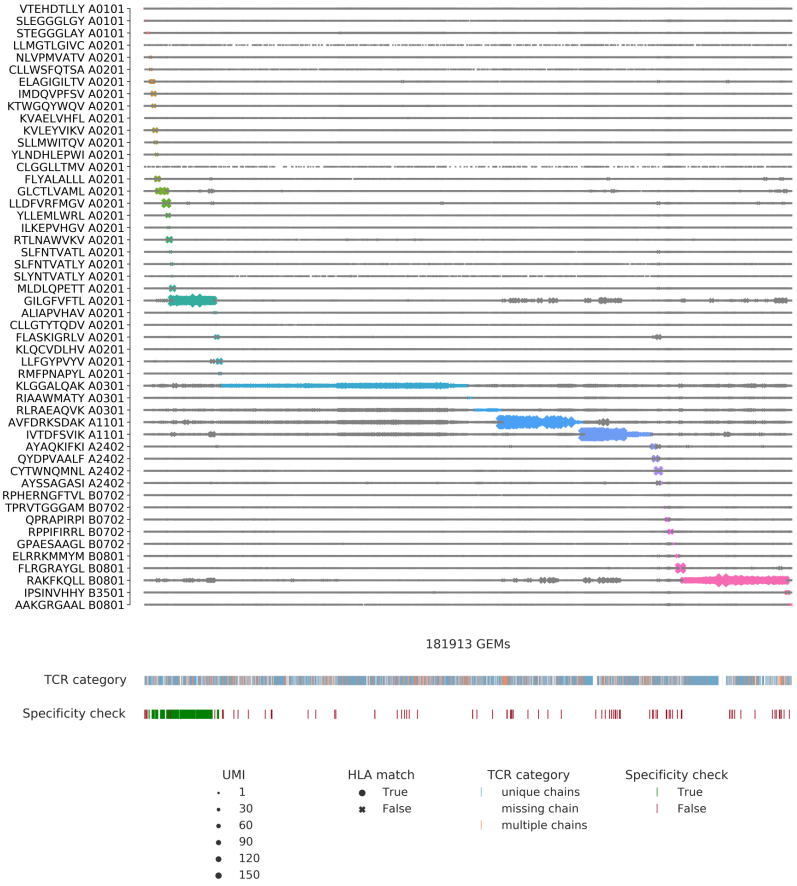


Figure 6.1: Scatterplot of all detected pMHC barcodes (y-axis) within each of the 181,913 GEMs (x-axis). In each GEM the most abundant pMHC is marked by a color, while the remaining pMHCs in the GEM are gray. The marker size reports the UMI count of the given pMHC and the shape recounts whether the HLA allele of the pMHC matches the HLA haplotype of the donor, which is provided in the experimental report [126]. The fraction of HLA matches within the GEMs displaying a given specificity is annotated to the right of the plot. The first colorbar indicates the type of TCR chain annotation; whether the TCR has a unique αβ-pair, is missing a chain or consists of multiple chains. The second colorbar is a specificity check against the specificity databases IEDB and VDJdb. Colors highlight the GEMs where the CDR3αβ sequences are contained in the databases. The green color represents a match between the database pMHC and the detected pMHC, while red indicates a mismatch.

barcode, due to inconsistencies in the suffix annotation. When stripping the barcode down to only the 16 nucleotides, we were able to align 39,806 GEM barcodes, as exemplified in figure 6.2a. We also observed inconsistencies of TCRαβ annotations in 3391 GEMs, as illustrated in figure 6.2b+c. 1854 GEMs were missing either an α- or a β-chain in the 10x data, but not in the ICON set, while 1537 GEMs were fully annotated, but had inconsistent TCR annotations between ICON and the 10x data. The inconsistencies in TCRαβ annotations may have arisen from imputations based on the 10x-provided clonotype summary. However, imputation is risky because the same CDR3 may form part of several different clonotypes. The example given in 6.2b represents an imputation likely based on the CDR3β sequence. In this example the CDR3β sequence is part of 42 distinct 10x clonotypes, all carrying the same CDR3β sequence, but paired with different CDR3α sequences. The same case is made for 6.2c and all the other inconsistent GEMs. Imputation by 10x clonotypes is further made difficult as their clonotype definition actually allows multiple α- or β-chains in one clonotype, perhaps a reflection of incomplete allelic exclusion. Thus, 116 of the fully annotated GEMs with mismatching TCRαβ annotations between ICON and 10x can be explained by a switch from one chain to the other, still within the same clonotype definition. This non-conformity has challenged the benchmark, however, we have proceeded assuming that there is a reasonable, however undocumented, explanation for their GEM assignments.

ATRAP - Revisiting clonotype assignment

For efficient utilisation of ATRAP, the 10x assigned clonotypes were redefined. The original annotations of clonotypes were based on unique nucleotide sequences of the T cell receptor to identify expansions of clonally related T cells. However, the somatic pedigree is not relevant to understand the biochemical properties of the TCR. In stead, we are interested in grouping T cells of TCRs with identical amino acid sequence including identical CDR3 sequences. This regrouping of GEMs results in larger clonotypes beneficial for statistical power. Thus, in ATRAP a clonotype is defined by a unique set of Vαβ- and Jαβ-genes as wells as CDR3αβ. Note, that

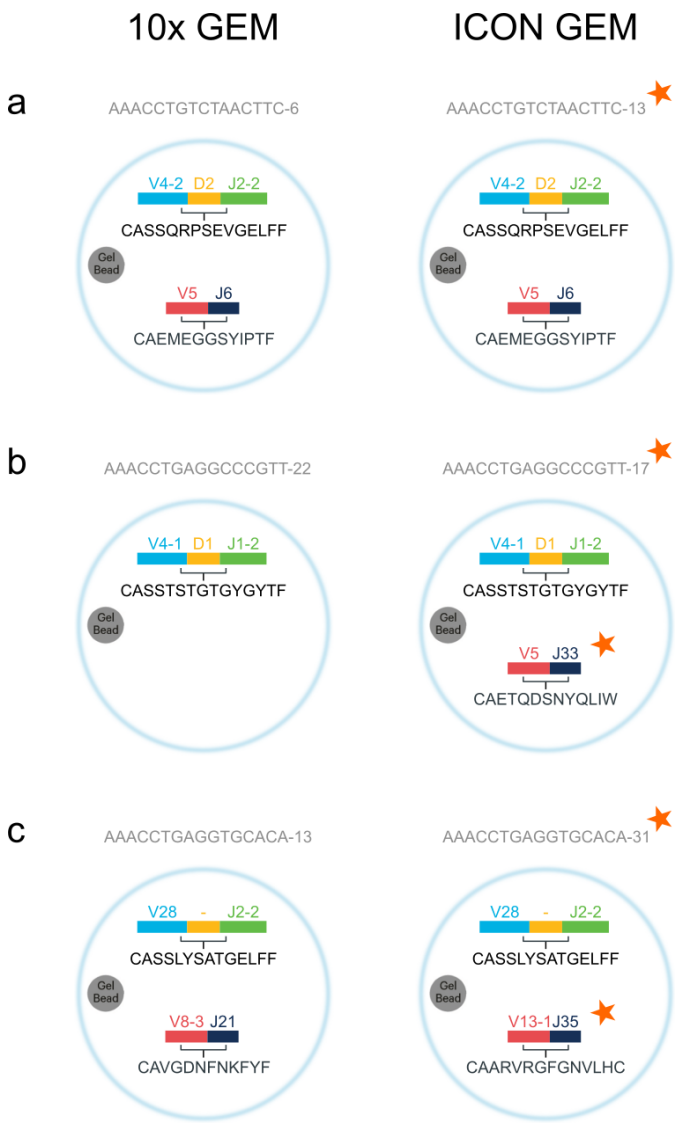


Figure 6.2: Illustrations of annotation inconsistencies. The figure shows examples of GEMs and their TCR annotations from 10x and ICON, respectively. The observed inconsistencies are grouped into three major groups. The inconsistencies are highlighted with a red star in each group. (a) 33,342 GEMs were mapped from the ICON set with inconsistent GEM barcode suffix. Mapping was based on the GEM barcode nucleotide sequence and TCR annotations. (b) 1854 GEMs were missing either an α- or a β-chain in the 10x data, but not in the ICON set. (c) 1537 GEMs were fully annotated, but the TCR annotations were inconsistent between ICON and the 10x data.

redefining clonotypes does not affect the individual GEM annotations of VαβJαβ-genes or CDR3αβ sequences. The redefinition only pertains to how GEMs are grouped and labeled.

The optimized ATRAP threshold on UMI counts

Of the complete data provided by 10x Genomics, we initially reduced the set to only include IUPAC encoded amino acids within CDR3 sequences and further only considered GEMs which contained both TCR and pMHC annotations, resulting in 181,913 GEMs. Redefining 10x clonotypes resulted in 76,627 unique VαβJαβ-CDR3αβpairs. Of those clonotypes, 1151 were represented with 10 or more GEMs, and for 1107 of them we were able to annotate an expected binder. The derived optimized UMI thresholds set a cutoff at minimum 5 UMI for any pMHC. For pMHC multiplets the most abundant pMHC must be 1.2 times greater in UMI counts than the second most abundant pMHC. A minimum of 1 UMI is required for TCR α- and β-chains. By this filter, the data set is reduced to 91,652 GEMs and 27,925 unique clonotypes. Additionally, filtering on matching HLA serves as the recommended minimum of filters for ATRAP.

Benchmark of ICON and ATRAP

The two filtering frameworks were benchmarked on four metrics, as described by Povlsen et al. [233]: Fraction of retained GEMs, accuracy of specificity, average binding concordance across all clonotypes, and AUC of CDR3αβ similarities. Accuracy is computed as the fraction of GEMs where the most abundant pMHC (by UMI counts) corresponds to the expected binder of a clonotype. An expected binder is defined for each clonotype as the pMHC which is distributed with a mean UMI count significantly higher (Wilcoxon, $\alpha = 0.05$) than the other pMHCs detected as binders for the given clonotype. Binding concordance is computed as the fraction of GEMs within a clonotype than binds a given pMHC and describes the dispersion of pMHC annotations within the clonotype. In a data set where no cross-reactivity is expected, the average binding concordance should be 100%. Finally, the similarity between two TCRs is defined as the summed score of

the pairwise CDR3α and CDR3β similarities each calculated using the kernel similarity method described in Shen et al., 2012 [201] and applied in Chronister et al., 2021 [191]. The AUC metric is computed based on the hypothesis that different TCRs binding the same pMHC (intra-specificity) are more similar to each other than to TCRs of other specificities (inter-specificity). The performance metrics are presented in figure 6.3.

The metrics presented in figure 6.3 reveal good performance from both frameworks. Figure 6.3a+b show the distribution of similarity scores between intra- and inter-specificity TCRs for each filtering step. Figure 6.3a show the individual effects of each filter, revealing that filtering specificity singlets away to only retain specificity multiplets, yields the greatest separation between intra- and inter-specificity distributions of all filtering steps. We define a specificity singlet as a TCR-pMHC pair only detected with a single GEM, which makes the pairing more susceptible to artifacts. The combined effect of each filter is visualized in figure 6.3b, which clearly shows how the separation of inter- and inter-specificity improves with more filters. To quantify the separation of distributions, we compute an AUC score from the principles that perfect intra-specificity scores are close to maximum value of 2, while inter-specificity resembles completely different TCRs of similarity close to 0. The exact numerical values of the individual specificities are not of interest and they do not affect the AUC. Note, that AUC here does not translate into a predictive performance, but rather reflects the extend to which intra-similarity can be distinguished from inter-similarity values.

