UiO : University of Oslo

High-throughput sequencing of gluten-specific T cells in celiac disease

Thesis submitted for the degree of Philosophiae Doctor

Ying Yao

Department of Immunology
Institute of Clinical Medicine
Faculty of Medicine

2020

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-801-4

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Table of Contents

Acknowledgement

The work of the thesis was performed at the Department of Immunology, Oslo university Hospital-Rikshospitalet and University of Oslo. The work is funded by Research Council of Norway, and grants from the Stiftelsen Kristian Gerhard Jebsen.

First of all, I want to thank my primary supervisor Shuo-Wang Qiao. She walked me through all the stages of the writing of this thesis; provided me many precious opportunities and constant support during my PhD. There are many things she taught me, from basic knowledge in immunology to scientific writing, but nothing was more precious than her enthusiasm and dedication to science. Another important person is my co-supervisor Gier Kjetil Sandve, who shared his expertise in bioinformatics when I struggled with problem. He is always inspiring me with his critical thinking and rigorous academic attitude.

I would like to express my gratitude to Asima, Lucasz, Noami for their worthy laboratory work; Victor Greiff for helpful comments and suggestions in data processing; Knut E. A. Lundin for providing patient material; Ralf, Milena and Gabriel for their contribution to the researches.

I would like to thank the members of the Gutfeeling group in the immunology department; Chakri, Knut, Ivar, Ankush, Boris, Dmytro in the bioinformatics department. We had a lot of interesting discussions and joyful time. Thanks for creating such a great working place.

Last but not least, special thanks must go to my beloved family for their loving considerations and great confidence in me all through these years.

Oslo, June 2020
Ying Yao

Abbreviations

APC	antigen presenting cell
AUC	area under the ROC curve
CCA	canonical correlation analysis
CD	celiac disease
CDR	complementarity-determining region
CMV	cytomegalovirus
HLA	human leukocyte antigen
MHC	major histocompatibility complex
MMN	mutual nearest neighbor
NGS	next generation sequencing
pMHC	peptide:MHC complex
PBMC	peripheral blood mononuclear cells
PCA	principal component analysis
PCs	principal components
ROC	receiver operating characteristics
SNN	shared nearest neighbor
TCR	T cell receptor
TCRα	T cell receptor α chain
TCRβ	T cell receptor β chain
Tfh	follicular helper T cell
Th	T helper cell
TPM	transcripts per million
Treg	regulatory T cell
tSNE	t-Distributed Stochastic Neighbor Embedding
UMAP	uniform manifold approximation and projection for dimension reduction
UMI	unique molecular identifier

List of publications

Paper I

Ying Yao, Asima Zia, Ralf Stefan Neumann, Milena Pavlovic, Gabriel Balaban, Geir Kjetil Sandve, Shuo-Wang Qiao. T Cell Receptor Repertoire as a Potential Diagnostic Marker for Celiac Disease. *Manuscript under review Clinical Immunology. 2021 Jan; 222*

Paper II

Ying Yao, Asima Zia, Łukasz Wyrożemski, Ida Lindeman, Geir Kjetil Sandve, Shuo Wang Qiao. Exploiting antigen receptor information to quantify index switching in single-cell transcriptome sequencing experiments. *PLoS One. 2018 Dec 5;13(12)*

Paper III

Ying Yao, Łukasz Wyrożemski, Knut E. A. Lundin, Geir Kjetil Sandve, Shuo-Wang Qiao. Differential expression profile of gluten-specific T cells identified by single-cell RNA-seq. *Manuscript under review*

Introduction

1.1 Immunity

The immune system protects organisms from invading pathogens and harmful substances. In higher organisms, it is commonly comprised of two components, namely innate immunity and adaptive immunity. As the first line of immune protection, innate immunity springs into action immediately after a pathogen enters the organism. In contrast, adaptive immunity acts through a slower but more effective, specific response and provides protection from later re-exposure to the same antigen. T cells and B cells are two main types of cells carrying out the adaptive immune response. T cells have two main subtypes, namely CD4 and CD8, characterized by the co-receptor expressed on the cell surface. A CD4 T cell, also known as T helper cell, has the function of activating the B cells and macrophages. A CD8 T cell, known as cytotoxic T cell, can destroy infected cells and cancer cells. Antibodies secreted by the activated B cells travel along the bloodstream and bind to the foreign antigens they recognize, thereby inactivating them or marking them attractive targets for the immune system to destroy. For the immune system to function properly, it has to detect a variety of harmful antigens as well as distinguish them from the normal tissue. Dysfunction of the immune system would result in either autoimmune disease when it raises immune response to a healthy tissue, or immunodeficiency when it fails to detect certain harmful antigens.

1.2 MHC class II

Major histocompatibility complex (MHC) molecules are a class of transmembrane proteins that are essential for presenting antigens on cell surface to T cells. As opposed to MHC class I which is expressed on all nucleated cells, the expression of MHC II is restricted on the surface of dendritic cells, macrophages, and B lymphocytes, due to which these three types of cells are collectively known as professional antigen-presenting cells (APCs). MHC class II are also expressed on epithelial cells in the thymus and interact with T lymphocyte precursors as they mature. In contrast to MHC class I which presents antigens from intracellular pathogens, MHC class II presents antigens from extracellular pathogens (1). Professional APCs generate antigenic peptides by the degradation of extracellular pathogens and present them bound to MHC class II on the cell surface, thus enabling the peptides to be recognized by the corresponding epitope specific CD4 T cells. The MHC class II molecules consists of two similarly sized transmembrane polypeptides, α chain and β chain. In humans, the MHC is called the human leukocyte antigen (HLA) complex. Among MHC class II proteins in human,

there are three highly polymorphic molecules which present peptide antigens to CD4 T cells, HLA-DP, HLA-DQ, HLA-DR. On each of the HLA locus where the three highly polymorphic molecules are encoded, denoted as DPA1, DPB1, DQA1, DQB1, DRB1, DRB3, DRB4, DRB5, there are numbers of functional alleles, except DRA that is invariable. The variation between MHC allotypes is concentrated in domains that bind peptide and T cell receptor, thereby determines the types of peptides that each MHC allotype binds, as well as the recognition by T cells. The polymorphisms in MHC genes are important for increasing the scope and strength of T cell immunity. Moreover, it has been shown that different HLA variant is related with different autoimmune and infectious diseases. For example, rheumatoid arthritis is found to be associated with certain alleles of HLA-DRB1 gene in European and Asian populations (2)(3)(4). Other diseases that show HLA association include celiac disease (CD), psoriasis, ankylosing spondylitis, systemic lupus erythematosus, and multiple sclerosis (5).

1.3 T cells

T cell is a type of lymphocyte that matures in the thymus and plays an important role in cell mediated immune response. Depending on the MHC class on the thymic epithetical cell it interacted during development, a lymphocyte could differentiate into either naive CD4 or naive CD8 T cell marked by the co-receptors CD4 or CD8 glycoprotein expressed on its surface. The lymphocytes interact with peptide-MHC II would become naive CD4 T cells, while those interact with peptide-MHC I would become naive CD8 T cells. Before the T cells mature and leave the thymus, they have to undergo positive selection and negative selection, where only a small subpopulation of thymocytes with successfully rearranged T cell receptor (TCR) that bind self-peptide:MHC complex with intermediate affinity would survive and differentiate into mature T cells. CD8 T cell, also known as cytotoxic T cell, can destroy infected cells and tumor cells mainly by releasing cytokines and interleukins upon recognizing peptides presented by MHC class I (6). The mature CD4 T cells have a variety of functions, such as helping B cells in the differentiation into plasma cells that produce antibodies, inducing macrophages to develop enhanced microbicidal activity, producing cytokines, as well as recruiting granulocytes to sites of infection (7).

1.4 CD4 T cells

In cell mediated immune response, naive CD4 T cells can be activated after recognition of its cognate antigenic peptides presented by MHC II on the surface of APC, such as dendritic cell. The activation process requires several signals. The primary one is TCR recognition of antigenic peptides presented by MHC molecules on the surface of APC, and this interaction is augmented by the CD4 co-receptor binding to the MHC class II. The second signal required occurs when co-activating molecules CD28 on the T cell bind costimulatory proteins on the APC. As a response of activation, naïve CD4 T cells proliferate by clone expansion and differentiate into effector T helper cells, where different cytokines (also sometimes known as signal 3) are required for inducing the differentiation into different subtypes of T helper (Th) cells, each with a distinct cytokine profile. The most well-established T helper cell subsets are Th1, Th2, Th17, regulatory T cells (Treg) and follicular helper T cells (Tfh) (8). Each of the subtypes secrete a set of specific cytokines with important pathogenic and protective functions (**Figure 1**).

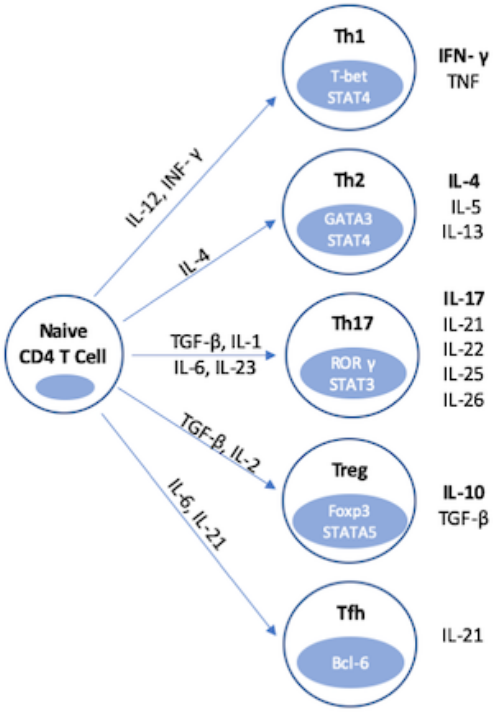


Figure 1. Different CD4+ T cell subsets. *Different CD4+ T cell subsets are differentiated from naive T cells under the influence of different cytokines. Each CD4+ subset produces a different set of cytokines controlled by the activation of different subset-specific transcription factors. The figure is modified from (8).*

A small fraction of the effector T cells further differentiates into either central memory T cells homing to the secondary lymphoid tissues or effector memory T cells in the infected tissues. Unlike the short-lived effector T cells, memory T cells has a potential for long-term survival, and therefore would be able to mount a faster and more vigorous immune response upon recognizing their cognate antigen in the future.

1.5 T cell receptor repertoire

T cell receptor (TCR) is a heterodimer composed of two highly variable protein chains on the surface of T cells. TCR of human T cells normally consists of an α chain and a β chain, whereas the TCR of around 1%-5% of human T cells consists of γ and δ chains (9). In the human genome, α chain locus consists of a constant region (C), approximately 70 variable (V) gene segments and 61 joining (J) segments. β chain locus consists of two constant regions (C), 52 variable (V) gene segments, 2 diversity (D) segments and 13 joining (J) segments (10). During T cell development, highly variable TCR chain is generated through genetic recombination of different V (D) J gene segments as well as random deletion and/or insertion of nucleotide at the junction regions. Specifically, recombination occurs between variable (V) and joining (J) segments for the α and γ chains, whereas for the β and δ chains, the recombination occurs between V, J and D segments (11). Each chain has three complementarity-determining regions (CDRs) in the variable domain which is known to be the structure for recognition specificity. In contrast to CDR1 and CDR2 that are encoded by germline sequences, CDR3 is the most highly variable region as a result of the random nucleotide deletion and insertion in the V(D)J junctions during the TCR generation process (**Figure 2**). CDR3 is the main region for recognizing the antigenic peptides (12).

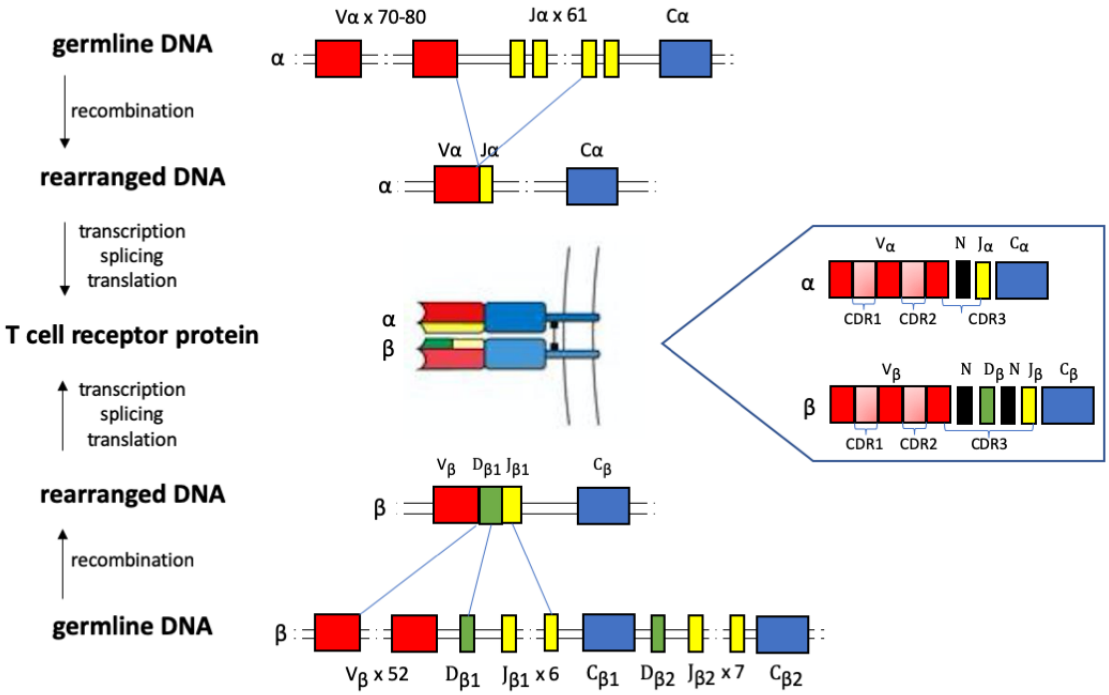


Figure 2. Generation of TCR. *For the generation of TCRα chain, DNA is rearranged through recombination of the germline V_α, J_α gene segments and some random deletion and/or insertion of nucleotide denoted by N at the junction region, which is then transcribed and spliced with C_α gene segment. Similarly, TCRβ chain is generated through recombination of the germline VDJ gene segments, with random deletion and/or insertion of nucleotide at the two junction regions. The two C_β gene and their associated D_β, J_β segments allow a further remedial attempt if rearrangement at one β chain locus is non-productive. For both TCR chains, CDR1 and CDR2 are encoded by germline sequences in the V region. CDR3 region straddle the VJ junction for α chain and VDJ junction for β chain. The figure is modified from Figure 5.3, The Immune system, 3rd edition (Garland Science 2009).*

All the highly variable TCRs in a human body constitute a highly diverse TCR repertoire, the diversity of which is the key to maximizing potential coverage of the protective immunity. Theoretically, the VDJ recombination of TCR genes can result in a total number of TCR clonotypes that ranges from 10^{15} to 10^{20}. However, due to the presumably non-random recombination process, the diversity of TCR repertoire in a human being was estimated to be at around 10^{13} clonotypes (13). By contrast, the total number of T cells in a human body is believed to be much smaller, i.e. at the order of 10^{12} (14)(15). The highly diverse TCR repertoire of an individual is both modulated by the MHC polymorphism during the thymic selection processes, and skewed towards some specificities resulting from past antigen exposure as a result of the T cell clonal expansion after antigenic stimulation. The clonal expansion in turn result in a large number of clones that share identical TCRs, some of which would transit into memory T cells persisting for decades after antigen clearance (16).

In recent years, advances in high-throughput immunosequencing allow millions of immune cell receptor sequences to be generated in parallel, thus largely improve the understanding of adaptive immune repertoires. Diagnostic of immunological diseases has become one of the potential clinical applications of this technology.

1.6 Celiac disease

Celiac disease (CD) is a chronic autoimmune disorder resulting from mis-appropriate immune response toward ingested gluten proteins found in wheat, barley and rye. In addition, oats are also reported to be unsafe for a small fraction of CD patients (17). Typical symptoms of CD include chronic diarrhea, abdominal distention and malabsorption (18). In the small bowel, the lesion caused by the abnormal inflammatory reaction is usually characterized by blunting of villi and lymphocyte infiltration (19). The most accurate diagnosis is the histopathological changes observed in small intestine biopsies. Serology-based diagnosis is also used in children as the combination of total IgA and IgA class antibodies against transglutaminase 2 (TG2) was shown to be more accurate than other test combinations (20).

The vast majority of CD patients have HLA-DQ protein either encoded by HLA-DQ2 allele or HLA-DQ8 allele, specifically HLA-DQ2.5 (HLA-DQA1*05/HLA-DQB1*02, expressed by 90% of CD patients). Patients with HLA-DQ8 (HLA-DQA1*03/HLA- DQB1*03:02) and HLA-DQ2.2 (HLA-DQA1*02:01/HLA-DQB1*02) (21) (22) (23) account for the remaining CD population. The strong HLA-DQ2.5 allele association with CD was explained by higher expression of HLA-DQA1*05/HLA-DQB1*02 genes than non-predisposing alleles (24). The other functional explanation is that these HLA class II molecules bind to gliadin peptides more tightly than other HLA class II molecules, thus increasing the risk of activating T cells (25).

A large number of gluten peptides have been identified as gluten T cell epitopes in CD. These epitopes are restricted by different HLA-DQ molecules, especially for HLA-DQ2.5 (26)(27)(28)(29). Among the large number of identified gluten T cell epitopes, a few of them were commonly recognized by a substantial proportion of gluten-specific T cells in CD patients with HLA-DQ2.5 (30). Moreover, different efficiency of these gluten epitopes to induce the T cell responses was reported to be influenced by factors such as resistance to proteolytic degradation, substrate affinity to TG2 and specificity to HLA molecules (31)(32).

TG2 deamidates the gluten peptides and the introduction of negatively charged glutamic residues enables the gluten peptides to bind better to HLA-DQ2 and HLA-DQ8 molecules. Thus, deamidated gluten T cell epitope peptides can stimulate immune system more efficiently (33). TG2 can also catalyze the formation of gliadin-TG2 complex by crosslinking of gliadin to itself, thus creates antigenic epitopes that can be recognized by TG2-specific B cells (34)(35). In such a way, it boosts the production of anti-TG2 Abs with the help of gluten-specific T cells (36). The above hapten-carrier model explains the observation that the presence of circulating anti-TG2 Abs coincides with gluten ingestion and clinical presentation of CD, and suggesting that the cooperation between T cells and B cells is crucial in the pathogenesis of CD (19). Taken together, the gluten peptides presented by the disease predisposing HLA-DQ proteins on the surface of APC to naive CD4 T cells provide signal for T cell activation and proliferation. CD associated B cells are primed upon binding either deamidated gluten peptides or gliadin-TG2 complex, so that they are able to differentiate to plasma cells producing anti-gluten or anti-TG2 antibodies with the help of the activated gluten-specific CD4 T cells. Moreover, such an amplified loop selects T cells specific for peptides that are good substrates for TG2 (37).

1.7 Gluten-specific T cells in CD

Although several studies (38)(39)(33)(40) have suggested that gluten peptide p31-43/49 may directly activate the innate immune system, CD4+ T cells that recognize gluten peptides bound to predisposing HLA-DQ protein are known as a key player in the pathogenesis of CD (41)(42)(43)(44). Importantly, the gluten-specific CD4+ T cells were only found in the small intestine of celiac disease patients, but not in healthy controls (43)(45). It has been suggested that the mucosal changes in untreated CD patients or treated patients after gluten challenge are initiated by activated gluten specific T cells that produce various cytokines (46). For instance, IFN-γ level was reported to be markedly increased in duodenal mucosa of CD patients after gluten exposure both in vivo and in vitro (47), As a marker of Th1 cell, IFN-γ alone or in combination with tumor necrosis factor (TNF-α) are cytotoxic to epithelial cells (47)(48), The other cytokines reported to be produced by CD4+ gluten-specific T cells upon activation include IL-2, IL-4, IL-6, IL-8, IL-10, IL-5, IL-21, and transforming growth factor (TGF)-β (47)(49)(50). In a recent mass cytometry analysis of CD4 T cells in the blood and intestines of CD patients, the gluten-specific cells display a distinct profile of surface proteins, including upregulated T cell activation makers such as CD38, CD161, CD28, HLA-DR,

OX40 and downregulated exhaustion marker killer cell lectin-like receptor subfamily G member 1 (KLRG1), the expression of the which was consistent with the result of bulk RNA sequencing (45).