The summary of both filtering frameworks across our selected performance metrics is presented by figure 6.3c. Both ICON and the combined ATRAP filters discard a large number of GEMs. The recommended filtering steps for ATRAP consist of filtering on UMI thresholds and matching HLA between annotated pMHC and HLA haplotype of donor, which yields 40,584 GEMs, which is slightly more than ICON (39,806). Filtering away specificity singlets only removes 5624 GEMs extra, but yields a gain in AUC, as we also saw in 6.3a+b. However, many of those GEMs represent unique clonotypes, so this filter also vastly reduces the total

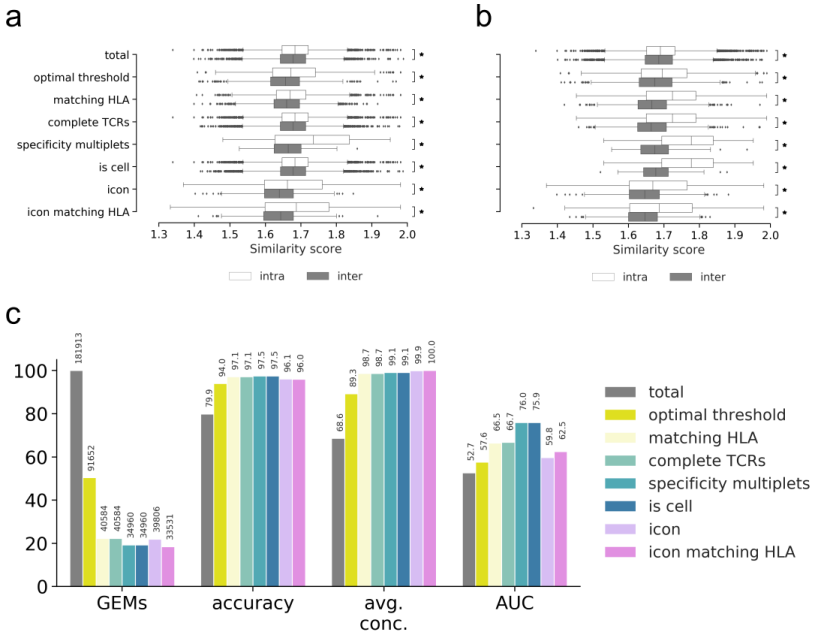


Figure 6.3: Performance metrics for evaluating the filtering steps of ATRAP with ICON. The ATRAP filtering steps consist of total (raw, unfiltered data), optimal threshold obtained from grid search, matching HLA, complete TCRs with a unique set of ∝- and β-chain, specificity multiplets i.e., TCR-pMHC pairs observed in two or more GEMs, and "is cell" defined by 10x Genomics Cellranger. ICON yields a single output, however, an addendum has been made to also filter ICON output on HLA match between pMHC and HLA haplotype of donor. (a) The boxplots show kernel similarity scores between CDR3β sequences of intra- (white) and inter- (dark) specificity for each of the filtering steps. A significant difference (Wilcoxon, $\alpha = 0.05$) of mean between inter- and intra-specificity is marked with an asterisk to the right (b) Here the boxplots show cumulative effect of ATRAP filters on similarity scores. (c) Performance is measured and summarized by a number of metrics: ratio of retained GEMs (GEMs), accuracy defined by proportion of GEMs where most abundant pMHC matches the expected binder (accuracy), average binding concordance (avg. conc.) and AUC of similarity scores (AUC). The ATRAP filters are also here cumulatively added to show increasing improvement in performance.

number of clonotypes.

As mentioned, ICON does not discard GEMs based on HLA match between pMHC and donor haplotype. However, we have tested

the impact of adding that filter to ICON, which reduces the yield to 33,531 GEMs. The performance measured by accuracy and average concordance is generally very high. ICON scores almost perfect binding concordance at every clonotype, which it essentially was designed to do, hence, we assume that the corrections of pMHC UMI counts and imputations of CDR3 sequences play a major role in this result. However, the slightly lower AUC of similarity scores of ICON suggest that some imputations might have been incorrect. Based on the AUC of similarity scores, the ATRAP-filters yield a slightly better performance, however, ICON yields specificity annotations of very little ambiguity, where each clonotype is assigned to only one pMHC

Visual inspection of ICON and ATRAP outputs

The differences in binding concordance between ATRAP and ICON are clearly visualized in figure 6.4 and figure 6.5. Figure 6.4 presents the ATRAP-filters of UMI threshold, HLA matching, and complete TCRs i.e., unique pairing of α- and β-chain.

With an average binding concordance of 98.7, we observe 407 GEMs with a binding concordance <50%, which we will refer to as outliers. A substantial proportion of these cross-binding events are across different HLA alleles. This contradicts the prevailing belief that T cells are restricted to the HLA for which they were positively selected during maturation. We thus suspect that some of these events are a result of random capture of ambient multimer barcode.

In 65 GEMs of the 407 outliers, an expected pMHC target had not been identified, due to the small sizes of the clones. Of the remaining 320 outliers, 76 GEMs exhibit a pattern which aligns with potential cross-reactivity.

Typically a TCR will have a single, preferred target, while allowing binding of other pMHCs to a lesser extend, i.e. clones of a clonotype may display a single dominant pMHC response of high binding concordance with few smaller responses of low binding concordance. For the clonotypes of these 76 GEMs, the dominant high-concordance pMHC coincides with the expected target

of the individual clonotypes. In 18 of these GEMs, the corresponding clonotypes showed divergent HLA restriction between the annotated low-concordance pMHC and the expected target for the given clonotype. In all of the 76 GEMs, the expected target was detected albeit at a lower UMI count than the annotated pMHC.

The remaining set of 266 GEMs consist of 80 clonotypes exhibiting highly dispersed binding to many different pMHCs, all with low binding concordance. All of these GEMs also contains multiplets of pMHCs. Based on these observations, we conclude that the majority of the 407 outliers are likely artifacts that have escaped the ATRAP filtering steps and thus not true cross-binding events.

Figure 6.5 presents the ICON retrieved specificities. With an average binding concordance of 99.9% most clonotypes are paired with a single specificity, and only 24 GEMs are categorized as outliers. 13 of the outliers are annotated with a pMHC that do not match the allele of the donor. 4 of the outliers contains CDR3 sequences that differ from the 10x annotation and may be a result of imputation.

Finally, a key difference between the two methods is that ATRAP retains 45 pMHCs from the staining whereas ICON retains 34 pMHCs. The 11 peptides retained by ATRAP and not ICON elicit small and few responses, but are primarily not involved in cross-binding events. With both filtering frameworks, the largest responses are toward KLG HLA*A-03:01, RKA HLA*B-08:01, and GIL HLA*A-02:01. ICON retains more GEMs and more clonotypes within these peptides, at the expense of other specificities, than ATRAP does.

Discussion

Single-cell screening assays may pave the way for better understanding of T cell specificity. The technology enables the study of binders, decisive non-binders and even cross-binding. However, de-noising single-cell specificity data is a critical bottleneck in studying T cell specificity. Here, we evaluate two methods, ATRAP and ICON, both aiming to resolve this bottleneck, filtering noise and putative artifacts from true binding events. Since

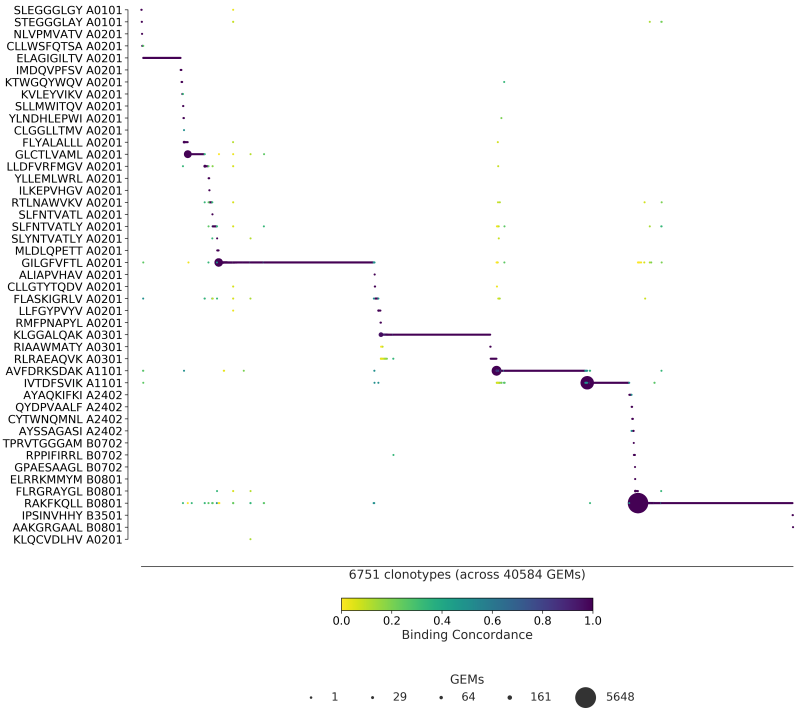


Figure 6.4: ATRAP derived specificity per clonotype. ATRAP-filters consist of UMI threshold, HLA matching, and complete TCRs i.e., unique pairing of α- and β-chain. The library peptides are listed on the y-axis and each clonotype is represented on the x-axis. Below the x-axis is annotated the total number of clonotypes and GEMs in the presented data. The marker size shows the number of GEMs supporting a given specificity. The color indicates the binding concordance which is calculated as the fraction of GEMs within a clonotype that support a given pMHC. The higher the concordance, the larger the fraction of supporting GEMs.

no golden-standard exist, the methods are evaluated via metrics designed for the purpose.

The two filtering frameworks both show very good performance, but with substantially different advantages and disadvantages. ICON excels at reducing ambiguous specificity annotations, such that the majority of clonotypes is annotated with exactly one pMHC target. The efficient reduction of outliers may, however, also become a hindrance for detecting cross-reactivity. The ATRAP

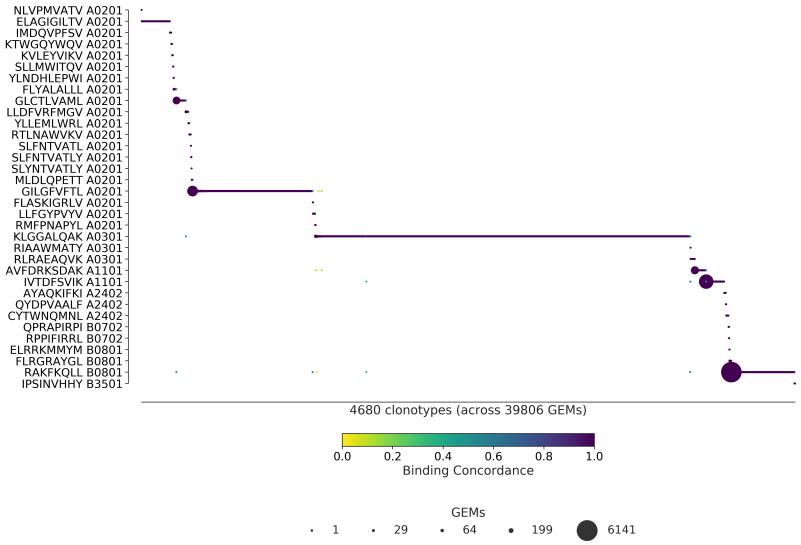


Figure 6.5: ICON derived specificity per clonotype. The library peptides are listed on the y-axis and each clonotype is represented on the x-axis. Below the x-axis is annotated the total number of clonotypes and GEMs in the presented data. The marker size shows the number of GEMs supporting a given specificity. The color indicates the binding concordance which is calculated as the fraction of GEMs within a clonotype that support a given pMHC. The higher the concordance, the larger the fraction of supporting GEMs.

method includes more GEMs across more pMHCs. A larger proportion of GEMs represent binding-events which resemble cross-reactivity, although after careful scrutiny, the majority of these are noisy observations having escaped filtering.

The filtering frameworks were evaluated on four metrics: retention of GEMs, binding accuracy guided by expected targets, average binding concordance, and AUC of kernel similarity scores. ATRAP scores the highest accuracy, however, binding accuracy may be a biased metric in this context as ATRAP was specifically designed to maximize this score. Similarly, we see ICON showing superior average binding concordance, favoring low dispersion of specificity within a clonotype, which ICON was purposefully designed to reduce. The AUC of kernel similarity scores is the only method-independent metric, which however does not account for

outliers, in favor of ATRAP.

Each framework has a set of requirements for the method to work optimally. ATRAP heavily relies on cell hashing, where HLA typing of donors is known, to validate specificities. In contrast, ICON relies on gene expression data to remove duplicates and negative control pMHC multimers to correct binding signal of positive pMHCs. The impact of gene expression data was previously tested for ATRAP, which showed only minute added performance [233]. Due to the low impact and the high expense of running gene expression sequencing, this filtering step was deprioritized. Cell hashing, of course, also confers an additional cost, however it further enables the study of immunodominant epitopes and individual T cell repertoires. The use of negative control pMHCs allows ICON to set a cutoff for pMHC UMI counts, similar to the accuracy optimizing threshold in ATRAP. The weakness of negative control pMHCs is that no one can yet define true negative targets. To circumvent this, utilizing empty multimer scaffolds containing only the DNA barcode as negative controls would reveal the level of ambient barcodes polluting the assay without risking rare but true binding.