Studies focusing on TCR repertoire of the CD patients discovered biased TCR V-gene usage, and in some instances preferred TCRα and TCRβ pairing, for gluten specific CD4 T cells towards each of immunodominant gluten epitopes, namely DQ8-glia-α1(51)(52), DQ2.5-glia-α1a, DQ2.5-glia-ω1(53), DQ2.5-glia-α2 (54)(55) and DQ2.5-glia-ω2(55). Particularly, T cell clonotypes with identical amino acid sequences were observed in multiple CD patients. These TCR shared across individuals are known as public TCR sequences.

1.8 HLA-DQ tetramer

Isolation of antigen-specific T cell was historically challenging due to the low affinity between TCRs and their pMHC counterparts. In 1996, the advance of MHC multimers, typically tetramers, facilitated the identification and isolation of antigen-specific T cells by using a complex of multiple MHC molecules covalently linked with antigenic peptides (56)(57). Among multiple HLA-DQ2.5-restricted gluten T-cell epitopes, a few epitopes were prevalently observed in CD patients, i.e. DQ2.5-glia-α1a, DQ2.5-glia-ω1, DQ2.5-glia-α2 and DQ2.5-glia-ω2 (43)(58). In recent years, HLA-DQ2.5-gluten tetramers carrying these four immunodominant gluten epitopes have been used to stain and visualize gluten-specific T cells directly from blood or small intestinal tissue from CD patients (59)(60).

1.9 DNA sequencing technologies

Advances in sequencing technologies have caused a revolution in many fields. The first Sanger sequencing method began in 1977 (61). A key technique was the use of fluorescence-labeled dideoxynucleotides (ddNTPs), which serve to randomly terminate the replicating chains. Many randomly terminated chains from the same DNA template then migrates on a polyacrylamide gel tray under electrostatic forces with a speed inversely proportional to their length. The sequence of the DNA template is recorded by capturing the color of fluorescence on ddNTP at the end of the gel tray. Roche's 454 Sequencer further simplified the preparation process and was introduced in 2005. It vastly improved the throughput by analyzing a large number of samples in parallel. Therefore, it is known as the second generation of sequencing technology. Now we have come to the era of next-generation sequencing (NGS) or high-

throughput sequencing. The cost of sequencing has decreased dramatically as more players joined the market. Illumina sequencers, which also rely on the chain termination method, distribute randomly fragmented chains of 200-300 base pair length to the surface of a flow cell with large number of adaptors for further replication through bridge PCR. Each fragment is replicated, forming a cluster of DNA chains. The sequences are then recorded by capturing the fluorescent signals emitted from each cluster as each new base is added. The sequencer became the most efficient at that time since each flow cell analyzes approximately 150 million clusters. In addition, there are several other sequencing technologies, such as long-read technologies from PacBio, Ion Torrent, SOLiD as well as Nanopore sequencing from Oxford Nanopore Technologies with lower throughput but user defined read length up to 2,272,580 base pair (62).

1.10 Transcriptomics

RNA as an intermediate product between DNA and protein is a useful tool in understanding the function of cells. Before RNA-seq, microarrays (63) was widely used in gene expression studies. But since it requires species-specific or transcript-specific probes, it is incapable to detect unknown changes, such as novel transcripts, gene fusions, and mutations. The range of gene expression measurements with the array hybridization technology is relatively narrow due to the limitation of background at the low end and signal saturation at the high end (64). With the advances of NGS, transcriptome can be profiled without any prior knowledge of novel changes. Moreover, it is free from cross-hybridization artifacts, and rare transcripts can be captured by increasing the sequencing depth (65).

1.11 Single cell transcriptome sequencing

Traditional bulk sequencing measures the expression of RNAs from large population of cells, therefore the heterogeneity between the cells is obscured. Single cell transcriptome sequencing has enabled the generation of transcriptomic data on the level of individual cells. The higher resolution of transcriptional data provides an unprecedented opportunity for exploring the cellular identity, dynamics and function, especially for cells of rare type. Many methods and protocols for sc-RNA sequencing have been developed in the last few years. Typically, protocol for sc-RNA sequencing consists of cell capture, mRNA reverse transcription, cDNA amplification, cDNA library preparation and sequencing. For single-cell capture strategy, low throughput manual approaches, such as mouth pipetting or laser capture

microdissection, are useful for sample containing rare targeted cells. By allowing cell enrichment by fluorescent labels and increasing the efficiency for cell capture, fluorescence-activated cell sorting has gained more popularity, and is widely applied in such as Smart-seq2 (66), CEL-seq/2 (67), MARS-seq (68), and STRT-seq (69). Fluidigm C1, a microfluidic based method with comparable efficiency for cell capture, provides a more integrated system enabling an automated process for cell capture and down-stream molecular biology steps. The droplet-based methods, such as 10x Genomics Chromium (70) and drop-seq (71), are of highest throughput but with a typically lower coverage, therefore are ideal for studies that require a large number of cells, for example, study for identifying cell subpopulations of complex tissues. Each protocol takes different strategies for mRNA reverse transcription and cDNA amplification. For instance, Tang-seq (72)(73) adopts poly(A) tailing followed by PCR; CEL-seq/2, MARS-seq use second-strand synthesis followed by in vitro transcription (IVT), where the sequencing covers only the 3'-end of mRNA due to the premature termination of reverse transcription; Likewise, 10x Genomics Chromium and drop-seq also adopt a tag-based method that only capture either 5'-end or 3'-end of mRNA; Smart-seq2 is a full-length protocol with higher sensitivity than the other methods evaluated in , including CEL-seq2, Drop-seq, MARS-seq, SCRB-seq and Smart-seq.

A well-known challenge of the single cell RNA-seq technique is the larger technical variations and dropout, i.e. a transcript cannot be not detected especially for those genes of low expression. The technical variations could arise from various phases of library preparation, such as reverse transcription and PCR amplification, as well as the process of sequencing. To alleviate the problem, the use of unique molecule identifier (UMI) and spike-ins, such as External RNA Control Consortium (ERCC) can be incorporated into certain scRNA-seq protocols. UMIs consisting of a certain number of random nucleotides serve as molecule tags that are added to cDNA segments during reverse transcription. The use of UMI allows for correcting the PCR amplification bias by collapsing reads with identical UMI, since they are considered to be PCR-amplified from the same mRNA molecule. ERCC is a collection of external RNAs with known sequences and concentration that can be added to each sample. Quite a few purposes of using ERCC have been published, including for quality control, read counts normalization and identification of highly variable genes. However, some usage is controversial considering that ERCCs often exhibit higher noise than endogenous genes due to contingent pipetting errors or mixture quality.

1.12 Index switching

Multiplexing is a common practice in single cell sequencing. By adding a unique index or combination of indices to each sample, it allows materials from large number of samples to be pooled and sequenced simultaneously in a single run. However, it has been reported that certain reads could be assigned to a wrong sample as a result of index switching or index hopping. Index switching is essentially a molecule recombination between endogenous sequences and free-floating indices during the sequencing run. While the problem of index switching is negligible on the older Illumina sequencing platforms such as MiSeq, NextSeq and HiSeq 2500, on the Illumina HiSeq 3000, 4000, X sequencing platforms using the patterned flow cell technique introduced in 2015, the level of index switching is reported to be increased to up to 10% (74)(75). The issue could be corrected by removing reads with non-existing combinations of indices if double indexing (76) was introduced with proper indexing strategy, such that each sample has a unique row index and a unique column index. However, it is not a practical solution for single cell studies considering the large number of cells usually processed and the limited number of indices available.

1.13 Bioinformatics

1.13.1 Quality control

Quality control is important for all next generation sequencing data, as samples can fail in a range of different ways. Quality control usually involves assessing the quality of data from many different perspectives, such as quality of base calling, GC content, duplicated reads, over-represented k-mers and presence of adaptors. Tools (77) and NGSQC (78) can be used to scan the sequences. Phred scores indicating the quality for base calling can be summarized by position in reads or tile of the sequencing lane, so that one could choose to trim low-quality bases from reads, or exclude all reads with low-quality. Specific contaminant such as adapter dimers might be visualized by displaying distribution of GC content in each read. Likewise, sequences of adapter dimers or rRNA may be identified by the plotting the proportion of each base position. In addition, summary of duplicated sequence usually indicates enrichments bias, such as PCR over-amplification. MultiQC (79) combines all these metrics of quality control from all individual samples and visualize them in an interactive way, therefore it is a practical tool for experiments with large number of samples, typically as in single cell studies. It often gives warning of the problematic batches or batch effects in the very early phases of bioinformatic analysis.

1.13.2 Genome alignment

Reads are typically aligned to either a genome or transcriptome in order to quantify the number of reads mapped to each gene or transcript. Various aligners have been developed, each with its own merits. Aligners including TopHat, STAR and Hisat2 have gained enormous popularity due to the ability to detect novel genes or transcripts by awareness of the gaps or splice junctions between exons. GSNAP (80), PALMapper (81), MapSplice (82) are optimized to identify SNPs (83). BWA-MEM, bbmap, Stampy, and NextGenMap deal with different issues when the reference genome is missing. Some K-mer based pseudo-alignment tools, such as Salmon, Sailfish and Kallisto are much faster, so that they are more practical for large datasets as in single cell studies and no extra step of transcript quantification is needed. Otherwise, raw counts of the mapped reads should then be aggregated on gene level or transcript level using, for instance, HTSeq-count (81) or featureCounts (85).

1.13.3 Single cell pre-processing quality filter

In single cell experiments, cells of poor-quality should be removed from the downstream analysis to mitigate the technical variation. Small number of reads or small number of detected genes usually indicate poor RNA capture efficiency, i.e. RNA molecules from the cell might have failed to be converted to cDNA or amplified. In addition, a large proportion of reads mapped to mitochondrial genes or spike-in molecules such as ERCC, could be a sign for cellular apoptosis. In addition, depending on the aligner used, some metrics from the alignment results can also be used to filter the low-quality cells, such as the mapping rate and proportion of reads mapped to exons.

1.13.4 Normalization

During library preparation and sequencing, samples are subject to different conditions causing bias in the measurement of gene expression. In contrast to data generated by bulk microarray or RNA-seq experiments, single cell RNA-seq data has higher level of these bias, which could be introduced by biological factor, such as endogenous mRNA content, as well as multiple technical factors, such as capture and reverse transcription efficiency, amplification factor, dilution factor, sequencing depth, etc. Most of the undesired factors are cell specific and some can be both cell-specific and gene-specific. Normalization is a preprocessing step in the analysis of transcriptome data that aims to correct the biases raised from those factors which are not of direct interest for the study. Ideally, the bias from technical effects should to be

corrected by normalization, while the biological signal of interests of the study should be preserved.

Following some within-sample normalization methods, such as fragments per kilobase of exon model per million mapped reads (FPKM), reads per kilobase of exon model per million reads (RPKM) and transcripts per million (TPM), a widely used class of normalization methods is based on estimating a scaling factor per library. Typical methods in this class are trimmed mean of M values (TMM), Upper Quartile and Full Quartile. Another class of methods is based on explicit regression on known confounding factors, such as batch and ERCC. Moreover, recently proposed methods such as BASiCS, remove unwanted variation (RUV) and surrogate variable analysis (SVA) are able to correct the unknown confounding factors given that suitable parameters are provided.

1.13.5 Data integration

Although the effect of technical factors can be alleviated with proper normalization, several methods for data integration were shown to outperform the traditional tools developed for microarray data, such as limma (86) and ComBat (87) in batch-effect correction (88). Aside from integrating single-cell data sets produced across different batches, individuals, protocols, more importantly, these methods allow harmonizing single cell data sets that measure distinct modalities, thus providing a better understanding. For example, integrating single-cell RNAseq data with single-cell ATAC sequencing improved the discovery of regulatory logic in different subpopulation of cells (89). One of the most popular methods was introduced by Butler.A et al. (90) in 2017, where canonical correlation analysis (CCA) was employed to identify shared correlation structures across single-cell datasets. After embedding cells from each data sets into the low-dimensional space that capturing the most correlated features, the CCA subspaces were aligned. However, due to the fact that CCA captures the most correlated features between the two data sets while ignore the others, the datasets were supposed to be based on the same set of genes. In other words, cells in each data sets should be of similar subpopulation. It is also necessary to remove non-expressed genes, as the CCA is sensitive to collinearity in the expression data. Haghverdi et al. (91) built connections between data sets by identifying Mutual Nearest Neighbors (MNN). These identified anchors were then used to calculate batch-correction vectors, which were subtracted from one dataset to integrate it into another. This MNN-based method improved the applicability as it only requires at least one cell population in both data sets, although a single small population might not suffice for

accurate estimation of local batch effects. However, another assumption that the batch-effect variation is much smaller than the biological-effect variation does not always hold in practice. In 2019, a strategy published in the third version of Seurat package (89) overcame the limitations of the two aforementioned methods. Cosine distances were used for measuring cell similarities as they were robust to differences in sequencing depth and capture efficiency between batches. Searching for MNNs in a shared low-dimensional space produced by CCA match cell successfully even in the presence of significant batch-effect, as CCA can effectively identify shared biological features and conserved gene correlation patterns.

1.13.6 Dimensionality reduction

Dimensionality reduction is useful in the analysis of large multidimensional datasets. Aside from the convenience for visualization, it also helps to reduce the noise and the cost of computing power. Principal component analysis (PCA), t-Distributed Stochastic Neighbor Embedding (tSNE) and Uniform Manifold Approximation and Projection (UMAP) are the most commonly used techniques for dimension reduction. Principal components in PCA is a linear recombination of uncorrelated variables, generated by the rule that each succeeding component has the highest variance possible under the constraint that it is orthogonal to the preceding ones. Therefore, PCA plot using only a few top components can usually capture the global structure by preserving the highly variable information. tSNE employs random walks and the nearest-neighbor network to map the high dimensional data to 2-dimensional space. Therefore, it is good at preserving local distances between cells, i.e. similar data points would be attracted in a cluster while the dissimilar ones would be repelled, in such a way that the distance between dissimilar cells or clusters are less meaningful. As with tSNE, UMAP is also a neighbor graph-based technique with improved algorithm that preserve both the local and global structure well. It can be applied on top n components of the PCA, and is much faster than tSNE.

1.13.7 Pathway enrichment analysis

The analysis of genome-scale data usually results in long list of interesting genes for certain phenotype, each associated with molecular function in a certain biological process, or expressed in certain cellular compartment. A typical challenge of these studies is to get systematic insight into the genes. Pathway enrichment analysis or gene set enrichment analysis is a computational method to identify the pathways that are enriched in the genes of interest more than could be expected by chance, thus providing a better understanding of the

biological functions of the genes and mechanisms of the disease studied. The most popular tools include DAVID, GSEA, Ingenuity and Reactome. One of the most commonly used databases Gene Ontology (GO) contains a large collection of standardized annotation of genes. By comparing the frequency of individual annotation in the gene list of interest with that in the background gene list, an enrichment score can be calculated for each pathway, indicating the significance level of enrichment. Tools as Camera (92) and Setrank (93) eliminate false positives in the significance test by incorporating some corrections based on the fact that the gene sets or pathways overlapped more with others are more likely to get significance due to the inter-gene correlation.

1.13.8 Immune adaptor sequencing from scRNAseq data

A large variety of immune cell receptor sequences can be generated by the recombination of different V, D, and J gene segments as well as random deletion and/or insertion of nucleotide at the junction regions. Due to the complexity, information about immune cell receptor constructed by using traditional alignment tools is very limited. Tools such as TraCer (94), TRAPeS (95), BraCer (96) and BASIC (97) were developed specifically to reconstruct the full-length immune cell receptor sequences from full-length scRNA-seq data and to annotate them in terms of V-, D-, J-gene usage, sequences of CDR3 region, thus facilitating the inference of clonal relationship. Therefore, single cell transcriptome data can be analyzed along with the reconstructed TCRs or inferred clonal type.

1.13.9 Methods for classification

Inference of a disease state is a typical classification problem, since the aim is to assign an individual to a category, for example, healthy or different states of disease development. Among many possible classification techniques, logistic regression, linear discriminant analysis and K-nearest neighbors are most popular. With the unprecedented improvement in computing power, some computing-intensive methods such as generalized additive models and support vector machines are also widely used.

The logistic regression is typically used when the response is binary, for example, an individual is diseased or healthy. In essence, in logistic regression and many other classification techniques, an individual is classified as being diseased if the probability of being healthy is calculated to be smaller than 0.5, and vice versa. As the log transform of odds have a nice feature that forcing the probability between 0 and 1, by assuming a linear

relationship between the log-odds (logit) of a probability and the predictor(s), the logistic function can be easily fitted and used for prediction.

Receiver operating characteristic (ROC) curve can be used to assess and illustrate the performance of a binary classifier. At various discrimination threshold settings, two operating characteristics, namely the true positive rate (TPR) and the false positive rate (FPR), are plotted in such a way that the detection probability of a classifier in the y-axis and the cumulative distribution function of the false detection probability on the x-axis are simultaneously conveyed in the ROC curve. ROC analysis is widely used to assess the performance of classifier and get the optimal models regardless of the distribution of the responses and predictors, especially in the case of diagnostic decision making.

Aims

The main goal of my PhD project is to characterize the gluten-specific T cells on the single cell transcriptome level. We used the HLA-DQ-gluten tetramers carrying the immunodominant gluten epitopes to identify and sort the disease-specific CD4+ T cells sampled from the lamina propria of duodenal biopsies and peripheral blood from untreated patients with CD. A secondary goal is to explore the potential of direct TCR sequencing in CD diagnostics.

Summary of papers

Paper I

In this proof-of-principle study, we demonstrated that the state of celiac disease could be inferred by unbiased direct TCR sequencing. Celiac disease affects around 1% of the population and the disease-associated gluten-specific TCRs are well characterized (51)(52)(53)(54)(55). By investigating the TCR repertoires of unsorted lamina propria T cells from 15 individuals, and comparing with a large database of nearly 6000 gluten-specific TCRα and TCRβ amino acid sequences, we showed that the states of celiac disease could be successfully inferred in the majority of the subjects. The result from this small study shows promise for the ultimate goal of inferring celiac disease state based on TCR sequencing of circulating T cells.

Paper II

We experienced a major set-back when it was reported that the most widely used sequencing platforms, HiSeq3000/4000/X from Illumina, are prone to erroneous read assignment due to adaptor switching of the identifying indices. Based on our single cell transcriptome data sequenced on HiSeq3000 and HiSeq4000 platforms, we utilized the unique expression of immune receptor of each T and B cell to quantify the impact of index switching on single cell RNA-seq experiments. We confirmed that index switching affects all samples run in multiplexed libraries on Illumina HiSeq3000 and HiSeq4000 platforms. By quantifying the spread-of-signal from 47 unique markers in 51 wells due to index switching, we estimated the median percentage of incorrectly detected markers to be 4.2% (interquartile range (IQR): 2.0%-8.7%). We did not detect any consistent pattern of some indices to be more prone for switching than others, suggesting that index switching is a stochastic process.

Paper III

Gluten-specific CD4+ T cells are the key drivers for the pathogenesis of celiac disease. To study the gluten-specific CD4+ T cells on the single cell transcriptome level, we conducted single cell transcriptome sequencing on CD4+ T cells sampled from peripheral blood of four untreated CD patients. Cells were sorted with a mix of HLA-DQ2.5:gluten tetramers presenting each of the four immunodominant gluten epitopes, i.e. DQ2.5-glia-α1a, DQ2.5-glia-ω1, DQ2.5-glia-α2 and DQ2.5-glia-ω2. We demonstrated that the transcriptome profiles

of the gluten specific cells were consistently different from the non-specific cells, and largely in accordant with Th1 and follicular helper T cells. Moreover, analysis on the reconstructed full-length TCRs of the CD4+ T cells suggested that cells that shared clonal origins did not show more similar transcriptional profiles compared with cells that were clonally unrelated.

Methodological considerations

4.1 Sample selection

Paper Ⅰ was a proof-of-concept study exploring the possibilities for inferring disease state by matching sequences from the sampled TCR repertoire against a priori known disease-associated TCR sequences. Although HLA-DQ2 and HLA-DQ8 present with identical clinical pictures, only HLA-DQ2-retricted TCR sequences were considered in this study. Therefore, only untreated HLA-DQ2 CD patients were included in the disease group, while HLA-DQ8 celiac patients as well as healthy individuals were grouped as controls. The diagnosis of CD was based on IgA-TG2 titer and histological Marsh score. We sampled two pieces of duodenal biopsies from each donor as it is advantageous to look at the T-cell response in the affected tissue where the frequency of disease-relevant T cells is much higher than in blood, i.e. 1-2% (98) compared to 1 per 100,000 (99)(60), respectively.

With the purpose of studying index switching, there was less restrictions on the sample selection for **paper II** as long as they were single cell samples, and full-length transcriptome was sequenced with sufficient depth, such that the reads with switched index could cover the receptor region.

To study the gluten-specific CD4+ T cells on the single cell transcriptome level, we conducted single cell transcriptome sequencing on CD4+ T cells sampled from peripheral blood and gut of untreated CD patients in study Ⅲ. By assuming that some known house-keeping genes are truly expressed in all cells, we modelled the detection failures as a logistic function of mean log transformed expression of the house-keeping genes, in line with the standard logistic model for drop-outs for quality control. The plot of the models (**Figure 3**) together with FastQC report indicated that quality of cells from three batches was significantly worse than the other one. All cells from gut were from the batches of poor quality, such that only 141 out of 371 of them passed quality control. Therefore, we mainly focused on the data of good quality and excluded the gut cells in the downstream analysis. Blood cells from the three batches of lower quality was integrated as an independent support cohort.

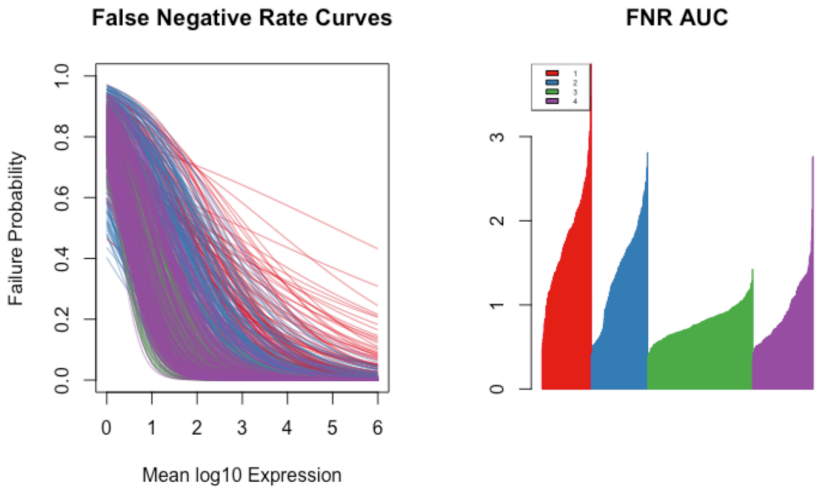


Figure 3. False negative rate for house-keeping genes for cells in each of the four batches (left). *For each cell colored by batch, detection failures of over 400 house-keeping genes were modelled as a logistic function of the mean log transformed expression. The house-keeping genes were extracted from scone (100). The corresponding area under the curve for each cell were visualised in the bar plot on the right panel sorted by batch and size. The height of each bar indicates the probability of drop-out in the corresponding single cell.*

4.2 TCR repertoire sequencing

For study Ⅰ, both bulk sequencing and single cell sequencing can be used to profile the TCR repertoire. In order to cover the gut TCR repertoire, we used bulk sequencing, as opposed to single-cell TCR sequencing which will be prohibitive for the same sequencing coverage. With single cell sequencing technique however, the paired TCRα and TCRβ information would be an advantage in terms of identifying the TCR of gluten-specific cells and therefore improve the performance.

4.3 Disease state inference

In study Ⅰ, we inferred the disease state of 15 donors by matching sequences from TCR repertoire against a priori known disease-associated TCR sequences. Although we chose to use logistic regression, a range of other methods were valid and applicable for the classification problem, such as linear discriminant analysis, K-nearest neighbors and support vector machine. ROC was an intuitive solution, as a single predictor was used. Without cross-validation, the performance tends to be over-estimated in terms of inferring the disease state of independent individuals. Together with the fact that the prediction could be less robust with the very limited sample size, we reported the result of ROC as a supplement for logistic regression.

4.4 Association between the frequency of an TCR in repertoire and its popularity rate

Concerning the association between how often an TCR clonotype was observed in our TCR repertoire data and the number of patients in the reference dataset who expressed this clonotype, correlation is typically used as a measurement. The different types of correlations are Pearson correlation (101), Kendall rank correlation (102), Spearman correlation (103), and the Point-Biserial correlation (104). Pearson correlation is typically used for measuring the association between two linearly related variables. Assumptions of two variables measured by Spearman correlation include linearity, homoscedasticity and normality, which is obviously violated in our data of frequency. Both Kendall's rank correlation and Spearman correlation are non-parametric tests for measuring the degree of association between two variables, as they do not rely on any assumptions on the distribution of the two variables. Both can be used for ordinal data as in our study. Considering the relatively higher frequency of ties in the frequency data, we chose to use Kendall's rank correlation over Spearmen.

4.5 Protocol for single cell RNA sequencing

Based on previous studies, gluten-specific cell is rarely detected in blood of CD patients. In the CD4 compartment, only 1 to 100 per million CD4 cells in blood would be expected to be specific to a given pMHC (99)(60). In order to capture sufficient gluten-specific cells for study Ⅲ, cell enrichment is required. Therefore, we chose to use bead enrichment of HLA-DQ:gluten-tetramer stained T cells prior to flow cytometry assisted cell sorting. Moreover, it is preferable for study Ⅱ and Ⅲ to quantify the expression of more genes with high accuracy. From the perspective of quantification, there are two categories of scRNA-seq methods. Methods such as Quartz-Seq, Smart-seq2 (66), SUPeR-seq and MATQ-seq are based on full-length, with the aim to achieve a uniform read coverage of each transcript. The other type of methods such as Drop-seq (71) , SEL-seq2 (67), InDrop-seq (105), MARS-seq (68), Seq-Well (106) and STRT-seq (69) only capture either the 5'- or 3'-end of each transcript. The full-length protocols provide the possibility to perform analysis on clonotype of the single cells in combination with the full-length transcriptomic profiles. Moreover, they were reported to have better sensitivity for detecting low-expressed genes (107). Considering all these issues as well as the cost, Smart-seq2 was used for single cell library preparation, as it is a full-length based protocol with high sensitivity (107)(108).

4.6 Quantification of genes

Traditional challenges for gene expression estimation include the requirement of substantial computational resources, the ambiguity in read mapping caused by alternative splicing, and non-uniform sampling of reads. For quantification of the transcript expression we chose to use Salmon (109), which is a k-mer based pseudo-aligner. It achieves a high speed in the same order of magnitude as some other aligners i.e. kallisto (110) and sailfish (111), by applying a lightweight alignment, which aims to get the original transcript of a read instead of the exact alignment. In the meanwhile, by using two phases of statistical inference and models accounting for sample-specific bias, better accuracy is achieved by accounting for alternative-splicing, sequence-specific bias, GC-content bias and positional bias. After testing some values of the parameter k on a few samples, we found that the default k-mer length of 31 worked well. Another benefit of Salmon is that we do not have to trim the reads before gene quantification, even though the base quality normally tends to drop at the end of the reads. The reasoning is that the pseudo alignment or lightweight alignment process of Salmon is based on super maximal exact matches (MEMs). In such a way that any nucleotide mismatch outside the region of the MEMs would not stop the read from being mapped as a whole, but only result in a proportional reduction of the estimated probability of the read originating from a transcript.

4.7 Cell selection for quality control

In single cell RNA-seq experiment, it is commonly observed that a portion of cells are of bad quality due to various reasons. For example, some cells could be apoptotic; RNAs in some samples are not efficiently converted into cDNA or amplified; and that some samples contain no cell or multiple cells. Those low-quality cells need to be excluded before analysis so that they would not distort the interpretation of the result. In **paper III**, we applied criteria from 5 metrics for excluding the low-quality cells based on their distribution. The 5 thresholds were chosen by assuming that most of the cells are of good quality. Specifically, cells with lower mapping rate than 30% were excluded, since a low mapping rate may indicate RNA degradation or empty wells. The library size, defined as the total number of counts over all genes in a single cell sample were also applied as a filter by removing samples with sizes smaller than 200,000. A low library size indicates a low RNA capture efficiency for the sample. Library size is in large correlated with the number of detected genes (including the

exogenous spike-ins). For the purpose of removing samples that potentially contained more than one cell, we also imposed a filter with upper boundary of the number of detected genes, in such a way that cells with number of detected genes ranged from 1,800 to 15,000 were retained. Another metrics for measuring sample quality is the proportion of read counts mapped to mitochondrial genes. High proportion of mitochondrial genes is usually indicative of low-quality, owing to the fact that the cytoplasmic RNAs tend to be lost while the mitochondrial RNAs were left in a broken or apoptotic cell (112)(113). Based on the overall distribution, samples with a proportion of read counts mapped to mitochondrial genes larger than 15% were excluded. Similarly, we also excluded samples with larger than 40% reads mapped to ERCC spike-ins. The amount of ERCCs added to each single cell sample were constant, therefore loss of endogenous transcripts would cause an increasing proportion of reads mapped to ERCCs.

We added ERCC spike-ins in two of the batches from the study, i.e. samples generated from patients CD1507 and CD1517, but stopped using it for the other batches based on the following three considerations. First, although it is useful for quality control on cells, when we applied the quality control scheme with the aforementioned five metrics, only 5 more cells were identified as low-quality compared to the scenario without considering ERCC where 256 low-quality cells were identified out of 576 cells in total; Second, the read counts from ERCC spike-in RNAs are noisier, as their variations were higher than majority of endogenous genes. This suspicious fact impacts the reliability of using analytical methods where ERCC spike-ins is used for normalization or selecting highly variable genes. Moreover, a considerable number of total reads, specifically around 7.4%, were generated from ERCC spike-ins.

4.8 Cell clustering and visualization

We chose to use Seurat 3.0.2, a widely used R package for single cell data analysis. It provides implementations of analytical methods for a variety of purposes, including identification of highly variable genes, data visualization, unsupervised clustering, and identification of differentially expressed genes.

Results and discussion

5.1 T Cell Receptor Repertoire as a Potential Diagnostic Marker for CD (Paper I)

Guided by the hypothesis that an individual's T cell receptor repertoire is skewed towards some specificities as a result of past antigen exposure, a few previous studies have shown that TCR clonotyping can be used as a diagnostic tool for infectious disease such as CMV serostatus (114). By profiling TCR repertoire of 15 donors, we explored the possibilities for inferring the state of CD, for which the T-cell response is less pronounced compared to the strong CMV response.

As a reference of CD associated TCRs, an external database comprised of 2,929 TCRα- and 2,662 TCRβ-clonotypes obtained from 6,808 single-cell TCR sequencing of HLA-DQ2.5-gluten-tetramer-sorted cells from 59 CD patients was utilized. From this large reference database, we generated a subset of public clonotypes (observed in at least two CD patients) that possibly represent a more CD-specific reference database. By matching TCRα- and TCRβ-clonotypes from TCR repertoires of the 15 donors against the above two references, 39 TCRβ clonotypes out of the 226 public TCRβ references were observed in our data. The number of matches increased only by 1.4-fold when we used the reference database comprised of all gluten-specific TCRβ, which was 14 times larger than public TCRβ (**Paper I, Figure1**). It confirmed our assumption that the public clonotypes were more specifically associated with CD.

In order to explore the effect of sequencing depth on the predictive accuracy in our gut repertoire data, we employed a bootstrap procedure. Each time, one individual was randomly set aside for testing and a logistic regression model was trained on 14 cases which were randomly sampled with replacement from the 14 remaining cases. After repeating this process for 1000 times, we calculated the average predictive accuracy for each individual that reflects the probability of being correctly inferred. By visualizing the repertoire size versus the predictive accuracy (**Figure 4**), the predictive accuracy was satisfactory for the individuals with a repertoire size over 2000. However, much more cautions are needed to generalized the observation due to the limited number of individuals.

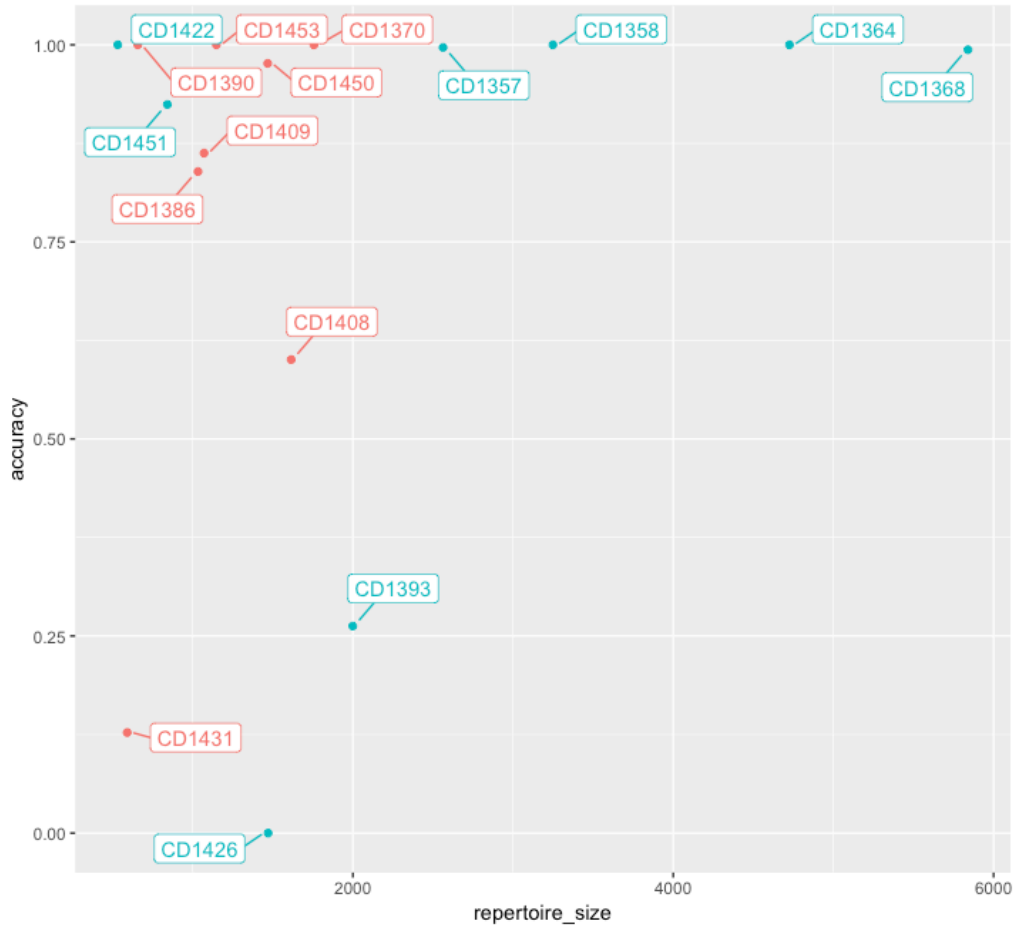


Figure 4. Effect of sequencing depth on the predictive accuracy. *Repertoire size is represented as number of unique TCR-beta chains detected from each individual. The accuracy on y axis is the frequency that the individual has been correctly inferred when it is selected as the testing case in bootstrapping.*

There is noticeable difference between the total number of detected TCRα and TCRβ clonotypes in our data, which could be a reason for the fact that the proportion of matches in the reference database of TCRβ clonotypes was higher than those in the TCRα compartment. In most published studies where both TCRα and TCRβ have been sequenced, there is more TCRβ data, either as more numerous clonotypes in bulk data or as more read counts and higher sequencing rate in single-cell data, compared to TCRα (115)(116)(117)(118). Most likely it reflects a biological inequality in the transcription rate, where there, for reasons not known, is more TCRβ transcripts than TCRα transcripts per cell. Consequently, the result of the inference (**Paper I, Figure 2**) suggested that matching TCRβ alone was sufficient for predicting celiac disease state, while the inference by TCRα clonotypes performed rather poorly both when used on its own and in combination with TCRβ. There are two hypotheses

for the cause, which can be tested in further studies: the TCRα chains are structurally less relevant for gluten-peptide recognition, or an TCRα clonotypes is disease-specific only if it is paired with certain TCRβ chains. The later hypothesis can be tested by single-cell TCR sequencing where the paired TCRα and TCRβ information is available.

Considering previously reported systematic differences in the human TCR repertoire with ageing (119), we have checked the association between age and our predictor variable. Based on the result, we do not believe that confounding by age played a role in our predictions. However, we cannot completely rule out the possibility, as our sample size was limited. Potential confounders should be examined in larger cohorts.

Since it is confirmed that the public clonotypes are more specifically associated with CD, we believe that the database of disease-associated TCRs can be improved by further step of validation. The current reference database for gluten-specific TCR clonotypes was sampled and generated from CD patients, and the public gluten-specific TCR clonotypes were treated equally regardless of the number of patients across which the given TCR was observed. The specificity of those gluten-specific TCR clonotypes in the reference could be further improved by involving more TCR repertoires from healthy controls for training and purifying. We believe the current reference database has covered the majority of the gluten-specific TCR clonotypes in HLA-DQ2.5+ CD patients, as the percentage of matching (0.22%-1.1%) was comparable to a recently published study (45) where the cells binding to the same tetramers comprised 0.3–1.5% of the total small intestinal CD4+ T cells. As public TCR sequences are continually accumulating by the inclusion of more patients, we believe the database of public gluten-specific TCRs would approach to saturation, where we would have a complete collection of public gluten-specific CD4+ TCR sequences when barely any new clonotypes were gained by including more patients.

5.2 Quantify Index Switching (Paper II)

The advance of next-generation sequencing technology comes with a risk of index switching, which is essentially molecular recombination occurring in Illumina Hiseq platform, resulting in certain amount of sequencing reads being assigned to an incorrect sample. We demonstrated that the immune cell receptor information can be utilized to detect and quantify index switching in single cell experiment. The estimated percentage of incorrectly detected markers ranges from 2.0% to 8.7%. It was in the same range as reported in (75), but higher

than it was reported in (74) (120), where the rate of index switching was up to 2% and 0.47% respectively. We speculate that those differences might be caused by the chosen method for measurement, the level of free-floating adaptors in library preparation and the variation of sequencer. We did not observe any significant difference in index switching level between Illumina HiSeq3000 and HiSeq4000.

We also confirmed that index switching is a stochastic process in which the occurrence is not biased towards any indices, which could be a fundamental basis for developing computational methods for correcting the effect of index switching. Although there is a published tool (121) for computationally correcting the index switching effect, it is more practical to either eliminate the possibility through experimental design, i.e. use a sequencing platform less susceptible to index switching, or apply an indexing strategy enabling reads with switched indices to be removed downstream, e.g. by using dual indexing with unique indices at both ends.