Both frameworks assume that the pMHC UMI count reflects the likelihood of a TCR-pMHC pair, and use the count either directly (ATRAP or corrected and normalized (ICON) to filter away GEMs. However, it is important to note, that the UMI count actually refers to the number of pMHC multimers captured together with a T cell in a GEM. The count may be affected by the extent of ambient multimers, T cell expression of TCRs, and binding affinity. Thus to improve the filtering strategies of ATRAP or ICON future methods may implement adjusted TCR-pMHC pairing scores.

Pairing of TCR and pMHC is further made difficult in the cases where a presumed single cell expresses two different α- or β-chains. The dual expression cannot simply be written off as capture of multiple cells, as multiple GEMs exhibit the same dual TCR profile, and is a known phenomenon [86–88]. Neither ICON nor ATRAP seeks to investigate the impact on specificities, but simply annotate the most abundantly expressed chain. To improve

specificity detection, this aspect should be investigated further. Moreover, CDR3 sequences are not unique, but exist in various combinations, despite the stochastic process under which they are produced. Therefore, imputing CDR3 chains for GEMs with either multiple chains or GEMs missing a chain, will often not result in unique pairing. We speculate that ICON has attempted this, since we have observed discrepancies in CDR3 annotations between 10x and ICON. The comparison was further complicated by inconsistent GEM barcodes between ICON and the 10x data. The altercation of barcodes is unaccounted for by the authors of ICON.

In conclusion, the two frameworks perform on par. ICON provides high specificity at the expense of sensitivity, whereas ATRAP provides high sensitivity to allow detection of cross-binding events, but at the expense of specificity.

Materials and Methods

Data retrieval

10x Genomics data used for this study were downloaded from: https://support.10xgenomics.com/single-cell-vdj/datasets.

Benchmark data was created by Zhang et al. employing their method, ICON (Integrative COntext-specific Normalization), for identifying reliable TCR-pMHC interactions. Data was downloaded from http://advances.sciencemag.org/cgi/content/full/7/20/eabf5835/DC1. This set contains 53,062 cells (here referred to as GEMs) which pass the ICON filtering with ICON-corrected pMHC and TCR annotations. The ICON output provided with the publication contains a fifth donor, donor V, which was removed from the set (14,052 GEMs).

Data curation

The data consists of four sets of single-cell RNA sequencing and immune profiling from four healthy donors. The HLA haplotype of each donor was manually added to each set. The sets were concatenated for one combined analysis. Few GEM-specific

10x barcodes (GEM barcodes) were duplicates across the donor sets, therefore the barcodes were additionally suffixed by donor, i.e. AAACCTGTCTAACTTC-6-s2. Cells (referred to as GEMs) were removed if the annotated CDR3αβsequences were not productive, full length, or contained non-IUPAC characters, resulting in 181,913 GEMs. Differently annotated clonotypes sharing VJ-CDR3αβannotations were aggregated, as described in [233]. GEMs will be included in a clonotype even if a chain is missing based on unique matches of the existing chain to only one fully defined clonotype. Clonotypes will be defined by only the α- or the β-chain if no GEMs contained an αβ-pair with a matching chain. Finally, some clonotypes contain multiplets of α- or the β-chain, however, only the most abundant chain is selected to represent the clonotype.

Data filtering

As described in [233], different types of filters can be applied single-cell immune profiling data to reliably identify TCR-pMHC interactions. The method handles multi-omics single-cell sequencing data generated from a multiplexed multimer binding platform such as 10x Genomics immune profiling. The accepted inputs include single-cell RNA sequencing, targeted T cell receptor sequencing, dCODE-Dextramer sequencing for DNA barcoded pMHC multimers, as well as CITE-seq sequencing of DNA barcoded cell hashing antibodies. The method includes the following major steps:

Step 1: Correction of 10x annotated clonotypes as described in [233]. Instead of limiting clonotypes to groups of GEMs with exact nucleotide sequence identity, clonotypes were defined based on VJαβ-gene annotation and the CDR3αβamino acid sequences. Clonotypes for GEMs containing only one TCR chain were imputed if the chain matched only one preestablished clonotype. GEMs containing multiple chains were annotated by the most abundant chain by UMI count.

Step 2: Filtering based on data-driven thresholds as described in [233]. For each clonotype consisting of more than 10 GEMs the expected target is identified if a pMHC has significantly higher UMI distribution than other pMHCs also captured in GEMs of the

given clonotype. Significance is tested by Wilcoxon, $\alpha = 0.05$. The pMHCs not declared as target are considered as background noise. An accuracy score based on the fraction of target pMHCs over background pMHCs guides the search for the optimal threshold to filter all data on.

Step 3: Match pMHC HLA allele with donor haplotype. The HLA-A, -B, and -C haplotypes were provided by an application note following the release of the single-cell sequencing of the four healthy individuals. Since the samples were sequenced individually the haplotypes were easily added to the data sets. GEMs consisting of mismatch between donor haplotype and pMHC were discarded.

Step 4: Selecting GEMs with paired αβ-chains. GEMs with only a single chain were removed. For GEMs with multiple α- and/or β-chains, the ones with highest UMI counts were assigned to each GEM.

Step 5: Filtering specificity singlets. If a TCR-pMHC pair is only observed once it is discarded as to increase confidence in matches.

Step 6: Selecting 10x annotated cells. Application of the 10x provided filter "is_cell".

Benchmark

The impact of above mentioned filters were compared to the ICON framework. ICON was applied on the public 10x data sets and the the result thereof was provided by the authors via the publication [127]. The annotation for each GEM between the two approaches was traced per donor via the 10x barcode, omitting the well suffix of the barcode. The two approaches were compared based on number of retained GEMS, average binding concordance across clonotypes, and AUC of kernel similarity scores.

Binding concordance

Binding concordance is defined per clonotype as the distribution of GEMs annotated with varying pMHCs, as described in [233]. In a clonotype, the more GEMs annotated with the same pMHC,

the larger the concordance for that specificity. The average concordance is a single measure of how much cross-binding the full data contains.

AUC of kernel similarity scores

Kernel similarity was implemented from work by Shen et al. [201] and applied by [191]. The method was adapted to handle αβpairs as described in [233]. Similarity scores were computed for sets of TCRs binding the same pMHC (intra-specificity) and sets of TCRs binding different pMHCs (inter-specificity). The similarity scores were converted to AUC from the assumption that intra-similarity would approach maximum score of 2, whereas inter-similarity would approach minimum score of 0. Only high-concordant specificities were included in the analysis.

CHAPTER 7

Epilogue

The research projects presented in this thesis are centered around T cell specificity. New frameworks were developed and tested to aid our ability to interrogate T cell recognition more accurately. The prospect of these frameworks is an unveiling of the rules governing T cell interaction with cognate pMHCs, which will advance the development of T cell based immunotherapeis and rational vaccines.

Identification of immunogenic targets.
DNA barcoded pMHC multimers has paved the way for large panel screening of T cell responses, enabling the study of immunogenic antigens across an entire viral genome. In **paper I**, we applied this method to screen the SARS-CoV-2 genome and map regions of individual epitopes and immunodominant epitopes. These findings are the first steps toward a vaccine design eliciting T cell mediated immunity for stronger and longer immunity [234]. Probing T cell repertoires of both healthy and infected individuals provided insight into pre-exposure responses, which may reflect the cross-reactive aspect of TCRs. We hypothesized that cross-reactivity could be conferred by T cells selected for common cold coronaviruses and showed short Hamming distance between SARS-CoV-2 epitopes and common cold peptides. Adding common cold corona-

viruses in the panel could have shown, if the responses were correlated with SARS-CoV-2. The knowledge acquired from **paper I** is limited to evaluations of the targets and distributions T cell responses within the sampled repertoire. At single-cell resolution, the repertoires of healthy donors responding to SARS-CoV-2 epitopes could further elucidate to what extent pre-exposure T cells exist and whether these are public clonotypes, commonly shared in populations. In hindsight, longitudinal studies of cohorts with mild diseases and severe diseases could shed light on the impact of initial repertoire composition as a measure of heterologous immunity [235].

The important aspects of cross-reactivity.
In order for our immune system to broadly protect us from any given pathogen, T cell cross-reactivity is pivotal. Cross-reactivity enhances the chance that a suitable TCR is represented in a T cell repertoire, ensuring timely and appropriate response to a given antigen. Moreover, the risk of escape-variants is reduced due to partially overlapping specificities of a repertoire [64, 74]. However, the downside of cross-reactivity is the risk of allergy or auto-immunity [71], which has been inflicted by a number of infections [236–239]. This feature of the immune system has been exploited in immune therapy of cancer, where neo-antigens are only slight variations of self [240]. To initiate proper response without undue consequential auto-reactivity, it is paramount to understand the scope of T cell specificity and cross-reactivity. The ideal scenario is a comprehensive model foreseeing potential cross-reactivity to avoid off-target toxicity. Inadvertent auto-immunity was a tragic result of a cancer treatment by affinity-enhanced T cells which were cross-reactive to titin, a self-peptide presented on cardiomyocytes causing cardiogenic shock [241]. Moreover, evidence suggest that vaccines do not act independently of other vaccines, and that vaccines influence infections caused by other pathogens than the target disease, which might also result in adverse events [242]. Thus, a deep understanding of the mechanisms guiding T cells specificity, which determine cross-reactivity is essential for therapeutic development.

Improved detection of cross-reactive events.
Deep insight into T cell specificity may come with single-cell data and appropriate de-noising methods as presented in **paper II** and **paper III**. However, the study of cross-reactivity is particularly challenging when the experiments are highly affected by a range of confounding factors which are difficult to discern from true biological signal. Dedicated efforts must be set to improve sensitivity to cross-reactivity without loosing general specificity, by studying scenarios where cross-reactivity is expected. A great technique for high resolution measurement of cross-reactivity is TCR-fingerprinting assays, which provides a hierarchy of binding preferences for each individual T cell [68, 69]. The technique can be used to design a peptide panel of known cross-binders to include in a single-cell sequencing setup for proper evaluation of the cross-reactive detection capabilities. Importantly, the fingerprint-reference would also aid in identifying false cross-reactive events. Selecting clones of orthogonal HLA-restriction is imperative for a broad control pool. Particularly HLA-restriction was a great source of confusion and frustration during development of the method presented in **paper II**. In order to limit the degrees of freedom, we decided to assume complete restriction of self-MHC, however, the potential of allo-reactivity impairs this approach. Thus, extending fingerprinting assays to investigate allo-reactivity would provide a great asset for improved distinction between cross/allo-reactivity and technical noise.

Another aspect of cross-reactivity may be reflected by dual TCRs. Dual TCR T cells are expected exceptions of TCR gene rearrangement during thymocyte development [86–88]. However, the role, or side effect, of dual TCRs is still unknown. If dual TCRs confers dual specificity, it may improve protective immunity by enabling T cell multitasking or cause inappropriate responses as auto-immunity. The first step, is to infer whether dual TCRs affect specificity. This aspect was barely recognized in **paper II**, however, with reliable single-cell data of TCR repertoires, the dual TCR specificities can be evaluated against clonotypes of unique $\alpha\beta$-pairing as a subset of the dual pairing.

Design of a golden standard.
The major challenge in developing de-noising frameworks for single-cell sequencing is the lack of a golden-standard benchmark set, as discussed in **paper II** and **paper III**. By combining a range of the current techniques, the individual sources of uncertainty may be averaged out, to obtain a high quality benchmark set. Assays that compliment each other and particularly high-throughput single-cell sequencing include multimer staining [123, 125], T cell repertoire sequencing [243], T cell fingerprinting [68, 69], and perhaps also sequencing of a subset of manually sorted single cells [244]. As shown in **paper II**, validating single-cell sequencing via multimer staining provides detailed insight into the precision and recall of T cell response distributions. Clonotype annotations and their relative frequencies within distinct targets may be evaluated from bulk repertoire sequencing of strictly sorted cells, striving for absolute specificity. Since bulk sequencing only provides a single chain, a small subset of manually sorted single cells may provide confidence in accurate pairing of α- and β-chains. To further strengthen the golden standard, the set should include a number of T cell clones with known cross-reactive targets from fingerprinting analysis, as described above. This component would serve as a sort of spike-in control when evaluating de-noising strategies and their retention of cross-reactivity. Based on this golden-standard set, a universal model or framework can be developed to use for future data sets.