5.3 Single cell transcriptomic analysis of gluten-specific T cells (Paper III)

Being aware of the issue of index switching on Illumina HiSeq3000 and HiSeq4000 platforms, we turned to Illumina Nextseq platform and conducted single cell transcriptomic sequencing on circulating gluten-specific CD4+ T cells identified by HLA-DQ:gluten tetramer staining and used tetramer-negative CD4+ T cells as controls. The majority (81%) of the sorted tetramer+ cells had an intestinal origin (positive for β7-integrin) and were activated (positive for CD38), while most (75%) of the tetramer- cells were negative for both markers. Considering the tetramers carrying the four immunodominant epitopes only capture less than a half of all gluten-specific T cells, gluten-specific T cells of other specificities may also be expected in a small population of the controls. In order to avoid high proportion of the gluten-specific T cells of other specificities in the control cells, we purposely did not match β7-integrin and CD38 expression in cell selection. Circulating gluten-specific CD4+ T cells that are only found in CD patients, showed a distinctive phenotype in a mass cytometry analysis with 43 antibody markers (45). Single cell transcriptomic analysis is not limited by the number of markers, thus providing a complete view of the gluten-specific CD4+ T cells. Based on the observations in the quality control process, including the number of cells passed and the model of detection failures as a logistic function of mean log transformed expression of the house keeping genes, we focused on the data of best quality, where 342 blood CD4+ T

cells from one patient passed quality control. By applying a graph-based clustering on the transcriptomic data of the 342 cells, we observed a cluster of the tetramer-binding gluten-specific T cells with distinct gene expression profile, together with some discordant cells mostly found in the adjacent area consisting of tetramer-negative T cells (**Paper III, Figure 1A**). It confirmed our hypothesis that gluten-specific CD4+ T cells harbor distinct phenotype not only on the protein level, but also on the transcriptome level.

We performed differentially expression analysis on the tetramer+ T cells and tetramer- T cells and compared the results obtained from single-cell RNA-seq study and bulk RNA-seq study. On one hand, the cells in our single-cell study can be well classified into either tetramer-positive or tetramer-negative group by using differentially expressed genes from a bulk study as markers (**Paper III, Figure 1C**). The consistency assured us that both experiments were well performed. On the other hand, we observed some cases where of two genes with similar biological function, one was only identified as differentially expressed in bulk study, while the other was only identified in single-cell study. We therefore believe that both bulk and single-cell RNA-seq are important to capture complete gene expression patterns.
Based on the result of differentially expression analysis, we then set out to probe into molecular function and biological process associated with the differentially expressed genes. The significantly enriched pathways in GSEA mainly include the following categories: 1) response mediated processes, such as antigen receptor-mediated signaling, TCR signaling and activating signal transduction; 2) those associated with co-stimulatory signal during T cell activation, such as CTLA4, CD28 family; all of the above showed an increased T-cell activation in tetramer+ T cells. 3) increased signal of cytokines, such as interleukin-1 family, a group of cytokines regulating immune and inflammatory responses. On the other hand, the widely down-regulated ribosomal genes indicate reduced overall translation and protein synthesis activities in tetramer+ T cells.

In contrast to a variety of activation makers that were upregulated, we observed down-regulation of the early activation marker CD69 in the tetramer+ cells. The down-regulation of CD69 is known to be required for cell egress from intestinal tissue (122)(123). It is in accordance with the mass cytometry data in (45), where the expression of CD69 was markedly lower in blood than in gut. Together with the report (124) that CD69 in T cell is up-regulated half an hour after stimulation, then rapidly down-regulated, and the up-regulated gut-homing markers such as CCR9 and α4 integrin in the tetramer+ cells, indicates that the

tetramer+ cells in blood have recently egressed from the intestinal tissue.

Our result (**Paper III, Figure 3A**) was in accordance with the previous observation in (60)(59)(125) that tetramer+ T cells are activated effector memory T cells. One previous in vitro study (126) showed that most gluten-specific cells had a phenotype of regulatory T cell. However, the ex vivo analysis (45) suggested that the gluten-specific T cells do not express a classical regulatory T cell phenotype due to the lack of markers CD25 and GARP. By applying VISION pipeline with gene signatures from (127), we observed that the tetramer+ T cells in our study showing features of either Th1 or Tfh cells (**Paper III, Figure 3C**). We also found a few cells with transcriptional burst of IFN-γ and IL-21, key cytokines secreted by T cells and important for inflammatory responses in the celiac lesion and differentiation of plasma cells. As suggested by a recent study (128), plasma cells might be the most abundant cells presenting gluten peptides to T cells in inflamed intestinal tissues from CD patients. Finally, we combined the clonotype inferred from the reconstructed TCRs and the transcriptional profile of the cells. As expected, the vast majority of the cells with clonal expansion were tetramer positive cells (**Paper III, Figure 4A**). Similar to the findings in (94), cells that shared the same clonal origin were not more transcriptionally similar than any other randomly selected pairs of tetramer-positive cells. The result indicate that the clonal origin has limited impact on the phenotype of the cells, if any.

Conclusion and future perspectives

The current diagnostic strategy of CD in adults is based on screening by serum test, followed by confirmation with biopsy. Based on knowledge established by many studies on characterization of gluten-specific T cells, e.g. gluten-specific T cells are only found in CD patients but not in healthy controls (43)(45); certain identical T cell clonotypes are observed in multiple CD patients (129), we hypothesized that TCR-based methods could also be used as a potential diagnostic tool and performed the proof-of-principle study in **Paper I**. Although we successfully inferred CD state of majority of the donors in the study with limited cohort size, there are some apparent obstacles for using TCR sequencing based test as a diagnostic tool in practice. The major concern is the difficulty for identifying sufficient number of circulating gluten-specific T cells at a reasonable cost, given that the frequency of these cells is much lower in blood than that in gut. Machine learning methods such as Support Vector Machine, Random Forest, Gradient Boosting Machine and Conversional Neural Network have been showing great potential in identifying disease-specific immune receptor sequence (130) and distinguishing immune repertoire in different disease status (131), such as multiple sclerosis (132); tumor or normal tissues from patients with gastric cancer (130), colorectal and breast cancer (133). These machine learning methods were very likely applied on large scale of immune receptor repertoire data from CD patients. In that event, the machine learning process should be able to generate signatures for discriminating the repertoires, which may not necessary be limited to gluten-specific T cells. Importantly, aside of sequential data, it is easy to incorporate other information such as CDR-length and physicochemical properties of amino acid sequences, thereby improving the predictive accuracy. In addition, clinical application of TCR-based methods would require further research focusing on how they can be combined with the existing serological tests of CD, e.g. serum IgA and TG2-IgA, DGP-IgG tests, to achieve sufficient specificity and sensitivity so that invasive biopsy-based tests will no longer be needed.

Many studies over the past decades pointed to gluten-specific CD4 T cell as a key player in the pathology of CD. In the single cell transcriptome analysis, we demonstrated that the gluten-specific CD4+ T cells displaying signatures consistent with activated effector memory T cells were more heterogenous in contrast to the naive CD4 T cells. For instance, despite the pervasive upregulated signals for TCR signaling observed in gluten-specific T cells, (co)activation marker such as CD28 was expressed among some of them, while the rest of

them expressed CTLA-4, a molecule dampening the immune response by competing with CD28. Further experiment, e.g. time-series study could explore if there is any regulon causing the differences or they were just in different phase of immune response. In addition, the global transcriptional profiles of the tetramer-positive cells showed that these cell harbor features of either Th1 or Tfh cells, of which Tfh cells have ability to support B cell antibody production (134)(135). It has been shown that Tfh cells increases in the blood of patients with autoimmune diseases, including systemic lupus erythematosus (136) and rheumatoid arthritis (137). It is a challenge but of major interest to understand how the interaction between T helper cells and variety of immune cells were regulated in different conditions.

We observed that differentially expressed genes between the gluten-specific CD4 T cells and the non-specific ones associated with several biological processes, such as TCR signaling, activating signal transduction, as well as some metabolic processes including fatty acid metabolism and redox potentials. In order to provide a rational basis for therapeutic intervention, further reconstruction of gene regulatory network is required. A more robust regulatory network can be expected if the network can be inferred by combined use of supportive information collected from multiple levels, i.e. gene co-expression, sequences of motif and documented function.

One of the major hindrances to identifying potential marker for therapeutic perturbation is variation in the expression of markers under certain condition, which might be caused by many factors including body compartments, function time, individual, experimental workflows in different labs as well as random dropout events. A single experiment is unlikely to handle all these factors, therefore is not robust enough for the purpose. In future, continually refined methods for data integration would allow collecting variety of data modalities produced from different experimental workflows and data produced by different labs. The joint analysis of them would be promising for overcoming the limitation and getting more reliable markers. Moreover, it would be valuable to compare molecular signatures of immune cells from different diseases and tissues. Take CD for example, the unique profile of gluten-specific CD4 T cells identified by CyTOF experiments was shown to be also found in multiple other autoimmune diseases, suggesting that therapies targeting CD associated T cells could also be useful in other autoimmune diseases, especially for that which disease-driving antigen have not been identified (138).

References

1. Parham P. The Immune System. 4th ed. New York: Garland Science; 2014. 625 p.

2. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. 2012 Mar;44(3):291–6.

3. Okada Y, Suzuki A, Ikari K, Terao C, Kochi Y, Ohmura K, et al. Contribution of a Non-classical HLA Gene, HLA-DOA, to the Risk of Rheumatoid Arthritis. Am J Hum Genet. 2016 Aug;99(2):366–74.

4. Okada Y, Kim K, Han B, Pillai NE, Ong RT-H, Saw W-Y, et al. Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. Hum Mol Genet. 2014 Dec 20;23(25):6916–26.

5. Kumar V, Robbins SL, editors. Robbins basic pathology. 8th ed. Vol. Table 5-7. Philadelphia, PA: Saunders/Elsevier; 2007. 946 p.

6. Katze MG, Korth MJ, Law GL, Nathanson N, editors. Viral pathogenesis: from basics to systems biology. 3rd ed. Amsterdam; Boston: Elsevier/Academic Press; 2016. 351 p.

7. Zhu J, Paul WE. CD4 T cells: fates, functions, and faults. Blood. 2008 Sep 1;112(5):1557–69.

8. Raphael I, Nalawade S, Eagar TN, Forsthuber TG. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. Cytokine. 2015 Jul;74(1):5–17.

9. Vantourout P, Hayday A. Six-of-the-best: unique contributions of γδ T cells to immunology. Nat Rev Immunol. 2013 Feb;13(2):88–100.

10. Dupic T, Marcou Q, Walczak AM, Mora T. Genesis of the αβ T-cell receptor. Chain B, editor. PLOS Comput Biol. 2019 Mar 4;15(3):e1006874.

11. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009 Oct;19(10):1817–24.

12. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. Annu Rev Immunol. 2015 Mar 21;33(1):169–200.

13. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos Trans R Soc B Biol Sci. 2015 Aug 19;370(1675):20140291.

14. Alberts B, Johnson A, Lewis J. Molecular Biology of the Cell [Internet]. 4th edition. Vol. Lymphocytes and the Cellular Basis of Adaptive Immunity. New York: Garland Science; 2002. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26921/

15. Arstila TP. A Direct Estimate of the Human T Cell Receptor Diversity. Science. 1999 Oct 29;286(5441):958–61.

16. Restifo NP, Gattinoni L. Lineage relationship of effector and memory T cells. Curr Opin Immunol. 2013 Oct;25(5):556–63.

17. Di Sabatino A, Corazza GR. Coeliac disease. The Lancet. 2009 Apr;373(9673):1480–93.

18. Fasano A. Clinical presentation of celiac disease in the pediatric population. Gastroenterology. 2005 Apr;128(4):S68–73.

19. Jabri B, Sollid LM. T Cells in Celiac Disease. J Immunol. 2017 Apr 15;198(8):3005–14.

20. Husby S, Koletzko S, Korponay-Szabó I, Kurppa K, Mearin ML, Ribes-Koninckx C, et al. European Society Paediatric Gastroenterology, Hepatology and Nutrition Guidelines for Diagnosing Coeliac Disease 2020: J Pediatr Gastroenterol Nutr. 2019 Oct;1.

21. Tosi R, Vismara D, Tanigaki N, Ferrara GB, Cicimarra F, Buffolano W, et al. Evidence that celiac disease is primarily associated with a DC locus allelic specificity. Clin Immunol Immunopathol. 1983 Sep;28(3):395–404.

22. Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J Exp Med. 1989 Jan 1;169(1):345–50.

23. Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, et al. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum Immunol. 2003 Apr;64(4):469–77.

24. Pisapia L, Camarca A, Picascia S, Bassi V, Barba P, Del Pozzo G, et al. HLA-DQ2.5 genes associated with celiac disease risk are preferentially expressed with respect to non-predisposing HLA genes: Implication for anti-gluten T cell response. J Autoimmun. 2016 Jun;70:63–72.

25. Bergseng E, Sidney J, Sette A, Sollid LM. Analysis of the binding of gluten T-cell epitopes to various human leukocyte antigen class II molecules. Hum Immunol. 2008 Feb;69(2):94–100.

26. Arentz-Hansen H, Körner R, Molberg Ø, Quarsten H, Vader W, Kooy YMC, et al. The Intestinal T Cell Response to α-Gliadin in Adult Celiac Disease Is Focused on a Single Deamidated Glutamine Targeted by Tissue Transglutaminase. J Exp Med. 2000 Feb 21;191(4):603–12.

27. Qiao S-W, Bergseng E, Molberg Ø, Jung G, Fleckenstein B, Sollid LM. Refining the Rules of Gliadin T Cell Epitope Binding to the Disease-Associated DQ2 Molecule in Celiac Disease: Importance of Proline Spacing and Glutamine Deamidation. J Immunol. 2005 Jul 1;175(1):254–61.

28. Sollid LM, Qiao S-W, Anderson RP, Gianfrani C, Koning F. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. Immunogenetics. 2012 Jun;64(6):455–60.

29. Sollid LM, Tye-Din JA, Qiao S-W, Anderson RP, Gianfrani C, Koning F. Update 2020: nomenclature and listing of celiac disease-relevant gluten epitopes recognized by CD4+ T cells. Immunogenetics. 2020;72(1–2):85–8.

30. Tye-Din JA, Stewart JA, Dromey JA, Beissbarth T, van Heel DA, Tatham A, et al. Comprehensive, Quantitative Mapping of T Cell Epitopes in Gluten in Celiac Disease. Sci Transl Med. 2010 Jul 21;2(41):41ra51.

31. Shan L, Qiao S-W, Arentz-Hansen H, Molberg Ø, Gray GM, Sollid LM, et al. Identification and Analysis of Multivalent Proteolytically Resistant Peptides from Gluten: Implications for Celiac Sprue. J Proteome Res. 2005 Oct;4(5):1732–41.

32. Dørum S, Arntzen MØ, Qiao S-W, Holm A, Koehler CJ, Thiede B, et al. The Preferred Substrates for Transglutaminase 2 in a Complex Wheat Gluten Digest Are Peptide Fragments Harboring Celiac Disease T-Cell Epitopes. Buckle AM, editor. PLoS ONE. 2010 Nov 19;5(11):e14056.

33. Araya RE, Gomez Castro MF, Carasi P, McCarville JL, Jury J, Mowat AM, et al. Mechanisms of innate immune activation by gluten peptide p31-43 in mice. Am J Physiol-Gastrointest Liver Physiol. 2016 Jul 1;311(1):G40–G49.

34. Sollid LM, Molberg O, McAdam S, Lundin KE. Autoantibodies in coeliac disease: tissue transglutaminase--guilt by association? Gut. 1997 Dec;41(6):851–2.

35. Iversen R, Roy B, Stamnaes J, Høydahl LS, Hnida K, Neumann RS, et al. Efficient T cell-B cell collaboration guides autoantibody epitope bias and onset of celiac disease. Proc Natl Acad Sci U S A. 2019 23;116(30):15134–9.

36. Dieterich W, Ehnis T, Bauer M, Donner P, Volta U, Riecken EO, et al. Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat Med. 1997 Jul 1;3(7):797–801.

37. du Pré MF, Sollid LM. T-cell and B-cell immunity in celiac disease. Best Pract Res Clin Gastroenterol. 2015 Jun;29(3):413–23.

38. de Ritis G, Auricchio S, Jones HW, Lew EJ-L, Bernardin JE, Kasarda DD. In vitro (Organ culture) studies of the toxicity of specific A-gliadin peptides in celiac disease. Gastroenterology. 1988 Jan;94(1):41–9.

39. Maiuri L, Troncone R, Mayer M, Coletta S, Picarelli A, Vincenzi MD, et al. In vitro Activities of A-Gliadin-Related Synthetic Peptides Damaging Effect on the Atrophic Coeliac Mucosa and Activation of Mucosal Immune Response in the Treated Coeliac Mucosa. Scand J Gastroenterol. 1996 Jan;31(3):247–53.

40. Gómez Castro MF, Miculán E, Herrera MG, Ruera C, Perez F, Prieto ED, et al. p31-43 Gliadin Peptide Forms Oligomers and Induces NLRP3 Inflammasome/Caspase 1-Dependent Mucosal Damage in Small Intestine. Front Immunol. 2019 Jan 30;10:31.

41. Halstensen TS, Brandtzaeg P. Activated T lymphocytes in the celiac lesion: Non-proliferative activation (CD25) of CD4+ α/β cells in the lamina propria but proliferation (Ki-67) of α/β and γ/δ cells in the epithelium. Eur J Immunol. 1993 Feb;23(2):505–10.

42. Halstensen TS, Scott H, Fausa O, Brandtzaeg P. Gluten Stimulation of Coeliac Mucosa In Vitro Induces Activation (CD25) of Lamina Propria CD4H T cells and Macrophages but no Crypt-Cell Hyperplasia. Scand J Immunol. 1993 Dec;38(6):581–90.

43. Molberg O, Kett K, Scott H, Thorsby E, Sollid LM, Lundin KEA. Gliadin Specific, HLA DQ2-Restricted T Cells are Commonly Found in Small Intestinal Biopsies from Coeliac Disease Patients, but not from Controls. Scand J Immunol. 1997 Jul;46(1):103–8.

44. Lundin KEA, Scott H, Fausa O, Thorsby E, Sollid LM. T cells from the small intestinal Mucosa of a DR4, DQ7/DR4. DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8. Hum Immunol. 1994 Dec;41(4):285–91.

45. Christophersen A, Lund EG, Snir O, Solà E, Kanduri C, Dahal-Koirala S, et al. Distinct phenotype of CD4 + T cells driving celiac disease identified in multiple autoimmune conditions. Nat Med. 2019 May;25(5):734–7.

46. Nilsen EM, Lundin KE, Krajci P, Scott H, Sollid LM, Brandtzaeg P. Gluten specific, HLA-DQ restricted T cells from coeliac mucosa produce cytokines with Th1 or Th0 profile dominated by interferon gamma. Gut. 1995 Dec 1;37(6):766–76.

47. Nilsen EM, Jahnsen FL, Lundin KEA, Johansen F, Fausa O, Sollid LM, et al. Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. Gastroenterology. 1998 Sep;115(3):551–63.

48. Deem RL, Shanahan F, Targan SR. Triggered human mucosal T cells release tumour necrosis factor-alpha and interferon-gamma which kill human colonic epithelial cells. Clin Exp Immunol. 2008 Jun 28;83(1):79–84.

49. Goel G, Tye-Din JA, Qiao S-W, Russell AK, Mayassi T, Ciszewski C, et al. Cytokine release and gastrointestinal symptoms after gluten challenge in celiac disease. Sci Adv. 2019 Aug;5(8):eaaw7756.

50. Bodd M, Ráki M, Tollefsen S, Fallang LE, Bergseng E, Lundin KEA, et al. HLA-DQ2-restricted gluten-reactive T cells produce IL-21 but not IL-17 or IL-22. Mucosal Immunol. 2010 Nov;3(6):594–601.

51. Broughton SE, Petersen J, Theodossis A, Scally SW, Loh KL, Thompson A, et al. Biased T Cell Receptor Usage Directed against Human Leukocyte Antigen DQ8-Restricted Gliadin Peptides Is Associated with Celiac Disease. Immunity. 2012 Oct;37(4):611–21.

52. Petersen J, van Bergen J, Loh KL, Kooy-Winkelaar Y, Beringer DX, Thompson A, et al. Determinants of Gliadin-Specific T Cell Selection in Celiac Disease. J Immunol. 2015 Jun 15;194(12):6112–22.

53. Dahal-Koirala S, Ciacchi L, Petersen J, Risnes LF, Neumann RS, Christophersen A, et al. Discriminative T-cell receptor recognition of highly homologous HLA-DQ2–bound gluten epitopes. J Biol Chem. 2019 Jan 18;294(3):941–52.

54. Qiao S-W, Christophersen A, Lundin KEA, Sollid LM. Biased usage and preferred pairing of α- and β-chains of TCRs specific for an immunodominant gluten epitope in coeliac disease. Int Immunol. 2014 Jan 1;26(1):13–9.

55. Dahal-Koirala S, Risnes LF, Christophersen A, Sarna VK, Lundin KE, Sollid LM, et al. TCR sequencing of single cells reactive to DQ2.5-glia-α2 and DQ2.5-glia-ω2 reveals clonal expansion and epitope-specific V-gene usage. Mucosal Immunol. 2016 May;9(3):587–96.

56. Altman JD, Moss PAH, Goulder PJR, Barouch DH, McHeyzer-Williams MG, Bell JI, et al. Phenotypic Analysis of Antigen-Specific T Lymphocytes. Science. 1996 Oct 4;274(5284):94–6.

57. Bakker AH, Schumacher TN. MHC multimer technology: current status and future prospects. Curr Opin Immunol. 2005 Aug;17(4):428–33.

58. Hardy MY, Girardin A, Pizzey C, Cameron DJ, Watson KA, Picascia S, et al. Consistency in Polyclonal T-cell Responses to Gluten Between Children and Adults with Celiac Disease. Gastroenterology. 2015 Nov;149(6):1541-1552.e2.

59. Raki M, Fallang L-E, Brottveit M, Bergseng E, Quarsten H, Lundin KEA, et al. Tetramer visualization of gut-homing gluten-specific T cells in the peripheral blood of celiac disease patients. Proc Natl Acad Sci. 2007 Feb 20;104(8):2831–6.

60. Christophersen A, Ráki M, Bergseng E, Lundin KE, Jahnsen J, Sollid LM, et al. Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. United Eur Gastroenterol J. 2014 Aug;2(4):268–78.

61. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci. 1977 Feb 1;74(2):560–4.

62. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019 Jul 1;35(13):2193–8.

63. Taub F E, DeLEO JM, Thompson EB. Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs. DNA. 1983 Dec;2(4):309–27.

64. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Zhang S-D, editor. PLoS ONE. 2014 Jan 16;9(1):e78644.

65. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Nov;2015(11):951-69.

66. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013 Nov;10(11):1096–8.

67. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016 Dec;17(1):77.

68. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science. 2014 Feb 14;343(6172):776–9.

69. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014 Feb;11(2):163–6.

70. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Apr;8(1):14049.

71. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May;161(5):1202–14.

72. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

73. Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc. 2010 Mar;5(3):516–35.

74. Effects of Index Misassignment on Multiplexing and Downstream Analysis [Internet]. Illumina, Inc.; Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

75. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples in Illumina HiSeq 4000 DNA Sequencing. bioRxiv. 2017 Apr 9;

76. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012 Jan;40(1):e3.

77. Simon Andrews. FASTQC. A quality control tool for high throughput sequence data [Internet]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

78. Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;11(Suppl 4):S7.

79. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047–8.

80. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873–81.

81. Jean G, Kahles A, Sreedharan VT, Bona FD, Rätsch G. RNA-Seq Read Alignments with PALMapper. Curr Protoc Bioinforma [Internet]. 2010 Dec [cited 2020 Feb 18];32(1). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1106s32

82. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010 Oct 1;38(18):e178.

83. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol [Internet]. 2016 Dec [cited 2017 Oct 15];17(1). Available from: http://genomebiology.com/2016/17/1/13

84. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.

85. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923–30.

86. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003 Dec;31(4):265–73.

87. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27.

88. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020 Dec;21(1):12.

89. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun;177(7):1888-1902.e21.

90. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 May;36(5):411–20.

91. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018 May;36(5):421–7.

92. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012 Sep 1;40(17):e133.

93. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017 Dec;18(1):151.

94. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016 Apr;13(4):329–32.

95. Afik S, Yates KB, Bi K, Darko S, Godec J, Gerdemann U, et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. Nucleic Acids Res. 2017 Sep 19;45(16):e148.

96. Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao S-W, et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat Methods. 2018 Aug 1;15(8):563–5.

97. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinformatics. 2017 Feb 1;33(1)425-7.

98. Bodd M, Ráki M, Bergseng E, Jahnsen J, Lundin KEA, Sollid LM. Direct cloning and tetramer staining to measure the frequency of intestinal gluten-reactive T cells in celiac disease. Eur J Immunol. 2013;43(10):2605–12.

99. Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, et al. Naïve CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. Immunity. 2007 Aug;27(2):203–13.

100. Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, et al. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. Cell Syst. 2019 Apr;8(4):315-328.e8.

101. Stigler SM. Francis Galton's Account of the Invention of Correlation. Stat Sci. 1989 May;4(2):73–9.

102. Kendall MG. A NEW MEASURE OF RANK CORRELATION. Biometrika. 1938 Jun 1;30(1–2):81–93.

103. Myers JL, Well A. Research design and statistical analysis. 2nd ed. Mahwah, N.J: Lawrence Erlbaum Associates; 2003. 508 p.

104. Glass GV, Hopkins KD. Statistical methods in education and psychology. 3rd ed. Boston: Allyn and Bacon; 1996. 674 p.

105. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015 May;161(5):1187–201.

106. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017 Apr;14(4):395–8.

107. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017 Feb;65(4):631-643.e4.

108. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017 Mar 6;14(4):381–7.

109. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417–9.

110. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525–7.

111. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014 May;32(5):462–4.

112. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016 Dec;17(1):63.

113. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Dec;17(1):29.

114. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet. 2017 May;49(5):659–65.

115. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-throughput pairing of T cell receptor α and β sequences. Sci Transl Med. 2015 Aug 19;7(301):301ra131.

116. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol. 2014 Jul;32(7):684–92.

117. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol. 2017 Dec;17(1):61.

118. Dash P, McClaren JL, Oguin TH, Rothwell W, Todd B, Morris MY, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. J Clin Invest. 2011 Jan 4;121(1):288–95.

119. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, et al. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. J Immunol. 2014 Mar 15;192(6):2689–98.

120. van der Valk T, Vezzi F, Ormestad M, Dalen L, Guschanski K. Low rate of index hopping on the Illumina HiSeq X platform. bioRxiv. 2017 Aug 22;179028.

121. Larsson AJM, Stanley G, Sinha R, Weissman IL, Sandberg R. Computational correction of index switching in multiplexed sequencing libraries. Nat Methods. 2018 May;15(5):305–7.

122. Shiow LR, Rosen DB, Brdičková N, Xu Y, An J, Lanier LL, et al. CD69 acts downstream of interferon-α/β to inhibit S1P1 and lymphocyte egress from lymphoid organs. Nature. 2006 Mar;440(7083):540–4.

123. Cibrián D, Sánchez-Madrid F. CD69: from activation marker to metabolic gatekeeper. Eur J Immunol. 2017;47(6):946–53.

124. Rosette C, Werlen G, Daniels MA, Holman PO, Alam SM, Travers PJ, et al. The Impact of Duration versus Extent of TCR Occupancy on T Cell Activation. Immunity. 2001 Jul;15(1):59–70.

125. du Pré FM, van Berkel LA, Ráki M, van Leeuwen MA, de Ruiter LF, Broere F, et al. CD62LnegCD38+ Expression on Circulating CD4+ T Cells Identifies Mucosally Differentiated Cells in Protein Fed Mice and in Human Celiac Disease Patients and Controls: Am J Gastroenterol. 2011 Jun;106(6):1147–59.

126. Cook L, Munier CML, Seddiki N, van Bockel D, Ontiveros N, Hardy MY, et al. Circulating gluten-specific FOXP3 + CD39 + regulatory T cells have impaired suppressive function in patients with celiac disease. J Allergy Clin Immunol. 2017 Dec;140(6):1592-1603.e8.

127. Crawford A, Angelosanto JM, Kao C, Doering TA, Odorizzi PM, Barnett BE, et al. Molecular and Transcriptional Basis of CD4+ T Cell Dysfunction during Chronic Infection. Immunity. 2014 Feb;40(2):289–302.

128. Høydahl LS, Richter L, Frick R, Snir O, Gunnarsen KS, Landsverk OJB, et al. Plasma Cells Are the Most Abundant Gluten Peptide MHC-expressing Cells in Inflamed Intestinal Tissues From Patients With Celiac Disease. Gastroenterology. 2019 Apr;156(5):1428-1439.e10.

129. Risnes LF, Christophersen A, Dahal-Koirala S, Neumann RS, Sandve GK, Sarna VK, et al. Disease-driving CD4+ T cell clonotypes persist for decades in celiac disease. J Clin Invest. 2018;128(6):2642–50.

130. Konishi H, Komura D, Katoh H, Atsumi S, Koda H, Yamamoto A, et al. Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning. BMC Bioinformatics. 2019 Dec;20(1):267.

131. Brown AJ, Snapkov I, Akbar R, Pavlović M, Miho E, Sandve GK, et al. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. Mol Syst Des Eng. 2019;4(4):701–36.

132. Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. BMC Bioinformatics. 2017 Dec;18(1):401.

133. Ostmeyer J, Christley S, Toby IT, Cowell LG. Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue. Cancer Res. 2019 Apr 1;79(7):1671–80.

134. Kim CH, Rott LS, Clark-Lewis I, Campbell DJ, Wu L, Butcher EC. Subspecialization of Cxcr5+ T Cells. J Exp Med. 2001 Jun 18;193(12):1373–82.

135. Campbell DJ, Kim CH, Butcher EC. Separable effector T cell populations specialized for B cell help or tissue inflammation. Nat Immunol. 2001 Sep;2(9):876–81.

136. Kim SJ, Lee K, Diamond B. Follicular Helper T Cells in Systemic Lupus Erythematosus. Front Immunol. 2018 Aug 3;9:1793.

137. Gensous N, Charrier M, Duluc D, Contin-Bordes C, Truchetet M-E, Lazaro E, et al. T Follicular Helper Cells in Autoimmune Disorders. Front Immunol. 2018 Jul 17;9:1637.

138. Christophersen A, Risnes LF, Dahal-Koirala S, Sollid LM. Therapeutic and Diagnostic Implications of T Cell Scarring in Celiac Disease and Beyond. Trends Mol Med. 2019 Oct;25(10):836–52.

I

T Cell Receptor Repertoire as a Potential Diagnostic Marker for Celiac Disease

Ying Yao[1,2], Asima Zia[4], Ralf Stefan Neumann[1,2], Milena Pavlovic[3], Gabriel Balaban[3], Geir Kjetil Sandve[2,3], Shuo Wang Qiao[1,2*]

[1] Department of Immunology, Institute of Clinical Medicine, University of Oslo, Norway
[2] K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, Norway.
[3] Department of Informatics, University of Oslo, Norway.
[4] Living Systems Laboratory, King Abdullah University of Science and Technology, Saudi Arabia.

*** Correspondence:**
Shuo Wang Qiao
s.w.qiao@medisin.uio.no

Keywords: Celiac Disease$_1$, CD4+ T cells$_2$, TCR repertoire$_3$, Disease Inference$_4$; High-throughput sequencing$_5$

Abbreviations

CD	celiac disease
TCR	T cell receptor
TCRα	T cell receptor α chain
TCRβ	T cell receptor β chain
MHC	major histocompatibility complex
pMHC	peptide:MHC complex
HLA	human leukocyte antigen
UMI	unique molecular identifier
CMV	cytomegalovirus
ROC	Receiver Operating Characteristics
AUC	area under the ROC curve

Abstract

An individual's T cell repertoire is skewed towards some specificities as a result of past antigen exposure. Identifying T cell receptor signatures associated with a disease is challenging due to the overall complexity of antigens. In celiac disease, the antigen epitopes are well characterised and the specific T-cell repertoire associated with the disease has been explored in depth. By investigating T cell repertoires of unsorted lamina propria T cells from 15 individuals, we provide the first proof-of-concept study showing that it could be possible to infer disease state by matching against a priori known disease-associated TCR sequences.

1. Introduction

T cell plays a central role in cell mediated immune response. The T cell receptor (TCR) is responsible for recognizing antigenic peptides bound to major histocompatibility complex (MHC) molecules. In 95% of human T cells, the TCR consists of the α chain (TCRα) and β chain (TCRβ). During T-cell development, each thymocyte generates its unique TCR variant through recombination of different V, D, J gene segments and random deletion and/or insertion of nucleotide at the junctions. This results in a highly diverse TCR repertoire and the diversity is important for maximizing potential coverage of the protective immunity. During thymic selection, only a small fraction of thymocytes that bind self peptide:MHC complex (pMHC) with intermediate affinity differentiate into mature T cells, therefore MHC polymorphism further modulates an individual's TCR repertoire by determining the collection of peptides that can be presented to T cells during development [1] [2]. Structural studies of TCRs binding to pMHC ligands [3] have shown that although there are some exceptions, as a rule, the variable regions of the TCRα chain largely contact the MHC molecule whereas the TCRβ chain makes most contact with the antigenic peptide. This notion is supported by genetic studies where the usage of V-gene segment of the TCRα is more closely associated with the human leukocyte antigen (HLA) profiles of the person [4].

In cell mediated immune response, naïve T cells are activated and clonally expand after recognition of foreign antigenic peptides presented by MHC. Thus, although the diversity is highest in the naive compartment, in the memory compartment it is skewed towards some specificities as a result of past antigen exposure and thereby antigen-driven selection and expansion. Since T cells directed against certain antigen in a disease setting are clonally expanded, a biased repertoire should be observed given enough sequencing power. With advancement of high-throughput immune receptor sequencing methods, TCR repertoire has the potential to be a diagnostic marker for infections or autoimmune disease. However, due to the complexity and diversity of TCR repertoires of different individuals, identifying the TCR signatures associated with an antigen is challenging. Somma et al.[5] identified a number of TCRβ clonotypes implicated in the pathogenesis of multiple sclerosis, which were clonally expanded in both the healthy and the affected twin. In contrast, studies on monozygotic twins [6] have demonstrated that different disease settings altered the TCR gene usage and TCR repertoire as a whole. Despite TCR complexities, TCR clonotyping has been used as diagnostic tool in Emerson et al.[7] where the exposure to cytomegalovirus (CMV) of 666 subjects could be inferred by their TCR repertoires. The TCR repertoire data was generated from peripheral blood, since CMV is a disease that elicit a particularly strong immune

response where an unusually large proportion of the T-cell response is CMV-related. However, the T-cell response is less pronounced for most other diseases. In the CD4 compartment, only 1 to 100 per million CD4 cells in blood would be expected to be specific to a given pMHC [8 9] whereas in affected tissue the frequency of antigen-specific TCR would be expected to be around 1 to 5 per hundred CD4 T cells, at least in celiac disease [10]. Therefore, it is advantageous to look at the T-cell response in the affected tissue where the frequency of disease-relevant T cells is much higher than in blood.

Celiac disease (CD) is a long-term HLA-associated autoimmune disorder that primarily affects the small intestine. Its pathogenesis is relatively slow compared to acute infections. The primary association is with MHC class II alleles encoding HLA-DQ2.5 (*HLA-DQA1*05*/*HLA-DQB1*02*, expressed by 90% of patients), HLA-DQ8 (*HLA-DQA1*03*/*HLA-DQB1*03:02*), and HLA-DQ2.2 (*HLA-DQA1*02:01*/*HLA-DQB1*02*) [11 12 13]. The MHC class II association shows that CD4 T cell plays an important role in pathogenesis of celiac disease. The epitopes of the causative antigen gluten are well defined and gluten-specific T cells that are only found in the small intestine of celiac disease patients, but not in healthy controls, have been isolated and extensively studied. All gluten-reactive T cells in the lesions are restricted by the disease-associated HLA-DQ2.5 molecule in HLA-DQ2.5-positive subjects [14]. HLA-DQ-gluten tetramers carrying the immunodominant gluten epitopes have been used to visualize gluten-specific T cells directly from blood or small intestinal tissue [15]. Studies have shown that public features, i.e. identical TCRα, TCRβ, or paired TCRαβ amino acid sequences found in different individuals, are frequently observed among gluten-specific T cells [15].

To explore whether disease state could be assessed from a limited number of tissue-derived cells, we started with around 10,000 T cells taken from the lamina propria of two duodenal biopsies per individual to assess the celiac disease state. Of these T cells, more than 80% are CD4. In order to find the best diagnostic biomarkers, we evaluated the usage of different types of prior information, i.e. all gluten-specific TCRs versus a smaller subset of public gluten-specific TCRs shared across multiple CD patients. The aim of this study is proof of principle to show the potential of using TCR-based diagnostics. This is the first step towards biopsy-free diagnostics of CD where TCR information would be collected directly from blood.

2. Materials and methods

2.1 Sample collection

The project was approved by the Regional Committee for Medical and Health Research Ethics South-East Norway (REK 2010/2720) and signed informed consent forms were obtained from all subjects. Intestinal biopsies from seven HLA-DQ2.5+ untreated celiac disease patients and eight non-celiac controls or HLA-DQ2-negative patients were collected in accordance with medical guidelines. Since TCR recognize peptide-HLA complexes, for the purpose of this study where we look for signature sequences of gluten:HLA-DQ2-reactive TCRs, we do not expect to find these TCR sequences in HLA-DQ2-negative patients whose gluten-reactive T cells are HLA-DQ8-restricted. Two pieces of duodenal biopsies were collected in ice-cold RPMI-1640. The epithelial layer that largely contain CD8+ intra-epithelial T cells was removed with two 5-minute incubation with PBS+2%FCS+2mM EDTA at 37C. After thorough washes with PBS to remove detached epithelial cells, the remaining lamina propria tissue was digested for 45 minutes with 1 mg/ml Collagenase (Sigma) and 0.1 mg/ml DNAse (Sigma). The resulting lamina propria single-cell suspension was counted and seeded directly in TCL buffer in four concentrations (108,000; 36,000; 18,000 and 9,000 cells per well) and eight biological replicates for each concentration. After thorough mixing to aid cell lysis, the lysates were kept frozen at -70C until processed. After defrosting, the cell lysate in TCL was transferred to 96-well plates precoated with dT-oligos from the TurboCapture 96 mRNA kit (Qiagen). mRNA extraction and cDNA synthesis using the plate-immobolized oligo-dT was carried out in accordance with the manufacturer's instructions with the modification of additional template switch oligo (Bio-d(AAGCAGTGGTATCAACGCAGAGTAGTNNNNNN)-r(GGG), where N denotes random nucleotides that serve as UMI). Following cDNA synthesis, two semi-nested TCRα- and TCRβ-specific PCR reactions were carried out as in [15].

2.2 TCR sequencing and data processing

Double indexing was applied in library preparation, thereby every pair of reads had an index composed of two barcodes on the forward and reverse read respectively, encoding the sample origin. Libraries were sequenced on the Illumina MiSeq platform with 250 nt pair-end sequencing at the Norwegian Sequencing Center (Oslo University Hospital).

All paired end reads were de-multiplexed based on the combination of their R1 and R2 barcode sequences. Both of the paired R1 and R2 reads were dropped if any of them had any nucleotide mismatch with the reference barcodes. On the paired reads assigned to each sample, we performed UMI tag extraction and UMI-guided assembly using the MIGEC pipeline [16], where all reads in a sample were grouped by their UMI and then each group with larger than 10 reads were assembled to generate a consensus sequence by multiple alignment. Both consensuses need to be successfully assembled for paired reads, otherwise the pair was dropped. Considering the relatively short UMI length and large expected number of cells in some wells in the study, the probability for a pair of similar UMI caused by sequencing error was relatively low. We have therefore not corrected sequencing errors in the UMI. The consensus sequences of samples from the same patient were then pooled and aligned with mismatches, inserts and deletions to the TCR database following the MiXCR pipeline [17], thereby TCRαβ chain and CDR3 repertoires were extracted from the assembled consensus sequences. Identical sequences were grouped in clonotypes, and the corresponding clonecounts were recorded. Consensus with poor quality were also collected and mapped to the grouped clonotypes for correction of PCR and sequencing errors. The default parameters of MiXCR were applied throughout this process.