Advancing de-noising frameworks
Single-cell data is loaded with information to capitalize from, which perhaps also clouds the vision of how to properly harness the features. The two frameworks for de-noising single-cell specificity data presented in **paper III** both incorporate pMHC UMI count as a feature for optimal threshold setting. An obvious, and also tested, feature to remove noisy observations, is gene expression data. Although widely employed to remove dying cells and duplicates, the implementation suffers from ad hoc thresholding guidelines, which can cause large variations in downstream data [245–247]. Without exhaustive analysis of thresholds, we concluded in **paper II** that the effect of filtering based on gene expression

was not sufficient to justify the high cost of full transcriptome sequencing. Instead, we recommend implementation of cell hashing combined with donor HLA haplotyping as a powerful tool to validate annotations. Even though cell hashing, as pMHC staining, is also affected by the inherent confounding factors of single-cell sequencing, the uncertainty of both components is independent and therefore assumed to reduce the overall uncertainty when combined. Several approaches to reduce noise in cell hashing have been proposed [161, 162, 248], and a benchmark to implement the optimal method could enhance the performance of specificity detection. A possibly untapped source of information is the UMI count of TCRs, although the method presented in **paper II** does search for TCR UMI cutoffs, but no influence was detected. Assuming efficient capture of TCR genes during sequencing, the expression level of TCRs may also quantitatively impact the cellular association with pMHC multimers, i.e. the pMHC UMI count.

To advance de-noising frameworks, we may learn from similar fields. Several de-noising and imputations approaches within single-cell gene expression data involves neural networks, leveraging the ability to capture higher-order correlations [154, 249–259]. Particularly unsupervised methods, such as variational autoencoders, are popular because labeled data is superfluous [154, 250, 251, 253]. However, T cell specificity screening is challenged by a vastly larger biological variability than observed in single-cell gene expression, which sets higher requirements for careful procedures and evaluation.

Future prospects of accurate pairing of TCR-pMHC

The method developed in **paper II** and benchmarked in **paper III** is paving the way for further studies scrutinizing the intricate interaction between TCR and pMHC. The type of data made available by the methods presented in **paper III** are particularly well suited for modelling TCR-pMHC interaction. Each screening, followed by adequate de-noising strategies, may provide >1000 specificities across a panel of pMHCs. The sheer number of specificities from a single experiment holds a promise of progress in the field, as more well-defined patterns of binding can be detected. Moreover,

the design of the assay provides additional pivotal improvements to the currently available data. Previously, assays have only focused on detecting interactions to describe the mechanism of binding, but equally as much information can be extracted from studying the examples of incompatibility between TCR and pMHC. As single-cell screening is initiated with a "one-pot" staining, the assay allow us to assume that every TCR-pMHC pair not detected is truly incompatible, and thus in modeling lingo, negative. Another key asset, as discussed above, is the potential for detecting cross-reactivity. In order for a model to learn the promiscuity of TCRs and what restricts this promiscuity, training data must encompass cross-reactivity. The first steps have been taken with the methods presented in **paper III**, and likely more will follow. With improved methods for de-noising single-cell data and reliable identification of cross-reactivity, we can push the boundaries for T cell specificity predictions even further. Hence, I hope the work presented here will serve as inspiration for further research.

Bibliography

[1] K. Murphy and C. Weaver. *Immunobiology. Janeway's.* Garland Science, 2007.

[2] E. Caron et al. "The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation". In: *Molecular systems biology* 7 (2011).

[3] M. M. Davis and P. J. Bjorkman. "T-cell antigen receptor genes and T-cell recognition". In: *Nature* 334.6181 (1988), pp. 395–402.

[4] A. J. Darmon, D. W. Nicholson, and R. C. Bleackley. "Activation of the apoptotic protease CPP32 by cytotoxic T-cell-derived granzyme B". In: *Nature* 377.6548 (Oct. 1995), pp. 446–448.

[5] Z. Hu, P. A. Ott, and C. J. Wu. "Towards personalized, tumour-specific, therapeutic vaccines for cancer". In: *Nature Reviews Immunology* 18.3 (Dec. 2017), pp. 168–182.

[6] J. Banchereau and R. M. Steinman. "Dendritic cells and the control of immunity". In: *Nature 1998 392:6673* 392.6673 (Mar. 1998), pp. 245–252.

[7] P. Guermonprez et al. "Antigen Presentation and T Cell Stimulation by Dendritic Cells". In: *Annual Review of Immunology* 20 (Nov. 2003), pp. 621–667.

[8] C. Théry and S. Amigorena. "The cell biology of antigen presentation in dendritic cells". In: *Current Opinion in Immunology* 13.1 (Feb. 2001), pp. 45–51.

[9] C. Reis E Sousa. "Activation of dendritic cells: translating innate into adaptive immunity". In: *Current Opinion in Immunology* 16.1 (Feb. 2004), pp. 21–25.

[10] R. S. Allan et al. "Migratory Dendritic Cells Transfer Antigen to a Lymph Node-Resident Dendritic Cell Population for Efficient CTL Priming". In: *Immunity* 25.1 (July 2006), pp. 153–162.

[11] D. J. Zammit et al. "Dendritic Cells Maximize the Memory CD8 T Cell Response to Infection". In: *Immunity* 22.5 (May 2005), p. 561.

[12] A. Grakoui et al. “The immunological synapse: A molecular machine controlling T cell activation”. In: *Science* 285.5425 (July 1999), pp. 221–227.

[13] G. Iezzi, K. Karjalainen, and A. Lanzavecchia. “The Duration of Antigenic Stimulation Determines the Fate of Naive and Effector T Cells”. In: *Immunity* 8.1 (Jan. 1998), pp. 89–95.

[14] C. Abraham, J. Griffith, and J. Miller. “The Dependence for Leukocyte Function-Associated Antigen-1/ICAM-1 Interactions in T Cell Activation Cannot Be Overcome by Expression of High Density TCR Ligand”. In: *J Immunol References* 4399 (1999), pp. 4399–4405.

[15] K. Ley et al. “Getting to the site of inflammation: the leukocyte adhesion cascade updated”. In: *Nature Reviews Immunology 2007 7:9* 7.9 (Sept. 2007), pp. 678–689.

[16] T. Gebhardt et al. “Different patterns of peripheral migration by memory CD4+ and CD8+ T cells”. In: *Nature 2011 477:7363* 477.7363 (Aug. 2011), pp. 216–219.

[17] R. S. Akondy et al. “Origin and differentiation of human memory CD8 T cells after vaccination”. In: *Nature 2017 552:7685* 552.7685 (Dec. 2017), pp. 362–367.

[18] B. Youngblood et al. “Effector CD8 T cells dedifferentiate into long-lived memory cells”. In: *Nature 2017 552:7685* 552.7685 (Dec. 2017), pp. 404–409.

[19] K. S. Kobayashi and P. J. Van Den Elsen. “NLRC5: a key regulator of MHC class I-dependent immune responses”. In: *Nature Reviews Immunology 2012 12:12* 12.12 (Nov. 2012), pp. 813–820.

[20] K. L. Rock et al. “Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules”. In: *Cell* 78.5 (Sept. 1994), pp. 761–771.

[21] B. Lankat-Buttgereit and R. Tampé. “The transporter associated with antigen processing TAP: structure and function”. In: *FEBS Letters* 464.3 (Dec. 1999), pp. 108–112.

[22] A. Blees et al. “Structure of the human MHC-I peptide-loading complex”. In: *Nature 2017 551:7681* 551.7681 (Nov. 2017), pp. 525–528.

[23] P. Cresswell et al. “The nature of the MHC class I peptide loading complex”. In: *Immunological Reviews* 172.1 (Dec. 1999), pp. 21–28.

[24] W. K. Suh et al. “Interaction of MHC Class I Molecules with the Transporter Associated with Antigen Processing”. In: *Science* 264.5163 (1994), pp. 1322–1326.

[25] E. T. Spiliotis et al. “Selective Export of MHC Class I Molecules from the ER after Their Dissociation from TAP”. In: *Immunity* 13.6 (Dec. 2000), pp. 841–851.

[26] J. Neefjes et al. “Towards a systems understanding of MHC class I and MHC class II antigen presentation”. In: *Nature Reviews Immunology 2011 11:12* 11.12 (Nov. 2011), pp. 823–836.

[27] M. Wieczorek et al. “Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation”. In: *Frontiers in Immunology* 8.MAR (Mar. 2017), p. 292.

[28] H. Xu and D. Ren. “Lysosomal physiology”. In: *Annual review of physiology* 77 (Feb. 2015), pp. 57–80.

[29] A. Rudensky and C. Beers. “Lysosomal cysteine proteases and antigen presentation”. In: *Cytokines as Potential Therapeutic Targets for Inflammatory Skin Diseases* 56 (2006), pp. 81–95.

[30] R. J. Riese et al. “Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading”. In: *Immunity* 4.4 (1996), pp. 357–366.

[31] L. K. Denzin and P. Cresswell. “HLA-DM induces CLIP dissociation from MHC class II alpha beta dimers and facilitates peptide loading”. In: *Cell* 82.1 (July 1995), pp. 155–165.

[32] H. Kropshofer et al. “Editing of the HLA-DR-peptide repertoire by HLA-DM.” In: *The EMBO Journal* 15.22 (Nov. 1996), p. 6144.

[33] C. C. Norbury et al. “Constitutive macropinocytosis allows TAP-dependent major histocompatibility compex class I presentation of exogenous soluble antigen by bone marrow-derived dendritic cells”. In: *European Journal of Immunology* 27.1 (Jan. 1997), pp. 280–288.

[34] M. Kovacsovics-Bankowski and K. L. Rock. “A Phagosome-to-Cytosol Pathway for Exogenous Antigens Presented on MHC Class I Molecules”. In: *Science* 267.5195 (1995), pp. 243–246.

[35] A. Rodriguez et al. “Selective transport of internalized antigens to the cytosol for MHC class I presentation in dendritic cells”. In: *Nature Cell Biology 1999 1:6* 1.6 (Sept. 1999), pp. 362–368.

[36] M. Grommé et al. “Recycling MHC class I molecules and endosomal peptide loading”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.18 (Aug. 1999), pp. 10326–10331.

[37] M. J. Kleijmeer et al. “Antigen Loading of MHC Class I Molecules in the Endocytic Tract”. In: *Traffic* 2.2 (Jan. 2001), pp. 124–137.

[38] P. A. MacAry et al. “Mobilization of MHC class I molecules from late endosomes to the cell surface following activation of CD34-derived human Langerhans cells”. In: *Proceedings of the National Academy of Sciences of the United States of America* 98.7 (Mar. 2001), pp. 3982–3987.

[39] M. Kovacsovics-Bankowski et al. “Efficient major histocompatibility complex class I presentation of exogenous antigen upon phagocytosis by macrophages”. In: *Proceedings of the National Academy of Sciences of the United States of America* 90.11 (1993), pp. 4942–4946.

[40] M. J. Wick and H. G. Ljunggren. “Processing of bacterial antigens for peptide presentation on MHC class I molecules”. In: *Immunological Reviews* 172.1 (Dec. 1999), pp. 153–162.

[41] M. Subklewe et al. “Dendritic Cells Cross-Present Latency Gene Products from Epstein-Barr Virus–Transformed B Cells and Expand Tumor-Reactive Cd8+ Killer T Cells”. In: *Journal of Experimental Medicine* 193.3 (Feb. 2001), pp. 405–412.

[42] M. L. Albert, B. Sauter, and N. Bhardwaj. “Dendritic cells acquire antigen from apoptotic cells and induce class I-restricted CTLs”. In: *Nature 1998 392:6671* 392.6671 (Mar. 1998), pp. 86–89.

[43] V. Russo et al. “Dendritic cells acquire the MAGE-3 human tumor antigen from apoptotic cells and induce a class I-restricted T cell response”. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.5 (Feb. 2000), pp. 2185–2190.

[44] F. Berard et al. “Cross-Priming of Naive Cd8 T Cells against Melanoma Antigens Using Dendritic Cells Loaded with Killed Allogeneic Melanoma Cells”. In: *Journal of Experimental Medicine* 192.11 (Dec. 2000), pp. 1535–1544.

[45] M. Nouri-Shirazi et al. “Dendritic Cells Capture Killed Tumor Cells and Present Their Antigens to Elicit Tumor-Specific Immune Responses”. In: *The Journal of Immunology* 165.7 (Oct. 2000), pp. 3797–3803.

[46] V. L. Crotzer and J. S. Blum. “Autophagy and its role in MHC-mediated antigen presentation”. In: *Journal of immunology (Baltimore, Md. : 1950)* 182.6 (Mar. 2009), p. 3335.

[47] D. J. Klionsky and S. D. Emr. “Autophagy as a regulated pathway of cellular degradation”. In: *Science* 290.5497 (Dec. 2000), pp. 1717–1721.

[48] V. Jurtz et al. “NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data”. In: *Journal of immunology (Baltimore, Md. : 1950)* 199.9 (Nov. 2017), pp. 3360–3368.

[49] A. Y. Rudensky et al. “Sequence analysis of peptides bound to MHC class II molecules”. In: *Nature* 353.6345 (1991), pp. 622–627.

[50] R. M. Chicz et al. “Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size”. In: *Nature* 358.6389 (1992), pp. 764–768.

[51] M. Bouvier and D. C. Wiley. “Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules”. In: *Science (New York, N.Y.)* 265.5170 (1994), pp. 398–402.

[52] L. Zhang et al. “Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools”. In: *Briefings in bioinformatics* 13.3 (2012), pp. 350–364.

[53] S. Beck and J. Trowsdale. “The human major histocompatability complex: lessons from the DNA sequence”. In: *Annual review of genomics and human genetics* 1.2000 (2000), pp. 117–137.

[54] J. Robinson et al. “IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex”. In: *Nucleic Acids Research* 31.1 (Jan. 2003), p. 311.

[55] J. Robinson et al. “The IPD-IMGT/HLA Database - New developments in reporting HLA variation”. In: *Human immunology* 77.3 (Mar. 2016), pp. 233–237.

[56] P. Parham and T. Ohta. “Population biology of antigen presentation by MHC class I molecules”. In: *Science (New York, N.Y.)* 272.5258 (Apr. 1996), pp. 67–74.

[57] P. J. Bjorkman et al. “The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens”. In: *Nature 1987 329:6139* 329.6139 (1987), pp. 512–518.

[58] A. Kumánovics, T. Takada, and K. Fischer Lindahl. “Genomic Organization of the Mammalian MHC”. In: *Annual Review of Immunology* 21 (Nov. 2003), pp. 629–657.

[59] C. Zou et al. “$\gamma\delta$ T cells in cancer immunotherapy”. In: *Oncotarget* 8.5 (Jan. 2017), p. 8900.

[60] “Why must T cells be cross-reactive?” In: *Nature Reviews Immunology 2012 12:9* 12.9 (Aug. 2012), pp. 669–677.

[61] S. Gras et al. “The Shaping of T Cell Receptor Recognition by Self-Tolerance”. In: *Immunity* 30.2 (Feb. 2009), pp. 193–203.

[62] D. K. Cole et al. “Germ Line-governed Recognition of a Cancer Epitope by an Immunodominant Human T-cell Receptor *”. In: *Journal of Biological Chemistry* 284.40 (Oct. 2009), pp. 27281–27289.

[63] J. L. Jorgensen et al. “Mapping T-cell receptor–peptide contacts by variant peptide immunization of single-chain transgenics”. In: *Nature 1992 355:6357* 355.6357 (1992), pp. 224–230.

[64] I. Y. Song et al. “Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8+ T cell epitope”. In: *Nature Structural and Molecular Biology 2017 24:4* 24.4 (Feb. 2017), pp. 395–406.

[65] K. M. Armstrong, K. H. Piepenbrink, and B. M. Baker. “Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes”. In: *Biochemical Journal* 415.Pt 2 (Oct. 2008), p. 183.

[66] F. E. Tynan et al. “T cell receptor recognition of a ’super-bulged’ major histocompatibility complex class I–bound peptide”. In: *Nature Immunology 2005 6:11* 6.11 (Sept. 2005), pp. 1114–1122.

[67] C. H. Coles et al. “TCRs with Distinct Specificity Profiles Use Different Binding Modes to Engage an Identical Peptide–HLA Complex”. In: *The Journal of Immunology* 204.7 (Apr. 2020), pp. 1943–1953.

[68] A. K. Bentzen et al. “T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes”. In: *Nature Biotechnology 2018 36:12* 36.12 (Nov. 2018), pp. 1191–1196.

[69] A. R. Karapetyan et al. “TCR fingerprinting and off-target peptide identification”. In: *Frontiers in Immunology* 10.10 (2019), p. 2501.

[70] J. Ekeruche-Makinde et al. “Peptide length determines the outcome of TCR/peptide-MHCI engagement”. In: *Blood* 121.7 (Feb. 2013), pp. 1112–1123.

[71] D. Mason. “A very high level of crossreactivity is an essential feature of the T-cell receptor”. In: *Immunology Today* 19.9 (Sept. 1998), pp. 395–404.

[72] M. K. Jenkins et al. “On the Composition of the Preimmune Repertoire of T Cells Specific for Peptide–Major Histocompatibility Complex Ligands”. In: *Annual Review of Immunology* 28 (Mar. 2010), pp. 275–294.

[73] J. Ishizuka et al. “Quantitating T Cell Cross-Reactivity for Unrelated Peptide Antigens”. In: *Journal of immunology (Baltimore, Md. : 1950)* 183.7 (Oct. 2009), p. 4337.

[74] K. F. Chan et al. “Divergent T-cell receptor recognition modes of a HLA-I restricted extended tumour-associated peptide”. In: *Nature Communications 2018 9:1* 9.1 (Mar. 2018), pp. 1–13.

[75] M. De Simone, G. Rossetti, and M. Pagani. “Single cell T cell receptor sequencing: Techniques and future challenges”. In: *Frontiers in Immunology* 9.JUL (July 2018), p. 1638.

[76] H. Sakano et al. “Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes”. In: *Nature* 290.5807 (1981), pp. 562–565.

[77] J. J. Lafaille et al. “Junctional sequences of T cell receptor $\gamma\delta$ genes: Implications for $\gamma\delta$ T cell lineages and for a novel intermediate of V-(D)-J joining”. In: *Cell* 59.5 (Dec. 1989), pp. 859–870.

[78] D. G. Schatz. “Antigen receptor genes and the evolution of a recombinase”. In: *Seminars in Immunology* 16.4 (Aug. 2004), pp. 245–256.

[79] A. Olaru et al. “Recombination signal sequence variations and the mechanism of patterned T-cell receptor-beta locus rearrangement”. In: *Molecular immunology* 40.16 (2004), pp. 1189–1201.

[80] M. F. Quigley et al. “Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.45 (Nov. 2010), pp. 19414–19419.

[81] M. P. Lefranc. “Unique database numbering system for immunogenetic analysis”. In: *Immunology today* 18.11 (1997), p. 509.

[82] M. Ruiz and M. P. Lefranc. “IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures”. In: *Immunogenetics* 53.10-11 (2002), pp. 857–883.

[83] M. P. Lefranc et al. “IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains”. In: *Developmental and comparative immunology* 27.1 (2003), pp. 55–77.

[84] M.-P. Lefranc. “IMGT Unique Numbering”. In: *Encyclopedia of Systems Biology* (2013), pp. 952–959.

[85] M.-P. Lefranc. “Complementarity Determining Region (CDR-IMGT)”. In: *Encyclopedia of Systems Biology* (2013), pp. 451–453.

[86] N. J. Schuldt and B. A. Binstadt. “Dual TCR T Cells: Identity Crisis or Multitaskers?” In: *The Journal of Immunology* 202.3 (Feb. 2019), pp. 637–644.

[87] J. I. Elliott and D. M. Altmann. “Dual T cell receptor alpha chain T cells in autoimmunity”. In: *The Journal of Experimental Medicine* 182.4 (Oct. 1995), p. 953.

[88] H. T. Petrie et al. “Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes”. In: *The Journal of Experimental Medicine* 178.2 (Aug. 1993), p. 615.

[89] L. Klein et al. “Positive and negative selection of the T cell repertoire: what thymocytes see (and don’t see)”. In: *Nature Reviews Immunology 2014 14:6* 14.6 (May 2014), pp. 377–391.

[90] K. J. Lafferty, I. S. Misko, and M. A. Cooley. “Allogeneic stimulation modulates the in vitro response of T cells to transplantation antigen”. In: *Nature* 249.454 (1974), pp. 275–276.

[91] J. M. Lumsden et al. “Differential requirement for CD80 and CD80/CD86-dependent costimulation in the lung immune response to an influenza virus infection”. In: *Journal of immunology (Baltimore, Md. : 1950)* 164.1 (Jan. 2000), pp. 79–85.

[92] J.-A. Gonzalo et al. “Cutting Edge: The Related Molecules CD28 and Inducible Costimulator Deliver Both Unique and Complementary Signals Required for Optimal T Cell Activation”. In: *The Journal of Immunology* 166.1 (Jan. 2001), pp. 1–5.

[93] S. Wang et al. “Costimulation of T cells by B7-H2, a B7-like molecule that binds ICOS”. In: *Blood* 96.8 (Oct. 2000), pp. 2808–2813.

[94] S. L. Gaffen. “Signaling domains of the interleukin 2 receptor”. In: *Cytokine* 14.2 (Apr. 2001), pp. 63–77.

[95] C. M. Smith et al. “Cognate CD4(+) T cell licensing of dendritic cells in CD8(+) T cell immunity”. In: *Nature immunology* 5.11 (Nov. 2004), pp. 1143–1148.

[96] N. K. Rajasagi et al. “CD4+ T cells are required for the priming of CD8+ T cells following infection with herpes simplex virus type 1”. In: *Journal of virology* 83.10 (May 2009), pp. 5256–5268.

[97] C. Kurts, B. W. Robinson, and P. A. Knolle. "Cross-priming in health and disease". In: *Nature Reviews Immunology 2010 10:6* 10.6 (May 2010), pp. 403–414.

[98] J. Gálvez, J. J. Gálvez, and P. García-Peñarrubia. "Is TCR/pMHC affinity a good estimate of the T-cell response? An answer based on predictions from 12 phenotypic models". In: *Frontiers in Immunology* 10.MAR (2019), p. 349.

[99] S. Zhong et al. "T-cell receptor affinity and avidity defines antitumor response and autoimmunity in T-cell immunotherapy". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.17 (Apr. 2013), pp. 6973–6978.

[100] D. L. Mueller. "Mechanisms maintaining peripheral tolerance". In: *Nature Immunology 2010 11:1* 11.1 (Dec. 2009), pp. 21–27.

[101] N. Ohkura, Y. Kitagawa, and S. Sakaguchi. "Development and Maintenance of Regulatory T cells". In: *Immunity* 38.3 (Mar. 2013), pp. 414–423.

[102] E. A. Tivol et al. "Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4". In: *Immunity* 3.5 (1995), pp. 541–547.

[103] C. Scheibenbogen et al. "Analysis of the T cell response to tumor and viral peptide antigens by an IFNγ-ELISPOT assay". In: *J. Cancer* 71 (1997), pp. 932–936.

[104] A. Schmittel, U. Keilholz, and C. Scheibenbogen. "Evaluation of the interferon-γ ELISPOT-assay for quantification of peptide specific T lymphocytes from peripheral blood". In: *Journal of Immunological Methods* 210.2 (Dec. 1997), pp. 167–174.

[105] Xi-Shan Hao et al. "Determination of human T cell activity in response to allogeneic cells and mitogens. An immunochemical assay for gamma-interferon is more sensitive and specific than a proliferation assay". In: *Journal of immunological methods* 92.1 (Aug. 1986), pp. 59–63.

[106] P. Carayon and A. Bord. "Identification of DNA-replicating lymphocyte subsets using a new method to label the bromo-deoxyuridine incorporated into the DNA". In: *Journal of Immunological Methods* 147.2 (Mar. 1992), pp. 225–230.

[107] J. Hasbold et al. "Quantitative analysis of lymphocyte differentiation and proliferation in vitro using carboxyfluorescein diacetate succinimidyl ester". In: *Immunology and Cell Biology* 77.6 (Dec. 1999), pp. 516–522.

[108] A. B. Lyons. "Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution". In: *Journal of Immunological Methods* 243.1-2 (Sept. 2000), pp. 147–154.

[109] M. Corr et al. “T Cell Receptor-MHC Class I Peptide Interactions: Affinity, Kinetics, and Specificity”. In: *Science* 265.5174 (Aug. 1994), pp. 946–949.

[110] K. Matsui et al. “Kinetics of T-cell receptor binding to peptide/I-Ek complexes: correlation of the dissociation rate with T-cell responsiveness.” In: *Proceedings of the National Academy of Sciences* 91.26 (Dec. 1994), pp. 12862–12866.

[111] J. Huang et al. “Kinetics of MHC-CD8 Interaction at the T Cell Membrane”. In: *The Journal of Immunology* 179.11 (Dec. 2007), pp. 7653–7662.