2.3 Reference database of gluten-specific TCR sequences

To search for disease-associated TCR sequences that were present in our data, we used a reference database comprised of TCRα- and TCRβ-clonotypes obtained from single-cell TCR sequencing of HLA-DQ2.5-gluten-tetramer-sorted cells from 59 celiac disease patients. Sequences belonging to donors in the present study were excluded. Overall, there were 2,929 TCRα- and 2,662 TCRβ-clonotypes originating from 6,808 tetramer-sorted cells. A clonotype is defined throughout the study as a unique amino acid sequence of the re-arranged variable regions of the TCRα (VJ) or TCRβ (VDJ). Within this large reference dataset that includes almost all known gluten-specific TCR clonotypes to date, there is a smaller subset that consists of public clonotypes, defined as identical amino acid sequences observed in at least two CD patients. This public TCR subset contains 151 TCRα and 226 TCRβ clonotypes that have been collapsed from 1,150 TCRα sequences and 1,436 TCRβ sequences from 2,003 gluten-specific T cells. The collaps of TCR sequences to clonotypes was caused by both in vivo clonal expansion (multiple cells expressing identical TCRαβ sequences in the same patient) and convergent recombination (different nucleotide sequences encoding identical amino acid sequence).

2.4 Inferring disease state

We used logistic regression to classify the subjects based on the presence of the aforementioned antigen-specific TCR sequences. Logistic regression was performed by sklearn.linear model.LogisticRegression function from scikit-learn v0.20.4 Python module [18], where either normalized unique match or normalized clonecount match was used a single predictor. All the other parameters were set as default except for the C (inverse of regularization strength) that was set at 1E+5 to eliminate the effect of penalty term since no simplified model was preferable with a single predictor. We also employed the R package ROCR1.0-7 to calculate the sensitivity and specificity while using the same single predictor ranged from 0 to the maximum in different experimental settings, as well as the corresponding AUCs. Test for association between the prevalence of a clonotype and its frequency in our data using Kendall's tau was done with R package stats 3.4.4.

3. Result

3.1 Data acquisition

We sampled one million cells from the lamina propria of two duodenal biopsies from each of 15 individuals and sequenced the rearranged TCRα and TCRβ variable region in all samples. Flow cytometric analysis showed that approximately 1% of the sampled cells were T cell, of which >80% were CD4+ T cells. Thus, we have sampled and sequenced approximately 10,000 T cells from each subject. The number of sequencing reads generated from each individual varied from 0.1 million to 2.7 million, with an average of 1.7 million, representing on average 5,821 molecules after deduplexing. The number of unique clonotypes we observed ranged from 861 to 8,778. Basic information of the donors and the libraries were summarised in Table 1.

3.2 Gut-derived TCR clonotypes in our data matched preferentially public gluten-specific TCRs

By collapsing TCRs with the same V gene, J gene and CDR3 amino acid sequences from repertoires from all donors, we had in total 17,261 unique TCRα and 26,820 unique TCRβ clonotypes in our dataset (Figure 1). To find an optimal set of disease-associated TCR clonotypes for inferring disease state, we employed an external database consisting of data from multiple single-cell TCR sequencing projects where HLA-DQ2.5:gluten tetramer was

used to stain T cells from CD patients in vitro, followed by sorting and sequencing of the sorted gluten-specific TCRs. Among the total 5,591 gluten-specific TCRα and TCRβ amino acid sequences in the database, 377 of them were observed in at least two celiac disease patients and were thus defined as public clonotypes. When we compared our dataset from the gut tissue with the reference database, we found that 93 of the TCRβ clonotypes in our data matched gluten-specific TCR sequences in the reference database, of which 39 matched the subset of public TCRβ sequences. While 58 of the TCRα clonotypes matched gluten-specific TCRα sequences, only 15 out of these matched the public TCRα sequences (Figure 1). Since the public clonotypes account for 5% and 8% of the total TCRα and TCRβ reference dataset, respectively, it is interesting to note that among the matches we found in the gut-derived TCR data, 26% and 42% of them were matched to the public TCRα and TCRβ sequences, respectively.

From published studies of gluten-specific TCR sequences, it is known that some TCR clonotypes such as the TRBV7-2/TRBJ2-3 clonotype with CDR3 amino acid sequence ASSxRxTDTQY are found in virtually all CD subjects [19] [15]. On the other hand, many public CD clonotypes were found in only two subjects of total 59 subjects from whom the reference database was made. We hypothesized that highly public clonotypes found in many individuals in the reference database were more likely to be observed in our test data derived from unsorted T cells from the gut. For all the matched TCR clonotypes found in repertoires of all CD patients, we calculated the Kendall's rank correlation to test for the association between how often the TCR clonotype was observed in our data and the number of patients in the reference dataset who expressed this clonotype (Supplementary Table 1). Result of the test showed that the matching frequency of a TCR was positively associated with the number of patients across which the TCRs was shared, with tau of 0.338 and P value of 0.0026. Therefore, we found that the most public clonotypes found in many CD patients were also more frequently observed in our gut repertoire dataset.

3.3 Matching TCRβ alone was sufficient for predicting celiac disease state

For each TCR-repertoire of the 15 individuals, we summed up the number of unique TCR clonotypes that matched the disease-associated reference TCR sequences, this is referred to as sum unique disease associated TCRs. Since clonal expansion is a feature associated with the gluten-specific T cells in earlier studies, we took into account the clone size of each matched sequence, measured by the clonecount. Therefore, we also calculated the sum of the

clonecount of the same matched TCR sequences. For each subject, we normalized the unique clonotype match by dividing it with the total number of unique clonotypes found in that individual (unique match). Similarly, we calculated for each individual the clonecount match where the total clonecounts of all matched sequences were divided by the total number of clonecounts in the repertoire (Figure 2, Supplementary Table 2).

The normalized unique match and clonecount match were then used as a predictor in a logistic regression model to infer the status of celiac disease. We performed the analysis by using all TCR sequences or by using only TCRβ repertoires. In every possible combination, a balanced predictive accuracy was evaluated based on leave one out cross-validation. In addition, either the normalized unique match or clonecount match was used as a nonparametric classifier which was evaluated by AUC of ROC plots to enhance the robustness of the result. Both the balanced accuracy for logistic regression and the AUC value for nonparametric classifier are metrics that evaluate the predictive performance through balancing sensitivity and specificity.

We did not observe any clear and consistent differences in the predictive performance when matching against all gluten-specific TCR sequences was compared with matching against the public TCR sequences (Figure 3). Also, the predictive performance was similar whether information of clonal expansion was used or not. We did in all experimental settings observe higher predictive performance when using only TCRβ repertoire data (AUC between 0.94 and 0.98; 12-13 correct predictions) compared to using the sum of TCRα and TCRβ matches (AUC between 0.66 and 0.88; 6-10 correct predictions).

Considering the fact that the individuals in the diseased group were older than those in the control group as a whole (see Supplementary Fig.S1 on line), we checked if age could be a confounding factor in the predictions. In each scenario, we calculated the Pearson correlation coefficient of age and the predictor, either the normalized unique match or the normalized clonecount match. In seven out of eight scenarios, we got a low correlation coefficient, ranged from -0.13 to 0.15. There is only one exception showing slightly stronger association between age and the predictor ($\rho = -0.34$), where the public sequences were used as the reference and the normalized clonecount matches of both TCRα and TCRβ was used as the predictor. (see Supplementary Fig.S2 on line). It is another support for using only TCRβ for prediction, since it was less likely to be confounded by age if any effect exists.

4. Discussion

With the rapid advances in sequencing technology and in particular the ability to sequence a large number of TCRs, the TCR signatures associated with the recognition of a particular antigen, and in its extension a particular disease, can conceivably be used to infer the disease state. In this paper, we have used celiac disease in which extensive information about the disease-specific TCRs exist, to do a proof-of-principle study showing that disease state can be inferred based on TCR sequences derived from the diseased tissue. Using a small set of a few thousand clonotypes sequenced from around 10,000 T cells sampled from each of the 15 donors, the CD status was correctly predicted for 13 out of the 15 donors by matching against known gluten-specific TCRs.

TCRα clonotypes performed rather poorly in our study both when used on its own and in combination with TCRβ. Two controls had relatively large expanded TCRα clones that matched with public CD associated TCRα clonotypes. We suspect that although these TCRα clonotypes are disease-specific when paired with certain TCRβ chains, the expanded TCRα clones we have observed in our two controls most likely are paired with different TCRβ chains that conferred the complete TCR some other celiac-unrelated specificities. This could be better tested by further study profiling paired TCRs at the single cell level.

We used two types of prior information; the list of gluten-specific TCRs that contains all the 5,591 clonotypes of gluten tetramer-sorted T cells, and its small subset of 377 public clonotypes observed in more than one CD patient in the same database. These two alternative choices for disease associated TCRs present a trade-off between the quantity and specificity for finding matches for disease associated TCRs, since the database of public TCR sequences is more reliable. For the TCRβ repertoires, despite that the reference database of all gluten-specific TCRβ was 14 times larger than the subset of public TCRβ, the number of clonotypes matched to non-public reference sequences was about the same as number of clonotypes matched to public TCRβ. The predictions were not improved by including those non-public gluten specific TCRβs. It suggested that the non-public gluten specific TCRβ might not be as powerful as the public ones for diagnosing celiac disease. In addition, the considerable smaller size of the public database would save computational power.

The public TCRβ sequence dataset can be further improved in two aspects. On one hand, although the majority of TCRs in this study that matched the public reference TCR sequences were from untreated celiac disease patients, a few of them were also observed occasionally in the controls, which is similar to findings in [20]. We therefore believe that specificity of the public clonotypes could be further improved by involving more TCR repertoires from controls for training and purifying. On the other hand, public TCR sequences among CD patients can be continually accumulated by including a larger cohort over time. In a previous published study[20] where 39 public TCRβ was used, only five of the 39 public TCRβ sequences were detected in 10 active CD patients. Comparing with the extremely low number of detected public sequences, 70 public TCRβ sequences were detected in 8 CD patients in our study where 226 public TCRβ was used. The public TCR sequences among CD patients might approach to saturation. The positive association between the prevalence and frequency for the public TCR clonotypes indicates that the both the sequencing depth and number of individuals included can be optimized.

In this study, the state of CD was successfully inferred for the majority of the donors. For CD, the antigen specific CD4+ T cells are restricted by the disease-associated HLA molecules, which facilitates the prediction task focusing on distinguishing the HLA-DQ2+ untreated CD patients from the others as control. As the number of antigen specific CD4 T cells varies in different tissue for different infectious diseases, the repertoire size should be carefully validated when using T cell repertoire information as diagnostic tool.

The number of subjects included in this study is rather small, and only a few thousand clonotypes were sequenced from unsorted lamina propria from each subject. Our results indicate that even with these limitations, it might be possible to infer disease state by matching against known disease-associated TCR sequences. It is to our knowledge the first time this was shown for CD. Age was relatively uncorrelated with our most predictive features, the clonecount match ($\rho = -0.026$) and unique count match ($\rho = 0.14$), to the TRB & public TCR set, so that we do not believe that confounding by age played a role in our predictions based on these features. However, previous studies[21] have shown systematic differences in the human TCR repertoire with ageing. As our sample size was limited, we cannot rule out the possibility of age-related confounding affecting the predictive value of our TRB & public TCR set for diagnosing CD in the general population. It needs to be tested in further studies with larger sample sizes whether some other factors aside from HLA type,

such as age and severity of tissue damage, affect the frequency of gluten specific T cells in CD patients.

Ultimately, in CD, we would like to infer the disease state from blood samples such that diagnosis can be given without the need of endoscopic biopsy. In addition, with the advances in the knowledge of specific TCRs and TCR sequencing, it is conceivable that TCR repertoire could be used for the diagnosis of other chronic immune-mediated inflammatory diseases.

References

1. Dyall, R, Messaoudi, I, Janetzki, S & Nikolić-Žugić, J. MHC Polymorphism Can Enrich the T Cell Repertoire of the Species by Shifts in Intrathymic Selection. *The Journal of Immunology* **164,** 1695–1698 (2000).

2. Bevan, MJ. In thymic selection, peptide diversity gives and takes away. *Immunity* **7,** 175–178 (1997).

3. Wucherpfennig, KW, Gagnon, E, Call, MJ, Huseby, ES & Call, ME. Structural biology of the T-cell receptor: insights into receptor assembly, ligand recognition, and initiation of signaling. *Cold Spring Harb Perspect Biol* **2,** a005140 (2010).

4. Sharon, E, Sibener, LV, Battle, A, Fraser, HB, Garcia, KC & Pritchard, JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet* **48,** 995–1002 (2016).

5. Somma, P, Ristori, G, Battistini, L, Cannoni, S, Borsellino, G, Diamantini, A, *et al.* Characterization of CD8+ T cell repertoire in identical twins discordant and concordant for multiple sclerosis. *J. Leukoc. Biol.* **81,** 696–710 (2007).

6. Fozza, C, Contini, S, Corda, G, Virdis, P, Galleu, A, Bonfigli, S, *et al.* T-cell receptor repertoire analysis in monozygotic twins concordant and discordant for type 1 diabetes. *Immunobiology* **217,** 920–925 (2012).

7. Emerson, RO, DeWitt, WS, Vignali, M, Gravley, J, Hu, JK, Osborne, EJ, *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* **49,** 659–665 (2017).

8. Moon, JJ, Chu, HH, Pepper, M, McSorley, SJ, Jameson, SC, Kedl, RM, *et al.* Naïve CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* **27,** 203–213 (2007).

9. Christophersen, A, Ráki, M, Bergseng, E, Lundin, KE, Jahnsen, J, Sollid, LM, *et al.* Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. *United European Gastroenterol J* **2,** 268–278 (2014).

10. Bodd, M, Ráki, M, Bergseng, E, Jahnsen, J, Lundin, KEA & Sollid, LM. Direct cloning and tetramer staining to measure the frequency of intestinal gluten-reactive T cells in celiac disease. *European Journal of Immunology* **43,** 2605–2612 (2013).

11. Tosi, R, Vismara, D, Tanigaki, N, Ferrara, GB, Cicimarra, F, Buffolano, W, *et al.* Evidence that celiac disease is primarily associated with a DC locus allelic specificity. *Clin. Immunol. Immunopathol.* **28,** 395–404 (1983).

12. Sollid, LM, Markussen, G, Ek, J, Gjerde, H, Vartdal, F & Thorsby, E. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J. Exp. Med.* **169,** 345–350 (1989).

13. Karell, K, Louka, AS, Moodie, SJ, Ascher, H, Clot, F, Greco, L, *et al.* HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. *Hum. Immunol.* **64,** 469–477 (2003).

14. Jabri, B & Sollid, LM. T Cells in Celiac Disease. *The Journal of Immunology* **198,** 3005–3014 (2017).

15. Risnes, LF, Christophersen, A, Dahal-Koirala, S, Neumann, RS, Sandve, GK, Sarna, VK, *et al.* Disease-driving CD4+ T cell clonotypes persist for decades in celiac disease. *Journal of Clinical Investigation* **128,** 2642–2650 (2018).

16. Shugay, M, Britanova, OV, Merzlyak, EM, Turchaninova, MA, Mamedov, IZ, Tuganbaev, TR, *et al.* Towards error-free profiling of immune repertoires. *Nature Methods* **11,** 653–655 (2014).

17. Bolotin, DA, Poslavsky, S, Mitrophanov, I, Shugay, M, Mamedov, IZ, Putintseva, EV, *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods* **12,** 380–381 (2015).

18. Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825−2830 (2011).

19. Qiao, S-W, Ráki, M, Gunnarsen, KS, Løset, G-Å, Lundin, KEA, Sandlie, I, *et al.* Posttranslational Modification of Gluten Shapes TCR Usage in Celiac Disease. *The Journal of Immunology* **187,** 3064–3071 (2011).

20. Ritter, J, Zimmermann, K, Jöhrens, K, Mende, S, Seegebarth, A, Siegmund, B, *et al.* T-cell repertoires in refractory coeliac disease. *Gut* gutjnl-2016-311816 (2017). doi:10.1136/gutjnl-2016-311816

21. Britanova, OV, Putintseva, EV, Shugay, M, Merzlyak, EM, Turchaninova, MA, Staroverov, DB, *et al.* Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *J.I.* **192,** 2689–2698 (2014).

Acknowledgments
The authors would like to thank Victor Greiff for helpful comments and suggestions that improved the data processing.

Funding
This work is funded by Research Council of Norway (project 179573/V40 through the Centre of Excellence funding scheme and project 233885), and grants from the Stiftelsen Kristian Gerhard Jebsen (SKGJ-MED-017).

Author Contributions
AZ and SQ contributed to the study design and performed the experiments, YY, RN and MP analysed and interpreted data, GS contributed to supervision and data analysis, GS and GB contributed to critical review of the manuscript, YY and SQ interpreted data and wrote the manuscript.

Competing Interests
The authors declare that they have no competing interests.

Figure Legends
Figure 1. Number of unique TCRα (A) and TCRβ clonotypes (B) that matched either public or non-public disease associated TCR sequences.

Figure 2. Normalized match to each reference dataset for donors grouped by disease status. The donors on the left side of each frame were controls, while donors on the right were untreated celiac disease patients (UCD). Colors indicate if a donor was correctly predicted by Logistic regression models trained on the others. (A) Using normalized unique match and the public sequences as reference (B) Using normalized clonecount match and the public sequences as reference (C) Using normalized unique match and all gluten specific sequences as reference (D) Using normalized clonecount match and all gluten specific sequences as reference.

Figure 3. Predictive performance in all experimental settings. (A) using normalized unique match as a single predictor in logistic regression, balanced accuracy was evaluated by leave one out cross-validation (B) Receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC) by using normalized unique match as classifier (C) using normalized clonecount match as a single predictor in logistic regression, balanced accuracy was evaluated by leave one out cross-validation (D) Receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC) by using normalized clonecount match as classifier.

Tables

Table 1. Basic information of the donors and the libraries

Subject ID	Age group	HLA	Histology (Marsh)	Serology (IgA-TG2)	CD status	Group	Reads	cDNA molecules (TCRα)	cDNA molecules (TCRβ)	Clonotypes (TCRα)	Clonotypes (TCRβ)
CD1357	50-54	DQ2	3a	n.a.	UCD	UCD	1 966 714	7 788	8 701	1 721	2 563
CD1358	35-39	DQ2	3c	93	UCD	UCD	1 783 764	3 857	6 945	1 866	3 249
CD1364	20-24	DQ2	3b	10	UCD	UCD	2 727 644	6 913	9 956	3 218	4 726
CD1368	40-44	DQ2	3c	66	UCD	UCD	2 519 295	6 293	13 086	2 938	5 840
CD1370	50-54	DQ8	1	<1	control	control	838 353	2 206	5 729	692	1 757
CD1386	30-34	DQ2	0	<1	control	control	1 696 311	4 302	4 548	677	1 033
CD1390	18-19	DQ8	3c	40	UCD	control	106 758	567	1 236	318	657
CD1393	25-29	DQ2	3b	9	UCD	UCD	2 269 730	9 746	6 693	2 041	1 999

CD1408	25-29	n.a	0	n.a.	control	control	2 001 839	8 792	4 298	2 051	1 615
CD1409	30-34	DQ2	0	<1	control	control	1 760 049	7 526	4 002	1 341	1 071
CD1422	30-34	DQ2	3a	6	UCD	UCD	1 901 639	3 881	2 998	500	532
CD1428	20-14	DQ2	0	<1	control	control	506 210	924	1 877	270	591
CD1450	35-39	DQ2	0	<1	control	control	1 559 849	2 158	4 035	471	1 468
CD1451	65-69	DQ2	3c	70	UCD	UCD	1 783 685	2 884	2 673	461	842
CD1453	30-34	DQ8	0	5	Potential *	control	1 880 982	4225	3728	808	1148

* Potential CD is defined as positive seology but normal histology.
n.a.: not available

Data Availability Statement:

4 Upon manuscript acceptance TCR sequences will be uploaded to the NCBI Sequence Read Archive (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?)