[112] J. Huang et al. “The kinetics of two-dimensional TCR and pMHC interactions determine T-cell responsiveness”. In: *Nature 2010 464:7290* 464.7290 (Mar. 2010), pp. 932–936.

[113] J. B. Huppa et al. “TCR–peptide–MHC interactions in situ show accelerated kinetics and increased affinity”. In: *Nature 2010 463:7283* 463.7283 (Feb. 2010), pp. 963–967.

[114] N. Jiang et al. “Two-Stage Cooperative T Cell Receptor-Peptide Major Histocompatibility Complex-CD8 Trimolecular Interactions Amplify Antigen Discrimination”. In: *Immunity* 34.1 (Jan. 2011), pp. 13–23.

[115] B. J. Cameron et al. “Identification of a Titin-Derived HLA-A1–Presented Peptide as a Cross-Reactive Target for Engineered MAGE A3–Directed T Cells”. In: *Science translational medicine* 5.197 (Aug. 2013), 197ra103.

[116] J. J. Adams et al. “Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity”. In: *Nature Immunology 2015 17:1* 17.1 (Nov. 2015), pp. 87–94.

[117] B. B. Beezley and N. H. Ruddle. “A critical analysis of the T cell hybrid technique”. In: *Journal of immunological methods* 52.3 (Aug. 1982), pp. 269–281.

[118] C. E. Sharrock, E. Kaminski, and S. Man. “Limiting dilution analysis of human T cells: a useful clinical tool”. In: *Immunology Today* 11.C (Jan. 1990), pp. 281–286.

[119] H. von Boehmer et al. “T cell clones: Their use for the study of specificity, induction, and effector-function of T cells”. In: *Springer Seminars in Immunopathology* 3.1 (1980), pp. 23–37.

[120] J. Leem et al. “STCRDab: the structural T-cell receptor database”. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D406–D412.

[121] J. D. Altman et al. “Phenotypic analysis of antigen-specific T lymphocytes”. In: *Science (New York, N.Y.)* 274.5284 (1996), pp. 94–96.

[122] R. S. Andersen et al. “Parallel detection of antigen-specific T cell responses by combinatorial encoding of MHC multimers”. In: *Nature Protocols 2012 7:5* 7.5 (Apr. 2012), pp. 891–902.

[123] S. R. Hadrup et al. “Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers”. In: *Nature Methods 2009 6:7* 6.7 (June 2009), pp. 520–526.

[124] E. W. Newell et al. “Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization”. In: *Nature Biotechnology 2013 31:7* 31.7 (June 2013), pp. 623–629.

[125] A. K. Bentzen et al. “Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes”. In: *Nature Biotechnology 2016 34:10* 34.10 (Aug. 2016), pp. 1037–1045.

[126] 10xGenomics. “A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype”. In: *Application Note* (2020).

[127] W. Zhang et al. “A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity”. In: *Science advances* 7.20 (May 2021).

[128] A. H. Bakker and T. N. Schumacher. “MHC multimer technology: current status and future prospects”. In: *Current Opinion in Immunology* 17.4 (Aug. 2005), pp. 428–433.

[129] G. Dolton et al. “Comparison of peptide–major histocompatibility complex tetramers and dextramers for the identification of antigen-specific T cells”. In: *Clinical and Experimental Immunology* 177.1 (July 2014), p. 47.

[130] J. Huang et al. “Detection, phenotyping, and quantification of antigenspecific T cells using a peptide-MHC dodecamer”. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.13 (Mar. 2016), E1890–E1897.

[131] W. A. Bonner et al. “Fluorescence Activated Cell Sorting”. In: *Review of Scientific Instruments* 43.3 (Nov. 2003), p. 404.

[132] J. L. Haynes. “Principles of flow cytometry”. In: *Cytometry* 9.S3 (Jan. 1988), pp. 7–17.

[133] V. Greiff et al. “A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status”. In: *Genome Medicine* 7.1 (May 2015), pp. 1–15.

[134] C. Linnemann et al. “High-throughput identification of antigen-specific TCRs by TCR gene capture”. In: *Nature Medicine 2013 19:11* 19.11 (Oct. 2013), pp. 1534–1541.

[135] J. Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. In: *Science (New York, N.Y.)* 357.6352 (Aug. 2017), pp. 661–667.

[136] N. Karaiskos et al. “The Drosophila embryo at single-cell transcriptome resolution”. In: *Science* 358.6360 (Oct. 2017), pp. 194–199.

[137] A. K. Shalek et al. “Single-cell RNA-seq reveals dynamic paracrine control of cellular variation”. In: *Nature* 510.7505 (2014), pp. 363–369.

[138] S. Aibar et al. “SCENIC: single-cell regulatory network inference and clustering”. In: *Nature Methods 2017 14:11* 14.11 (Oct. 2017), pp. 1083–1086.

[139] A. C. Villani et al. “Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors”. In: *Science (New York, N.Y.)* 356.6335 (Apr. 2017).

[140] L. Velten et al. “Human haematopoietic stem cell lineage commitment is a continuous process”. In: *Nature Cell Biology 2017 19:4* 19.4 (Mar. 2017), pp. 271–281.

[141] C. A. Herring et al. “Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut”. In: *Cell Systems* 6.1 (Jan. 2018), 37–51.e9.

[142] A. M. Klein et al. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. In: *Cell* 161.5 (May 2015), pp. 1187–1201.

[143] E. Z. Macosko et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. In: *Cell* 161.5 (May 2015), pp. 1202–1214.

[144] S. Bose et al. “Scalable microfluidics for single-cell RNA printing and sequencing”. In: *Genome Biology* 16.1 (June 2015), pp. 1–16.

[145] A. B. Rosenberg et al. “Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding”. In: *Science (New York, N.Y.)* 360.6385 (Apr. 2018), pp. 176–182.

[146] T. Kivioja et al. “Counting absolute numbers of molecules using unique molecular identifiers”. In: *Nature methods* 9.1 (Jan. 2011), pp. 72–74.

[147] A. Raj and A. van Oudenaarden. “Stochastic gene expression and its consequences”. In: *Cell* 135.2 (Oct. 2008), p. 216.

[148] F. Buettner et al. “Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells”. In: *Nature Biotechnology 2014 33:2* 33.2 (Jan. 2015), pp. 155–160.

[149] M. J. Zhang, V. Ntranos, and D. Tse. “Determining sequencing depth in a single-cell RNA-seq experiment”. In: *Nature Communications 2020 11:1* 11.1 (Feb. 2020), pp. 1–11.

[150] V. Sarangi et al. “SCELLECTOR: ranking amplification bias in single cells using shallow sequencing”. In: *BMC Bioinformatics* 21.1 (Dec. 2020), pp. 1–10.

[151] B. Phipson, L. Zappia, and A. Oshlack. “Gene length and detection bias in single cell RNA sequencing protocols”. In: *F1000Research* 6 (2017).

[152] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. “Bayesian approach to single-cell differential expression analysis”. In: *Nature Methods 2014 11:7* 11.7 (May 2014), pp. 740–742.

[153] S. Yang et al. “Decontamination of ambient RNA in single-cell RNA-seq with DecontX”. In: *Genome Biology* 21.1 (Mar. 2020), pp. 1–15.

[154] N. J. Bernstein et al. “Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning”. In: *Cell Systems* 11.1 (July 2020), 95–101.e5.

[155] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner. “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors”. In: *Cell Systems* 8.4 (Apr. 2019), 329–337.e4.

[156] S. L. Wolock, R. Lopez, and A. M. Klein. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. In: *Cell Systems* 8.4 (Apr. 2019), 281–291.e9.

[157] H. M. Kang et al. “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation”. In: *Nature Biotechnology 2017 36:1* 36.1 (Dec. 2017), pp. 89–94.

[158] A. S. Bais and D. Kostka. “scds: Computational Annotation of Doublets in Single Cell RNA Sequencing Data”. In: *bioRxiv* (Feb. 2019), p. 564021.

[159] J. Xu et al. “Genotype-free demultiplexing of pooled single-cell RNA-seq”. In: *Genome Biology* 20.1 (Dec. 2019), pp. 1–12.

[160] C. S. McGinnis et al. “MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices”. In: *Nature Methods 2019 16:7* 16.7 (June 2019), pp. 619–626.

[161] M. Stoeckius et al. “Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics”. In: *Genome Biology* 19.1 (Dec. 2018), pp. 1–12.

[162] J. T. Gaublomme et al. “Nuclei multiplexing with barcoded antibodies for single-nucleus genomics”. In: *Nature Communications 2019 10:1* 10.1 (July 2019), pp. 1–8.

[163] R. Vita et al. “The Immune Epitope Database (IEDB): 2018 update”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D339–D343.

[164] D. V. Bagaev et al. “VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium”. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D1057–D1062.

[165] N. Tickotsky et al. “McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences”. In: *Bioinformatics* 33.18 (Sept. 2017), pp. 2924–2929.

[166] R. Gowthaman and B. G. Pierce. “TCR3d: The T cell receptor structural repertoire database”. In: *Bioinformatics* 35.24 (Dec. 2019), pp. 5323–5325.

[167] S. Nolan et al. “A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2”. In: *Research Square* (Aug. 2020).

[168] M. Hebeisen et al. “Molecular insights for optimizing T cell receptor specificity against cancer”. In: *Frontiers in immunology* 4.JUN (2013).

[169] M. Hebeisen et al. “SHP-1 phosphatase activity counteracts increased T cell receptor affinity”. In: *The Journal of clinical investigation* 123.3 (Mar. 2013), pp. 1044–1065.

[170] G. Dolton et al. “Optimized peptide-MHC multimer protocols for detection and isolation of autoimmune T-cells”. In: *Frontiers in Immunology* 9.JUN (June 2018), p. 1378.

[171] G. P. Dunn, L. J. Old, and R. D. Schreiber. “The immunobiology of cancer immunosurveillance and immunoediting”. In: *Immunity* 21.2 (Aug. 2004), pp. 137–148.

[172] G. P. Dunn, L. J. Old, and R. D. Schreiber. “The Three Es of Cancer Immunoediting”. In: *Annual Review of Immunology* 22 (Mar. 2004), pp. 329–360.

[173] A. Montemurro et al. “NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data”. In: *Communications Biology* 4.1 (2021), pp. 1–13.

[174] D. S. Fischer et al. “Predicting antigen specificity of single T cells based on TCR CDR3 regions”. In: *Molecular Systems Biology* 16.8 (Aug. 2020), e9416.

[175] E. Jokinen et al. “Determining epitope specificity of T cell receptors with TCRGP”. In: *bioRxiv* (Aug. 2019), p. 542332.

[176] I. Springer et al. “Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs”. In: *Frontiers in Immunology* 11 (Aug. 2020), p. 1803.

[177] J. W. Sidhom et al. “DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires”. In: *Nature Communications 2021 12:1* 12.1 (Mar. 2021), pp. 1–12.

[178] P. Dash et al. “Quantifiable predictive features define epitope-specific T cell receptor repertoires”. eng. In: *Nature* 547.7661 (July 2017), pp. 89–93.

[179] S. C. Boutet et al. “Scalable and comprehensive characterization of antigen-specific CD8 T cells using multi-omics single cell analysis”. In: *The Journal of Immunology* 202.1 Supplement (2019).

[180] B. Reynisson et al. “NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data”. In: *Nucleic Acids Research* 48.W1 (July 2020), W449–W454.

[181] T. J. O'donnell, A. Rubinsteyn, and U. Laserson Correspondence. "MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing". In: *Cell Systems* 11 (2020), pp. 42–48.

[182] M. Bassani-Sternberg et al. "Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity". In: *PLoS computational biology* 13.8 (Aug. 2017), e1005725.

[183] J. Racle et al. "Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes". In: *Nature biotechnology* 37.11 (Nov. 2019), pp. 1283–1286.

[184] J. Glanville et al. "Identifying specificity groups in the T cell receptor repertoire". In: *Nature 2017 547:7661* 547.7661 (June 2017), pp. 94–98.

[185] V. Isabell Jurtz et al. "NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks". In: *bioRxiv* (Oct. 2018), p. 433706.

[186] Y. Tong et al. "SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction". In: *Computational Biology and Chemistry* 87 (Aug. 2020), p. 107281.

[187] P. Moris et al. "Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification". In: *Briefings in Bioinformatics* 22.4 (July 2021).

[188] S. Gielis et al. "Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires". In: *Frontiers in Immunology* 10 (Nov. 2019), p. 2820.

[189] N. De Neuter et al. "On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition". In: *Immunogenetics* 70.3 (Mar. 2018), pp. 159–168.

[190] E. B. Wong et al. "TRAV1-2+ CD8+ T-cells including oligoconal expansions of MAIT cells are enriched in the airways in human tuberculosis". In: *Communications Biology 2019 2:1* 2.1 (June 2019), pp. 1–13.

[191] W. D. Chronister et al. "TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors". In: *Frontiers in Immunology* 12 (Mar. 2021), p. 673.

[192] A. Weber, J. Born, and M. Rodriguez Martínez. "TITAN: T-cell receptor specificity prediction with bimodal attention networks". In: *Bioinformatics* 37.Suppl 1 (July 2021), p. i237. arXiv: 2105.03323.

[193] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et biophysica acta* 405.2 (Oct. 1975), pp. 442–451.

[194] K. Pearson. "Determination of the coefficient of correlation". In: *Science* 30.757 (July 1909), pp. 23–25.

[195] V. I. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710.

[196] A. Madi et al. “T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences”. In: *eLife* 6 (July 2017).

[197] E. Miho et al. “Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires”. In: *Frontiers in Immunology* 9.FEB (Feb. 2018), p. 224. arXiv: 1711.11070.

[198] S. Henikoff and J. G. Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 89.22 (1992), pp. 10915–10919.

[199] N. De Neuter et al. “Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus”. In: *Genes and Immunity* 20.3 (2019), pp. 255–260.

[200] P. Meysman et al. “On the viability of unsupervised T-cell receptor sequence clustering for epitope preference”. In: *Bioinformatics* 35.9 (May 2019), pp. 1461–1468.

[201] W.-J. Shen et al. “Towards a Mathematical Foundation of Immunology and Amino Acid Chains”. In: (May 2012). arXiv: 1205.6031.

[202] M. Nielsen et al. “NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence”. In: *PLoS ONE* 2.8 (Aug. 2007).

[203] B. Alvarez et al. “NNAlign_MA; MHC Peptidome Deconvolution for Accurate MHC Binding Motif Characterization and Improved T-cell Epitope Predictions”. In: *Molecular & Cellular Proteomics : MCP* 18.12 (2019), p. 2459.

[204] I. Springer, N. Tickotsky, and Y. Louzoun. “Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction”. In: *Frontiers in Immunology* 12 (Apr. 2021), p. 1436.

[205] G. Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of Control, Signals and Systems 1989 2:4* 2.4 (Dec. 1989), pp. 303–314.

[206] B. Alipanahi et al. “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning”. In: *Nature Biotechnology 2015 33:8* 33.8 (July 2015), pp. 831–838.

[207] W. Rawat and Z. Wang. “Deep convolutional neural networks for image classification: A comprehensive review”. In: *Neural Computation* 29.9 (Sept. 2017), pp. 2352–2449.

[208] Z. Chen et al. “Feature selection may improve deep neural networks for the bioinformatics problems”. In: *Bioinformatics (Oxford, England)* 36.5 (Mar. 2020), pp. 1542–1552.

[209] R. Akbar et al. “A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding”. In: *Cell Reports* 34 (2021).

[210] W. Nelson et al. “To embed or not: Network embedding as a paradigm in computational biology”. In: *Frontiers in Genetics* 10.5 (2019), p. 381.

[211] N. Arsov and G. Mirceva. “Network Embedding: An Overview”. In: (Nov. 2019). arXiv: 1911.11726.

[212] A. Vaswani et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems* 2017-Decem (June 2017), pp. 5999–6009. arXiv: 1706.03762.

[213] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: (Oct. 2018). arXiv: 1810.04805.

[214] Y. Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: (Sept. 2016). arXiv: 1609.08144.

[215] T. Kudo and J. Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 66–71.

[216] J. Cheng et al. “BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning”. In: *Bioinformatics* (2021), pp. 1–8.

[217] J. Jin et al. “Attention mechanism-based deep learning pan-specific model for interpretable MHC-I peptide binding prediction”. In: *bioRxiv* 1655740 (2019).

[218] D. M. Reddy and S. Reddy. “Effects of padding on LSTMs and CNNs”. In: (Mar. 2019). arXiv: 1903.07288.

[219] V. Jurtz et al. “NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data”. In: *Journal of immunology (Baltimore, Md. : 1950)* 199.9 (Nov. 2017), p. 3360.

[220] L. Van Der Maaten and G. Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[221] L. McInnes, J. Healy, and J. Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: (Feb. 2018). arXiv: 1802.03426.

[222] M. S. Krangel. “Mechanics of T cell receptor gene rearrangement”. In: *Current Opinion in Immunology* 21.2 (Apr. 2009), pp. 133–139.

[223] E. Mahe, T. Pugh, and S. Kamel-Reid. “T cell clonality assessment: past, present and future”. In: *Journal of Clinical Pathology* 71.3 (Mar. 2018), pp. 195–200.

[224] N. R. Gascoigne et al. “TCR Signal Strength and T Cell Development”. In: *http://dx.doi.org/10.1146/annurev-cellbio-111315-125324* 32 (Oct. 2016), pp. 327–348.

[225] D. Jung and F. W. Alt. “Unraveling V(D)J Recombination: Insights into Gene Regulation”. In: *Cell* 116.2 (Jan. 2004), pp. 299–311.

[226] K. J. Jackson et al. “The shape of the lymphocyte receptor repertoire: Lessons from the B cell receptor”. In: *Frontiers in Immunology* 4.SEP (2013), p. 263.

[227] V. I. Zarnitsyna et al. “Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire”. In: *Frontiers in Immunology* 4.12 (2013), p. 485.

[228] Y. Elhanati et al. “repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data”. In: *Bioinformatics* 32.13 (July 2016), pp. 1943–1951. arXiv: 1511.00107.

[229] P. Marrack et al. “T cell receptor specificity for major histocompatibility complex proteins”. In: *Current Opinion in Immunology* 20.2 (Apr. 2008), pp. 203–207.

[230] “T Cell Receptor–MHC Interactions up Close”. In: *Cell* 104.1 (Jan. 2001), pp. 1–4.

[231] M. Wieczorek et al. “Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation”. In: *Frontiers in Immunology* 8.MAR (Mar. 2017), p. 292.

[232] N. L. La Gruta et al. “Understanding the drivers of MHC restriction of T cell receptors”. In: *Nature Reviews Immunology 2018 18:7* 18.7 (Apr. 2018), pp. 467–478.

[233] H. Povlsen et al. “ATRAP - Accurate T cell Receptor Antigen Pairing through data-driven filtering of sequencing information from single-cells [Manuscript submitted for publication]”. unpublished. 2022.

[234] J. Chen et al. “Decline in neutralising antibody responses, but sustained T-cell immunity, in COVID-19 patients at 7 months post-infection”. In: *Clinical and Translational Immunology* 10.7 (Jan. 2021), e1319.

[235] R. M. Welsh and L. K. Selin. “No one is naive: the significance of heterologous T-cell immunity”. In: *Nature Reviews Immunology 2002 2:6* 2.6 (2002), pp. 417–426.

[236] C. Benoist and D. Mathis. “Autoimmunity provoked by infection: how good is the case for T cell epitope mimicry?” In: *Nature Immunology 2001 2:9* 2.9 (2001), pp. 797–801.

[237] S. C. Clute et al. “Broad Cross-Reactive TCR Repertoires Recognizing Dissimilar Epstein-Barr and Influenza A Virus Epitopes”. In: *The Journal of Immunology* 185.11 (Dec. 2010), pp. 6753–6764.

[238] R. M. Welsh et al. “Heterologous immunity between viruses”. In: *Immunological reviews* 235.1 (May 2010), p. 244.

[239] M. Partinen et al. “Narcolepsy as an autoimmune disease: the role of H1N1 infection and vaccination”. In: *The Lancet Neurology* 13.6 (June 2014), pp. 600–613.

[240] J. Couzin-Frankel. “Cancer immunotherapy”. In: *Science* 342.6165 (Dec. 2013), pp. 1432–1433.

[241] G. P. Linette et al. “Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma”. In: *Blood* 122.6 (Aug. 2013), pp. 863–871.

[242] F. Shann. “The Nonspecific Effects of Vaccines and the Expanded Program on Immunization”. In: *The Journal of Infectious Diseases* 204.2 (July 2011), pp. 182–184.

[243] F. Rubelt et al. “Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data”. In: *Nature Immunology 2017 18:12* 18.12 (Nov. 2017), pp. 1274–1278.

[244] F. Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature Methods 2009 6:5* 6.5 (Apr. 2009), pp. 377–382.

[245] C. Ziegenhain et al. “Comparative Analysis of Single-Cell RNA Sequencing Methods”. In: *Molecular Cell* 65.4 (Feb. 2017), 631–643.e4.

[246] K. Bach et al. “Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing”. In: *Nature Communications 2017 8:1* 8.1 (Dec. 2017), pp. 1–11.

[247] N. A. Krentz et al. “Single-Cell Transcriptome Profiling of Mouse and hESC-Derived Pancreatic Progenitors”. In: *Stem Cell Reports* 11.6 (Dec. 2018), p. 1551.

[248] H. Xin et al. “GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing”. In: *Genome Biology* 21.1 (July 2020), pp. 1–35.

[249] K. R. Moon et al. “Visualizing structure and transitions in high-dimensional biological data”. In: *Nature Biotechnology 2019 37:12* 37.12 (Dec. 2019), pp. 1482–1492.

[250] G. Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. In: *Nature Communications 2019 10:1* 10.1 (Jan. 2019), pp. 1–14.

[251] X. Zhang et al. “NISC: Neural Network-Imputation for Single-Cell RNA Sequencing and Cell Type Clustering”. In: *Frontiers in Genetics* 0 (May 2022), p. 953.

[252] C. Arisdakessian et al. “DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data”. In: *Genome Biology* 20.1 (Oct. 2019), pp. 1–14.

[253] H. Li, C. R. Brouwer, and W. Luo. “A universal deep neural network for in-depth cleaning of single-cell RNA-Seq data”. In: *Nature Communications 2022 13:1* 13.1 (Apr. 2022), pp. 1–11.

[254] J. Wang et al. “scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses”. In: *Nature Communications 2021 12:1* 12.1 (Mar. 2021), pp. 1–11.

[255] Y. Zeng et al. “A robust and scalable graph neural network for accurate single-cell classification”. In: *Briefings in Bioinformatics* 23.2 (Mar. 2022).

[256] H. Wen et al. “Graph Neural Networks for Multimodal Single-Cell Data Integration”. In: (Mar. 2022). arXiv: 2203.01884v2.

[257] M. Amodio et al. “Exploring single-cell data with deep multitasking neural networks”. In: *Nature Methods 2019 16:11* 16.11 (Oct. 2019), pp. 1139–1145.

[258] S. Gigante et al. “Modeling Global Dynamics from Local Snapshots with Deep Generative Neural Networks”. In: *2019 13th International Conference on Sampling Theory and Applications, SampTA 2019* (Feb. 2018). arXiv: 1802.03497.

[259] D. B. Burkhardt et al. “Quantifying the effect of experimental perturbations at single-cell resolution”. In: *Nature Biotechnology 2021 39:5* 39.5 (Feb. 2021), pp. 619–629.

APPENDIX A

Paper I Appendix

Supplementary information

Supplementary tables

Supplementary Table 1. **Samples**

Hashing-ID	Donor	HLA-A	HLA-B
1	BC-300	0201	0702
2	BC-326	0201	
3	BC-126	0201	
4	BC-328	0301	0702
5	BC-62	0301	0702
6	BC-355	0201	0702
7	BC-360		0702
8	BC-314	0301	0702
9	BC-353	0301	0702
10	BC-311, BC-11, BC-83, BC-88, BC-341, BC-342, BC-76	0201, 0301	0702

Supplementary Table 1: Overview of which samples contain cells from which donors and the relevant donor haplotypes.