Figure 1

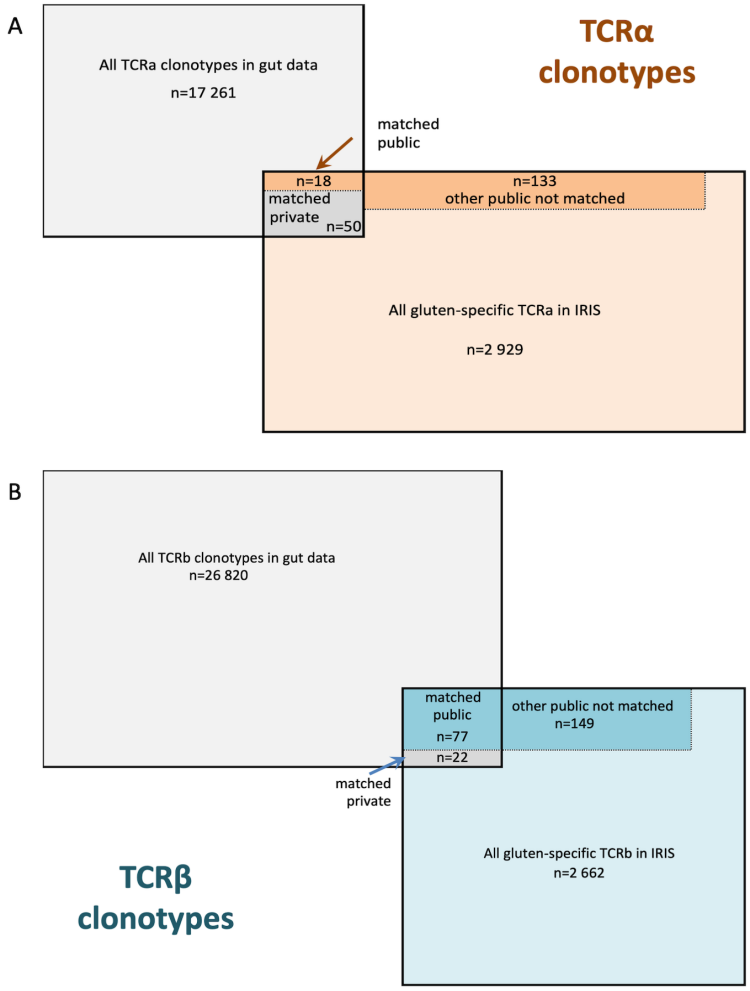


Figure 2

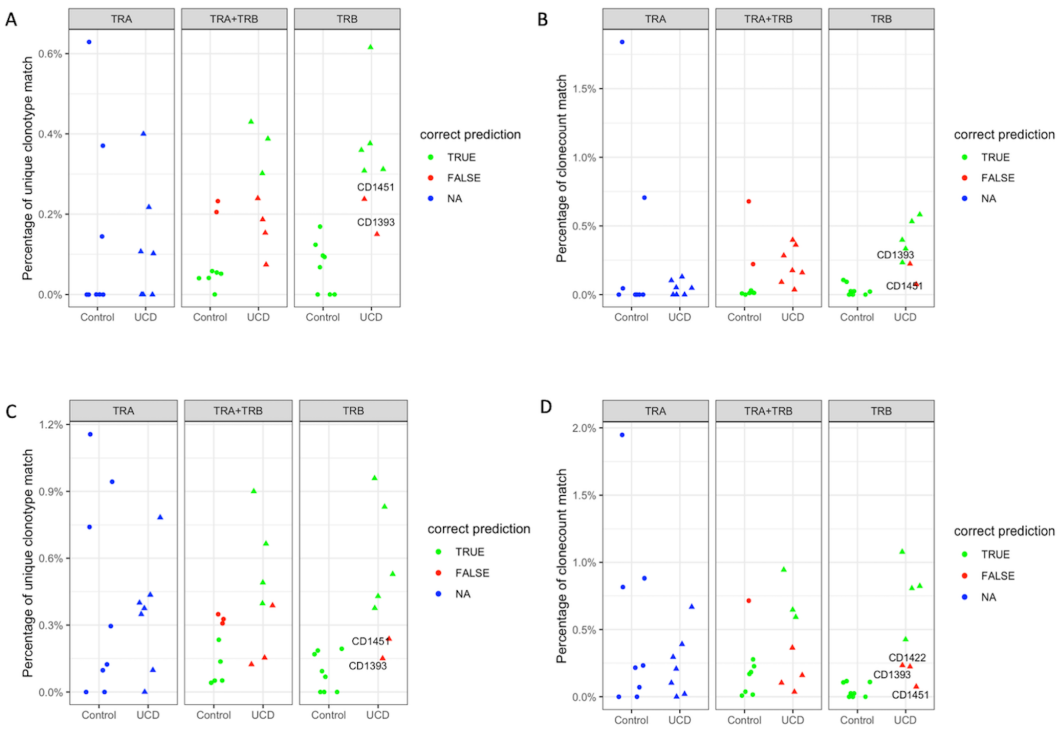
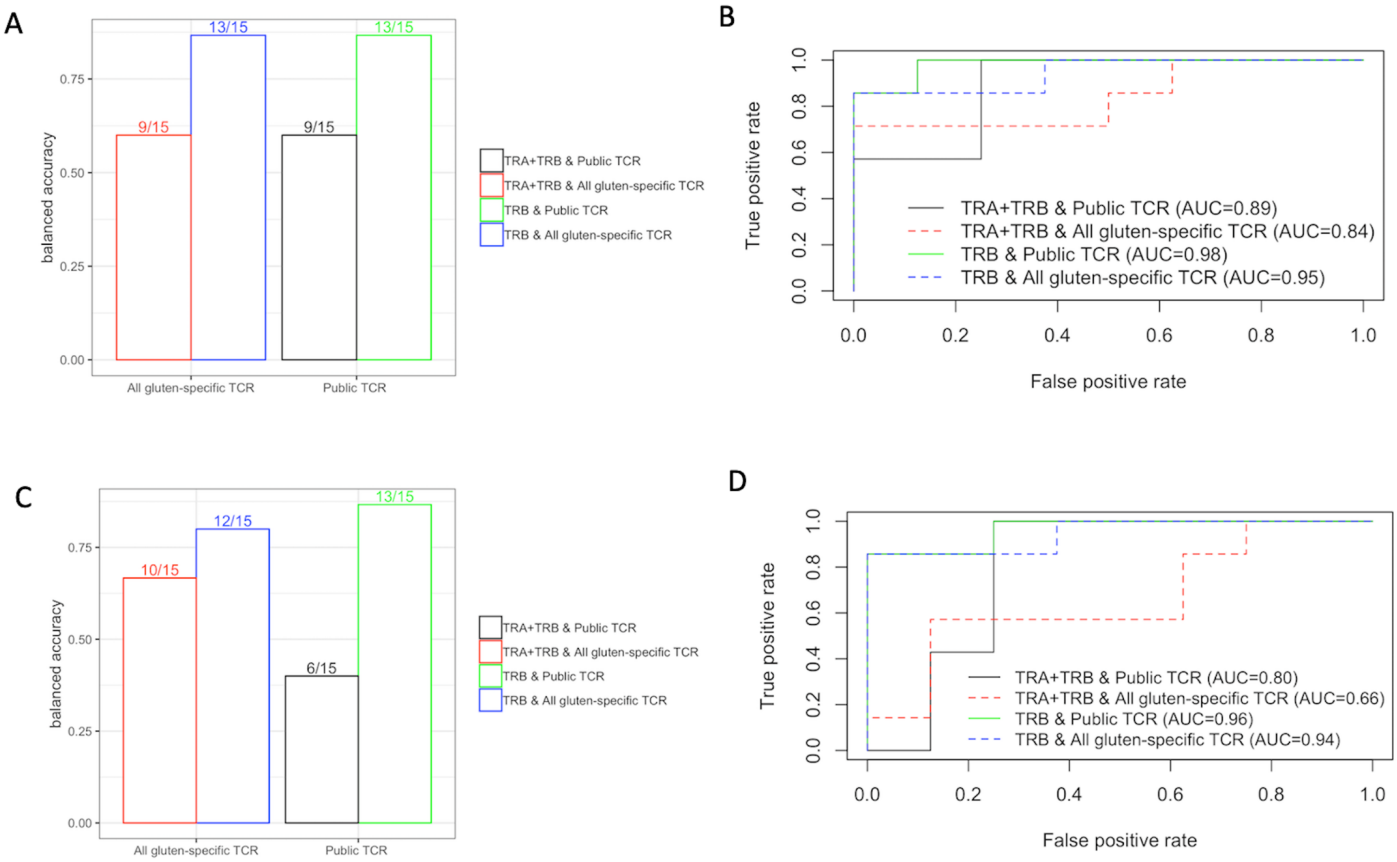


Figure 3



Supplementary Figures

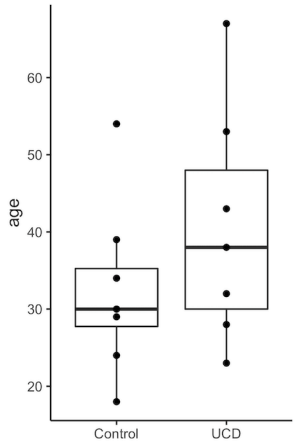


Fig S1. Age of donors in the group of untreated celiac disease (UCD) and the control group.

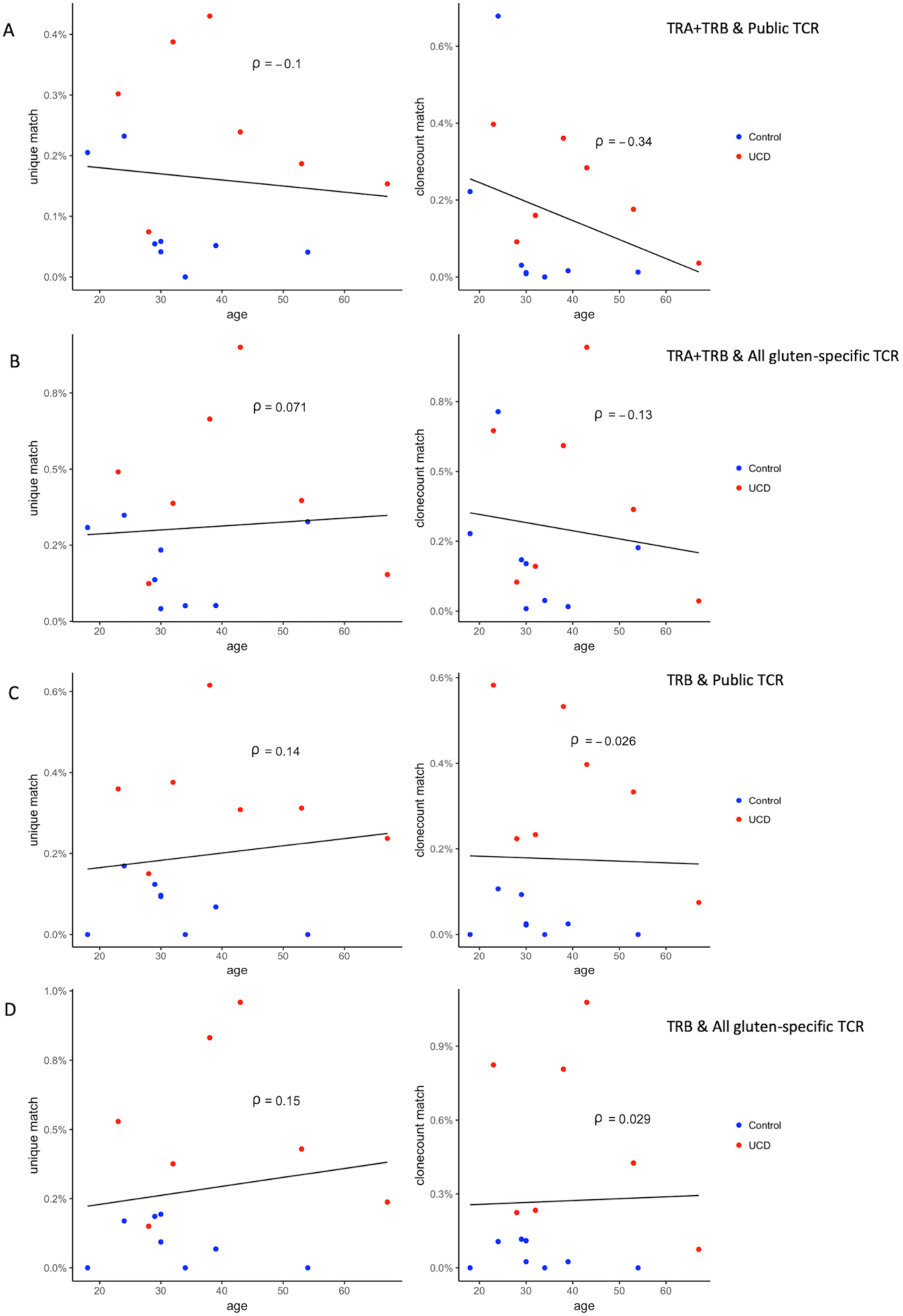


Fig S2. Correlation of age and the predictor, either the normalized unique match (left on each panel) or the normalized clonecount match (right on each panel) in each scenario with fitted line and value of Pearson correlation coefficient. Red represent untreated celiac disease patients (UCD), blue represent the controls.

II

PLOS ONE

RESEARCH ARTICLE

Exploiting antigen receptor information to quantify index switching in single-cell transcriptome sequencing experiments

Ying Yao[1]*, Asima Zia[1☯], Łukasz Wyrożemski[2☯], Ida Lindeman [1,2], Geir Kjetil Sandve[2,3‡], Shuo-Wang Qiao [1,2‡*]

1 Department of Immunology, Centre for Immune Regulation, University of Oslo, Oslo, Norway, **2** K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, Oslo, Norway, **3** Department of Informatics, University of Oslo, Oslo, Norway

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* ying.xiaodai2@gmail.com (YY); s.w.qiao@medisin.uio.no (SWQ)

Citation: Yao Y, Zia A, Wyrożemski Ł, Lindeman I, Sandve GK, Qiao S-W (2018) Exploiting antigen receptor information to quantify index switching in single-cell transcriptome sequencing experiments. PLoS ONE 13(12): e0208484. https://doi.org/10.1371/journal.pone.0208484

Editor: Stephen J Turner, Monash University, Australia, AUSTRALIA

Received: November 18, 2017

Accepted: November 19, 2018

Published: December 5, 2018

Data Availability Statement: Human full-length transcriptomics data is considered to inherently contain sensitive information and is protected by rules from the Regional Committee for Medical and Health Research Ethics South-East Norway and the Norwegian Data Protection Authority. The study is registered in the European Genome-phenome Archive https://www.ebi.ac.uk/ega/studies/EGAS00001002911. All data associated with the study is archived in EGAD00001004081 with an access control. Any data access requests will be

Abstract

By offering high sequencing speed and ultra-high-throughput at a low price, Illumina next-generation sequencing platforms have been widely adopted in recent years. However, an experiment with multiplexed library could be at risk of molecular recombination, known as "index switching", which causes a proportion of the reads to be assigned to an incorrect sample. It is reported that a new advance, exclusion amplification (ExAmp) in conjunction with the patterned flow cell technology introduced on HiSeq 3000/HiSeq 4000/HiSeq X sequencing systems, potentially suffers from a higher rate of index switching than conventional bridge amplification. We took advantage of the diverse but highly cell-specific expression of antigen receptors on immune cells to quantify index switching on single cell RNA-seq data that were sequenced on HiSeq 3000 and HiSeq 4000. By utilizing the unique antigen receptor expression, we could quantify the spread-of-signal from many different wells (n = 55 from total of three batches) due to index switching. Based on full-length T cell receptor (TCR) sequences from all samples reconstructed by TraCeR and TCR gene expression quantified by Kallisto, we found index switching in all three batches of experiments investigated. The median percentage of incorrectly detected markers was estimated to be 3.9% (interquartile range (IQR): 1.7%-7.3%). We did not detect any consistent patterns of certain indices to be more prone to switching than others, suggesting that index switching is a stochastic process. Our results confirm that index switching is a problem that affects samples run in multiplexed libraries on Illumina HiSeq 3000 and HiSeq 4000 platforms.

Introduction

Next-generation sequencing technology has been improved greatly in terms of its data output, sequencing speed, read quality and cost. As the throughput of sequencing platforms is increasing, sample multiplexing becomes practical and widely adopted by many studies. A common

reviewed by Data Access Committee (DAC): EGAC00001000877.

Funding: This study was supported by grants from Stiftelsen Kristian Gerhard Jebsen (project SKGJ-MED-017) to S.W.Q, G.K.S., and L.A.W.; and the Research Council of Norway (project 179573/V40 through the Centre of Excellence funding scheme and project 233885) to S.W.Q., A.Z. and Y.Y.

Competing interests: The authors have declared that no competing interests exist.

way to perform multiplexing is to introduce unique indices into each adaptor during library preparation, so that multiple samples can be pooled and sequenced in parallel. After sequencing, reads can be assigned to their original samples based on the indices. However, the process comes with a risk of molecular recombination, known as "index switching", which causes some of the reads to be assigned to an incorrect sample. In 2015, Illumina adopted a new cluster generating mechanism, known as exclusion amplification (ExAmp), as well as patterned flow cells in their HiSeq 3000/4000/X instruments. By increasing cluster density on the flow cell, these new advances facilitate higher data output, more sequencing reads and shorter running time, thereby reducing sequencing cost substantially. However, because of the increased concentration of free primers produced during library preparation, it has been suggested that this promising chemistry results in a higher level of index switching than conventional bridge amplification [1]. Index switching is essentially a molecular recombination event, which is mainly driven by the presence of free-floating adaptors. A free adaptor could potentially hybridize to remaining complementary sequence of adaptor inside of a cluster and get extended. The index-switched strand can also float off and seed another nanowell on the same lane and generate a new cluster there [1]. The level of index switching depends on the purity of a library, library storage and the method of library preparation. In the Illumina white paper, the level of index switching is shown to increase in a linear fashion with the concentration of free adaptors or primers. PCR-free libraries show higher switching rates compared to library preparation that includes PCR [2].

Index switching may compromise applications that rely on accurate profiling of gene expression, especially those with the need to profile lowly expressed genes. For example, in differential expression analysis of RNA-seq data, it is usually considered to be a good practice to place samples randomly over a plate in order to minimize any potential confounding, such as the edge effect, a common phenomenon in immunoassays. To balance samples from the test group and the control group, sequencing pooled samples from each group on one lane is also preferred for the same reason. However, in an experiment impacted by index switching, those practices make it possible that reads originally from a sample in a test group might be incorrectly assigned to another sample in a control group, which weakens and distorts any real signal. In a single-cell RNA-seq experiment that mainly focuses on detecting heterogeneity among single cells, it is commonly required that hundreds or even thousands of cells are analyzed, and every one of them needs a unique index to be sequenced in a single run. Double indexing, introduced by Kircher et al. [3], increases the scope of indices by using a combination of indices for both forward and reverse adaptors. In an experiment with few samples, both indices can be unique for each sample. Reads assigned to illegitimate combinations of indices are considered to be a result of index switching, and can be easily excluded before downstream analysis. However, even the enlarged index scope with double indexing can hardly satisfy the number of cells to be sequenced in a single cell experiment. When all the combinations of indices are exhaustively used to allow more cells to be sequenced in a single run, there is no way to distinguish the incorrectly assigned reads.