Supplementary table 2. **Peptide-MHC multimers**

Peptide	HLA	Origin	Barcode sequence	Fluorochrome
CLGGLLTMV	A0201	EBV LMP2	TATGAGGACGAATCT	APC
FLYALALLL	A0201	EBV LMP2	CCGATGTTGACGGAC	APC
YVLDHLIVV	A0201	EBV BRLF1	TAGTAGTTCAGACGC	APC
VLEETSVML	A0201	CMV IE-1	CCGTACCTAGATACA	APC
RVRAYTYSK	A0301	EBV BRLF1	GGTATGGCACGCCTA	APC
RPHERNGFTVL	B0702	CMV pp65	GGATGCATGATCTAG	APC
TPSVSSSISSL	B0702	EBV BFRF3	GATTCAATATGTGTC	APC
RPPIFIRRL	B0702	EBV EBNA3A	GGTAACTGCGCATAG	APC

TPRVTGGGAM	B0702	CMV pp65	GGTACAGTAAGTGAG	APC
RPHERNGFTV	B0702	CMV pp65	GCCACCTTAACACGC	APC
GILGFVFTL	A0201	Flu M1	TTCTATTCTAAGCCG	PE
GLCTLVAML	A0201	EBV BMFL1	TCCAAGTTAGCTTAC	PE
NLVPMVATV	A0201	CMV pp65	CTGTTAATTAGGCTC	PE

Supplementary Table 2. Information on the applied pMHC multimers. The full oligonucleotide tag are designed as follows: Biotin-C6-CGGAGATGTGTATAAGAGACAGNNNNNNNNNNXXXXXXXXXXXXXXXNNNNNNNNNCCCATATAAGAAA, with the barcode sequence indicated by 15 purple X's. C6 indicates a six carbon spacer with a hydroxyl to the 5' end of an oligonucleotide. Read2N is indicated by the black sequence, UMI's are indicated in grey, and the capture oligo is indicated in turquoise.

Supplementary Table 3. **Database cross-referencing specificities**

ct	CDR3 TRA	genes TRA	CDR3 TRB	genes TRB	peptide MHC	# GEMs	DB Match
76	CATEGDSGYSTLTF	TRAV17; TRAJ11; TRAC	CASSYQGGNYGYTF	TRBV6-5;;TRBJ1-2; TRBC1	FLYALALLL A0201	4	T
278	CALYNTDKLIF	TRAV9-2;TRAJ34; TRAC	CASSPTSGSVYEQYF	TRBV3-1;;TRBJ2-7; TRBC2	GLCTLVAML A0201	1	T
478	CAEDNNARLMF	TRAV5; TRAJ31; TRAC	CSARDGTGNGYTF	TRBV20-1;TRBD1; TRBJ1-2; TRBC1	GLCTLVAML A0201	1	T
574	CAESIGKLIF	TRAV5; TRAJ37; TRAC	CSVGAGGTNEKLFF	TRBV29-1;;TRBJ1-4;TRBC1	RVRAYTYSK A0301	1	F
1140	CATEGDSGYSTLTF	TRAV17; TRAJ11; TRAC	CASSLQGGNYGYTF	TRBV6-5;;TRBJ1-2; TRBC1	FLYALALLL A0201	1	T
1984	CIRDNNNDMRF	TRAV26-2;TRAJ43; TRAC	CASSLAPGATNEKLFF	TRBV7-6;;TRBJ1-4; TRBC1	NLVPMVATV A0201	1	T
1985	CILDNNNDMRF	TRAV26-2;TRAJ43; TRAC	CASSLAPGATNEKLFF	TRBV7-6;;TRBJ1-4; TRBC1	NLVPMVATV A0201	1	T

Supplementary Table 3. Information on the CDR3 sequences which matched the CDR3 sequences of the IEDB and VDJ databases presented in fig. 2d. Six different clonotypes (ct)

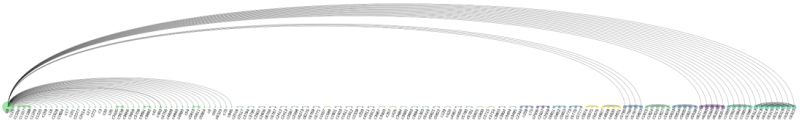
had CDR3 sequence matches. Five of the clonotypes also matched (T:True) the database on the annotated pMHC (DB Match), while one clonotype (ct 573) had conflicting annotations.

Supplementary table 4. **Multimer staining responses**

All responses reported in Fig. 7. See table enclosed

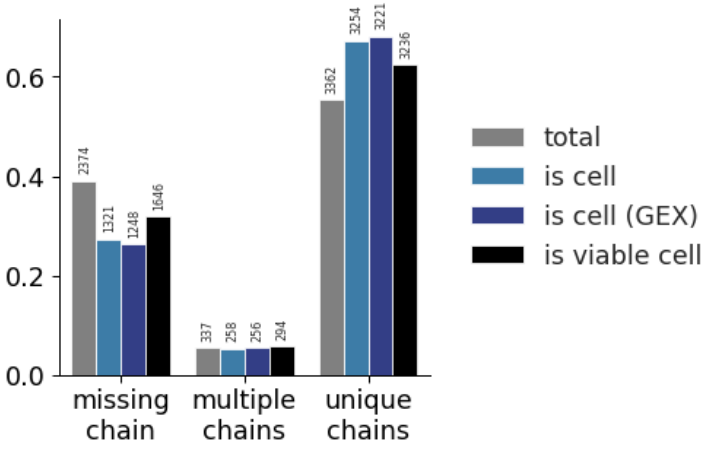
Supplementary figures

Supplementary figure 1



Supplementary Figure 1: Clonotype replicas sharing VJ-CDR3ab. Arc diagram revealing shared VJab-genes and CDR3ab sequences across clonotypes defined by 10x Cellranger. Each node is a clonotype and the size reflects the magnitude GEMs in that clonotype sharing VJab-genes with GEMs of other clonotypes. The first node (c0, green) consists of the GEMs with no 10x clonotype annotation, while the remaining (c>0) are annotations by 10x Cellranger. The diagram reveals clonotype duplets (single arc connections), triplets (2 arcs), quadruplets, quintuplets, and even a single sextuplet. Since node c0 is a mixture of GEMs that were not annotated, the GEMs in this group will match many different clonotypes. Once a c0 VJ-CDR3 matches a clonotype which already is a replicate, the GEM will of course match all of them.

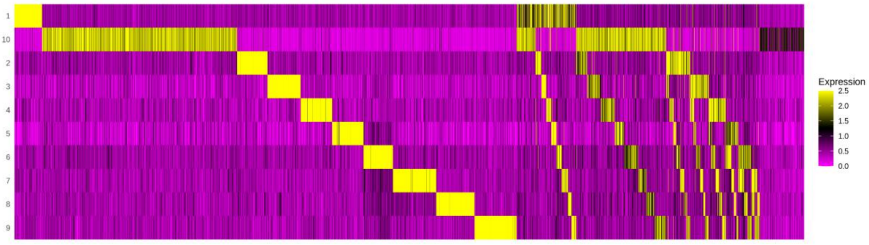
Supplementary figure 2



Supplementary figure 2: Distribution of the three categories of TCR chains across different methods of filtering. GEMs are categorized in one of three categories based on the detection of α- and β-chains: TCRs missing any chain, TCRs with multiple α- and/or β-chains, and TCRs with a unique set of one α- and one β-chain. The colors each represent a filtering step. The grey bars present the raw, total data with no filtering. The light blue bars present filtering on 10x Genomic's Cellranger "is cell" call based on transcript level of TCR sequences only.

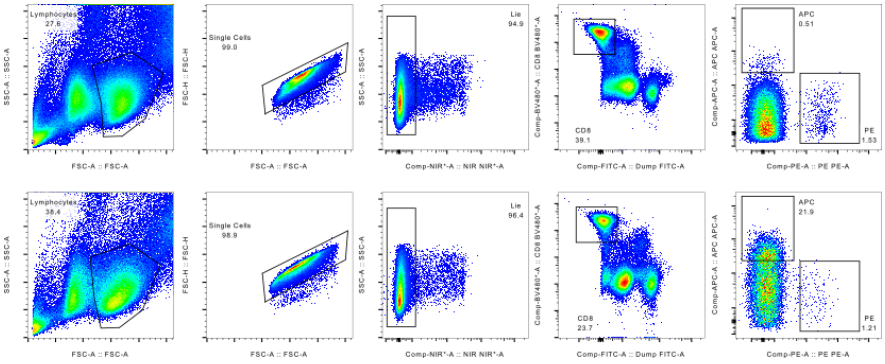
The dark blue bars present filtering on 10x Genomic's Cellranger "is cell" call based on transcript level of gene expression (GEX) sequencing. The black bars present filtering of GEX data on mitochondrial load and gene counts. For each step of filtering the counts within each category are normalized and the total value is listed above the bar. The raw data has a larger proportion of missing chain TCRs than the filtered sets. Filtering on "is cell" based on GEX data yields the largest proportion of unique chains. None of the filters completely nor substantially reduces the proportion of TCRs missing a chain or with multiple chains. See also supplementary note.

Supplementary figure 3



Supplementary figure 3: Demultiplexing cell hashing using Seurat. The GEMs (x-axis) are evaluated by the abundance of each sample barcode of 10 possible hashings (y-axis). The first section of the heatmap contains GEMs with unambiguous annotation to one sample. The second section illustrates how some GEMs contain barcodes for two samples, which might indicate a doublet, i.e. a capture of two T cells in one GEM. The last section reveals GEMs where no barcodes above a certain threshold were detected, and hence must be a result of leakage and can be discarded as noise.

Supplementary figure 4.



Supplementary figure 4: Gating strategy employed for sorting out pMHC binding MHC multimers isolated for single-cell processing.

Supplementary note

Additional filters to confidently assign the GEMs containing a cell

Removal of potential multiplets, leakage events, and dead cells

We set out to investigate how filters designed to remove potential multiplets, leakage events and dead cells would affect the distribution of GEMs between the three TCR categories: missing chain, multiple chains, and unique chains (Fig. 3). The 10x Genomics Software has a built-in method for flagging GEMs that are unlikely to contain a cell based on the transcript level of that GEM. Applying this filter based on the VDJ transcripts would reduce the set from 6073 to 4833. According to the software provider, the cell flagging method is more robust when including gene expression (GEX) data (10x Genomics 2022), which instead would reduce the set to 4725. Alternatively, the GEMs were filtered independently of Cellranger and directly on GEX data based on mitochondrial load and a minimum and maximum gene count per GEM, resulting in 5176 GEMs. The persisting GEMs should then be more likely to each contain a single viable T cell. Fig. 3 presents how filtering by the cell flag ('is cell') and viable cells affects the distribution of GEMs between the three TCR categories: missing chain, multiple chains, and unique chains. The filtered GEMs particularly contained TCRs which are missing an α- or a β-chain, however, the increased stringency of filtering did not substantially change the distribution of TCRs with unique chains relative to TCRs with missing or multiple chains. However, the filters substantially reduced the number of included GEMs (from 6073 without filters to 4725 when applying the most stringent filter).

such that the y-value of all three categories sum to 1. The distributions are shown for the unfiltered total GEMs (*total*), GEMs annotated as true cells by 10x Genomics Cellranger based on VDJ transcripts only (*is cell*), GEMs annotated as true cells when including GEX data (*is cell (GEX)*), and GEMs identified as viable cells from mitochondrial load and gene counts (*is viable cell*).

Applying hashing

The sample hashing component was predominantly observed as multiplets. In fact all, but one GEM, contained multiple sample hashing barcodes. An acknowledged method for demultiplexing SCseq data via sample hashing barcodes is the Seurat package: hashtag oligo (HTO) demultiplexing (Stoeckius et al. 2018). In short the method infers a threshold per sample barcode and thereby annotates GEMs as negative of any barcode, as singlets or doublets if multiple barcodes exceed their threshold.

Demultiplexing yielded 4,315 singlets, 1,580 doublets/multiplets, and 287 negatives (S.Fig. 9). Based on the large degree of ambient cell hashing barcode capture (Fig. 1a+b), we suspect that many of the 1,580 labeled doublets might be due to high contamination levels. Since demultiplexing was performed on the 6,073 GEMs containing both TCR and pMHC it is not surprising that only few GEMs are labeled negative.

In 320 GEMs, the demultiplexing method suggested another sample annotation than obtained from annotating by the most abundant barcode. Of the 320 GEMs, 283 were categorized as doublets and 37 singlets. The majority (227 & 35 GEMs) were originally annotated with sample 10. Since all three HLA alleles are contained in sample 10, any pMHC will inadvertently appear as having an HLA match. In only 8 GEMs the demultiplexing resulted in a different HLA profile, which corrected 7 GEMs from mismatches to matches between pMHC and sample HLA.

Technical University of Denmark
Health Technology
Section of Bioinformatics

Kemitorvet 204, 257
2800 Kgs. Lyngby

www.healthtech.dtu.dk