To allow immune cells to be able to recognize a wide range of antigens, each naïve B and T cell expresses a unique immunoglobulin receptor (consisting of heavy and light chain) or T cell receptor (consisting of α and β chain), respectively, through somatic recombination. For human immunoglobulin heavy chain region, there are 123–129 Variable (V) gene segments [4], 27 Diversity (D) gene segments and six Joining (J) gene segments [5]. During B-cell development in the bone marrow, a stochastic process results in a recombined VDJ gene that uses one gene segment of each type. In addition to the combinatorial diversity, there is random deletion of nucleotides as well as insertion of non-germlined encoded P or N nucleotides in the junctions thus creating further diversity. A similar gene rearrangement takes place for the

light chain, except for the absence of the D gene segments. Overall, the total diversity of the immunoglobulin receptor in humans is 5 x 10^{13}. Likewise, T cells deploy a similar strategy for creating receptor diversity and an estimated 10^{18} different T-cell receptors can exist [6].

By using unique pairs of i5 and i7 index adaptors, the rate of index switching was estimated to be up to 2% for sequencing libraries that run on patterned flow cells using ExAmp cluster generation (2). Van der Valk et al. [7] found an average index switching rate of only 0.47% by using inline barcode ligated to reads. Sinha et al. [1] discovered that 5–10% of total mapped reads were assigned to incorrect samples in their single cell RNA-seq study using Smart-seq2 [8] for library preparation and HiSeq 4000 for sequencing. Aside from those inconsistent measurements, it remains unclear whether certain indices are more prone for switching than others. The use of unique immune receptors in immune cells provides us enough diversity for inspecting index switching in single cell RNA-seq data. From three single-cell transcriptomics libraries of human T and plasma cells (terminally differentiated B cells that express B-cell receptor) that were sequenced on the HiSeq 3000 and HiSeq 4000 platforms, we looked for index switching in the expression pattern of immune receptors. Since the receptor might be considered as a cell-specific marker, we have substantial data to quantify the impact of index switching in a single cell experimental setting.

Method

Ethics statement

Study participants provided written informed consent. The study was approved by Regional Committee for Medical and Health Research Ethics South-East Norway (2010/2720).

Datasets

The data used for this study was originally generated to profile the transcriptomics of disease-specific immune cells from intestinal biopsies or peripheral blood from celiac disease patients. We used the Smart-seq2 protocol (8) to perform single-cell RNA-sequencing on 465 immune cells. The immune cells analysed included 215 HLA-DQ2: gluten-(DQ2.5-glia-α1, -α2, -ω1, and -ω2) tetramer-sorted T cells, 247 transglutaminase 2 (TG2)-specific plasma cells, and three unassigned cells in three batches. Correspondingly, Batch 1 contained one plate of 96 wells; Batch 2 contained two (Plate 2 and Plate 3); while Batch 3 contained two plates (Plate 4 and Plate 5). Except Plate 2 which only contained T cells from peripheral blood, T cells and plasma cells were randomly placed into 96-well plates as single cells by index sorting on a FACS Aria II cell sorter. The cells were retrospectively identified based on each cell's high-dimensional immune phenotype. Double indexing was applied in library preparation, thereby every read had an index composed of two segments, indicating which row and column its source well was located. The indices were incorporated in PCR primers that were used at a final concentration of 125 nM. Free adaptors were removed by two rounds of purification with AMPure XP beads. The final sequencing library shows undetectable or negligible number of free adaptors assessed by High sensitivity DNA assay run on the BioAnalyzer (S2 Fig). Samples in Plate 1 were indexed using eight row indices and 12 column indices. After exclusive amplification (ExAmp) process, the samples were sequenced on a single lane of Illumina HiSeq 3000. In the following two batches, we used one set of indices for two plates. 12 indices were used for indexing samples on 12 columns in both Plate 2 and Plate 3, whereas 16 unique row indices were used for eight rows in Plate 2 and eight rows in Plate 3. Each plate was processed independently and the pooled indexed library from the 96 wells in each plate was combined and kept at -20C without further manipulation until sequenced on a single lane of Illumina HiSeq

4000. Likewise, Plate 4 and Plate 5 were prepared in the same way (S1 Table). The output reads were 150 bp paired-end. Average number of reads per cell was around 1.3 million.

Quantifying expression of TR_V gene

To acquire more complete reference transcriptome for immune receptors, we downloaded human immunoglobulin (IG) and T cell receptor (TR) gene sequences in fasta format from IMGT [4], and used them to replace those in human transcriptome GRCh38. Transcript expression levels of all V gene segments were quantified by Kallisto [9]. TPM (Transcripts Per Million) values were used to represent expression for downstream analysis.

Reconstructing T-cell and B-cell receptor

We applied TraCeR [10] to reconstruct full-length T cell receptor sequences from all samples. TraCeR is capable of extracting TCR-derived reads and mapping them to a custom-made reference database that contains all the possible combinations of V and J segments. Instead of retaining at most two TCR sequences with the largest expression per TCR locus in each single cell, which is the default setting based on a reasonable assumption in a single-cell scenario, all reconstructed TCR sequences and their corresponding expressions were used in downstream analysis to capture index switching artifacts. Similarly, we used BraCeR [11] to reconstruct full-length B cell receptor (BCR) sequences form all the samples.

Quantifying spread of signal within and outside of the cross pattern

By using a double indexing strategy, all reads contain two indices that denote respectively a specific column and row. During demultiplexing, a read with a switched index would be incorrectly assigned to a well corresponding to the switched index. Since all index combinations were used exhaustively in all three batches, an index-switched read could not be easily identified. However, the marker gene affected by switching of one of the indices would be detected in wells in the same row or column of the origin well, but at a low expression level. Since TCR sequences are only shared between T cells belonging to the same clonotype, we expect the large majority of cells in a batch to have a unique TCR. This can be exploited to detect spread of signal originating from a well with a given TCR sequence to other wells in the same batch. For each TCR sequences detected in more than two wells in a batch, we defined the well with the highest expression in each batch as origin well. The relative expression of all wells that expressed identical TCR sequences was defined as the ratio of expression in each well to the origin well. Wells that used identical column or row indices as the origin wells were defined as within the cross, and all remaining wells were defined as outside of the cross. To show that index switching was a major cause of signal spread, we counted the number of wells within the cross that seemingly expressed the same TCR sequence as the origin well, and compared it with the number of wells expressing the same TCR sequence outside the cross.

Testing if the spread of signal is more in the source plate than the other plate in the same batch

For each TCR marker in Batch 2 and Batch 3, we divided the above calculated relative expression in wells along the column of the origin well into those in the source plate and those in the other plate, then average the relative expressions in the two groups respectively. A paired sample t-test was conducted for testing if the estimated level of signal spreading in the source plate was higher than that in the other plate in the same batch, which was sequenced on the same lane.

Identifying source well of signal spreading caused by index switching

For each marker used for quantifying index switching, either V gene or full-length TCR sequence, we identified the source well where the cell expressing the marker was located. Based on the expression data, we identified a well as a source of a certain marker for index switching if it met all the following criteria. 1) A well was in a position where the marker was also detected in wells in the same row or the same column (center of the cross). 2) Expression level of the marker was five times higher in this well than in any other well in the same column or the same row. 3) The cell type of the cell that was placed in the well during library preparation must match the cell type of the corresponding marker. For example, a source must be T cell while using TR_V gene segment or TCR as a marker. While using a BCR as a marker, the source cell must be a plasma cell. 4) The expression level of the marker was the highest among all the corresponding TR_V alpha or beta markers expressed in this well.

Criteria for choosing marker, V gene segment or TCR

By assuming that the expression levels for each marker are independent, a marker was eliminated in the process of quantifying the rate of index switching if it complies to any of the following three requirements. 1) A marker was detected in less than three wells. 2) A marker was detected in more than three wells outside of the cross pattern. 3) More than two sources were identified for the same marker in a batch. In cases where both TCR sequence and TR_V gene used by it were qualified, we primarily used TR_V gene segments as markers since they are more readily detected than full-length TCR sequences and most TR_V gene segments were each only expressed by one T cell in the batch. However, in cases where the same TR_V gene segment was used by different TCRs or that cells belonging to the same clone and thus expressing the same TCR were present in the same batch, the V gene segment would be observed in wells outside of the cross pattern. Therefore, we used full-length TCR sequence reconstructed by TraCeR instead of V gene segment as a marker if a V gene segment was detected in more than one well outside of the cross pattern or more than two source wells were identified for a V gene segment.

Calculating the rate of index switching for each index used

If a marker was detected in more than three wells outside of the cross pattern, or more than two sources were identified for the same marker in a batch, it was eliminated from further study for quantifying the rate of index switching. The reasoning was that their spreading signals are more likely to be confounded with each other's, thus compromising the accuracy of quantification of index switching arising from a given well. When two sources of the same marker were identified in one batch, expressions of the marker on presumably overlapped positions were allocated proportionally to the sources' expression, and then the level of signal spread was estimated for the two sources separately. The relative expression values in recipient wells (ratio of expression in the well to the expression in its source well) in the crosses for every identified marker were summed up by columns and rows, and then adjusted by dividing the number of times the well was exposed as a recipient of the identified marker.

Results

High rate of common TCR sequences observed in wells within the same batch

We used TraCeR (9) to reconstruct full length TCR sequences from all samples, including samples of plasma cells, because as a result of index switching we found reads mapped to TCR

also in wells originally containing plasma cells. All the reconstructed receptors, including both productive and unproductive ones, were included. TCR sequences should be unique for most cells, although occasionally two cells share identical TCR sequences due to in vivo clonal expansion, these cases are relatively rare. However, we observed in our data that 182 unique full-length TCR sequences were reconstructed from 288 wells and that 76 of them were detected in more than one well. Closer inspection revealed that the shared TCR expression were most noticeable in plates belonging to the same batch and sequenced on the same lane (Table 1). This lead us to suspect that an unwanted spread-of-signal has occurred.

There could be two major causes of signal spreading, either by index switching or by contamination. Index switching is a known problem associated with libraries sequenced on the Illumina HiSeq 3000 and 4000 platforms.

Index switching is the main cause of the signal spreading

Since switching of both indices is extremely rare, index switching will in most cases cause the signal to spread to wells in the same row (when column index is switched) or the same column (caused by switching of the row index). In other words, index switching will typical spread the signal from the source well to the recipient wells in a cross pattern. To verify that index switching is the main cause of the observed signal spreading, we quantified the spread of signal within the cross pattern and compared it with the signal spreading to wells outside the cross pattern. After excluding TCRs detected in less than three wells, 45 wells with the highest expression of each TCR markers were identified as origin wells. Note that these excluded TCR markers had generally very low expression—the median expression of these markers was 239 TPM (IQR 3–1531), while the median expression of the included markers was 7658 TPM (IQR 3658–14870). The same TCR sequence as in the origin wells was found in 320 out of 1104 wells (29%) in the same row or column of the origin wells, i.e. within the virtual cross patterns. In comparison, among the 6820 wells outside of the cross pattern, only 13 of them (0.19%) were found to express any of the 45 TCR markers (Fig 1A). After normalizing the counts to adjust the different number of wells within and outside a cross pattern, 99.35% of identical TCR sequences observed outside the origin wells were found within the cross patterns.

Furthermore, we excluded the origin wells and wells with 0 relative expression and then looked closely at the relative expression levels of the remaining wells. They were in most cases very low (Fig 1B). 97.7% of them were smaller than 0.05, indicating a spread-of-signal rather than true TCR duplication due to clonal expansion in vivo. Hence index switching was shown to be the main cause for identical TCRs detected in multiple wells. In addition, in a minority of wells, i.e. 7 out of 310, we observed considerable higher relative expressions ranging from 0.1 to 0.99 both within and outside of the cross patterns, indicating that there might exist in a few cases in vivo clonal expansion that resulted in identical TCR sequence in different wells.

Contamination contributes little to the signal spreading

In two of the batches, there were two 96-well plates that were each individually sorted and processed independently, including cDNA amplification, fragmentation and index barcoding by PCR. Purified libraries from each of the plates were kept frozen without further PCR rounds. The final pooled indexed libraries from each of the plates were mixed immediately before sequencing. Thus, contamination should only cause spread of signal of wells within the same plate, whereas index switching would occur both within and across the plate boundary. Hence, we assume that all spread of signal across plate boundaries is caused by index switching whereas the spread of signal within the plate is caused by index switching and possibly in

Table 1. Number of common TCRs constructed by TraCeR.

	Plate 1	Plate 2	Plate 3	Plate 4	Plate 5
Plate 1	17	1	1	2	1
Plate 2	1	49	14	3	3
Plate 3	1	14	35	2	2
Plate 4	2	3	2	26	15
Plate 5	1	3	2	15	27

Number on diagonal position is number of unique TCRs detected in every plate. Others are numbers of common TCRs in two corresponding plates. Samples in Plate 2 and Plate 3 (double outline) were indexed with the same set of column indices and sequenced on the same lane. Similarly, samples in Plate 4 and Plate 5 (double outline) were indexed with the same set of column indices. TCRs observed only in the 5 positive control wells (A1) were excluded.

https://doi.org/10.1371/journal.pone.0208484.t001

addition by contamination. To test if contamination aggravated the signal spreading, we calculated the rate of TCR signal spread from the well with the highest expression in a batch to wells that shared the same column index. The spread of signal within the source plate was not significantly more than across the plates (Fig 2) by a paired samples T test (p-value = 0.24). This indicates that contamination, if any, contributes to a minor fraction of the signal spread we observe.

Two main causes for detecting identical TCR sequences across wells

For every reconstructed TCR sequence, we profiled its expressions in all wells. Indeed, when we closely examined the expression pattern of several TCR sequences that were frequently observed in our data, we found clear cross patterns in most cases (Fig 3A). Apart from incorrectly assigned TCRs reads due to index switching, there were also a few identical TCR sequences that were truly expressed by different cells because of clonal expansion in vivo, such as cells D8 and D10 in Plate 2 (Fig 3B). In this scenario, the randomly located multiple sources give rise to spreading signals in wells along their crossing lines.

Since we have many different TCR sequences that were uniquely expressed by one or very few cells in each batch, we decided to use the immune receptor information to quantify index switching. In addition to TCR sequences, we also explore the probability for using TR_V gene segments and BCR sequences.

TR_V gene segments are more readily detected than full-length TCR sequences

Expression of 84 unique TR_V gene segments for TCR alpha and beta chains were detected and quantified from the 215 T cells by using Kallisto. For every single TR_V gene segment, we profiled its expression in all wells except for well A1 of every plate, where multiple cells were added as control. The results for 23 out of 84 V gene segments showed clear cross patterns in wells within the same batch (Fig 4A). The signal we can get from expression of TR_V genes is generally more complete than that from full-length TCR sequences (Fig 4B) since the TR_V genes are more readily detected.

Reconstructed full-length antigen receptors sequences helps to separate overlapped signals, but depends on sufficient read coverage

Full-length TCR receptor sequences reconstructed by TraCeR were used to distinguish those commonly used V gene segments based on additional information such as J-chain usage and the highly diverse junctional sequence. Fig 5 shows a typical case of this. Profiling the

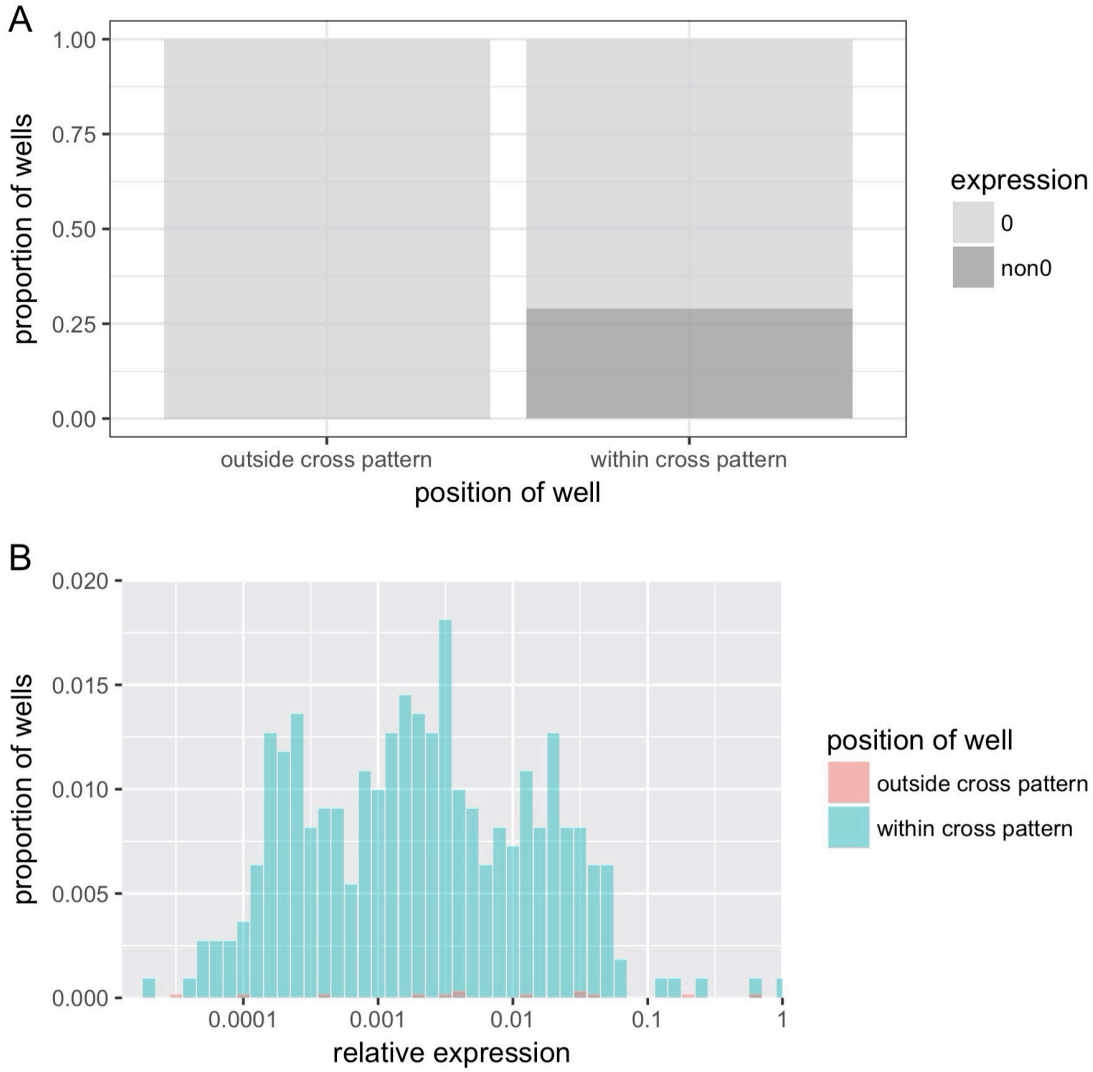


Fig 1. Identical TCR detected outside the origin well was found within the cross pattern. (A) Proportion of wells with detected TCR marker among wells within and outside of "cross patterns" centered at the origin wells for all TCR markers. TCR markers were detected in 320 out of 1104 (29%) within-cross wells and 13 out of 6820 (0.19%) outside-cross wells. (B). Overlaid histogram of relative expression within and outside of "cross patterns" centered at the origin wells for all TCR markers. X axis is shown on a log scale with base of 10. Wells with relative expression 0 were excluded.

https://doi.org/10.1371/journal.pone.0208484.g001

expression of the TR_V gene segment *TRBV20-1* was too uncertain to be used for quantifying index switching. By using TraCeR, we reconstructed four unique TCR sequences that used this TR_V gene segment, and then profiled the expression of each of the four TCR sequences. Comparing with the overlapped expression while using only TR_V gene segment *TRBV20-1* (Fig 5B), the four expression plots using full-length TCR sequences (Fig 5A, 5C, 5E and 5F) were cleaner. All of them showed a clear cross pattern although one of the four TCR sequences (Fig 5A) was detected in fewer wells than the others. It was also noticeable that *TRBV20-1* was expressed at a relatively high level in well G*4 in Plate 5 with a spreading signal in wells along the same row and column. However, the expression level was not as high as in the other three sources where TCR sequences were reconstructed. It was very likely that the TCR in the well used *TRBV20-1*, and spread the signal via index switching. Due to the low number of total

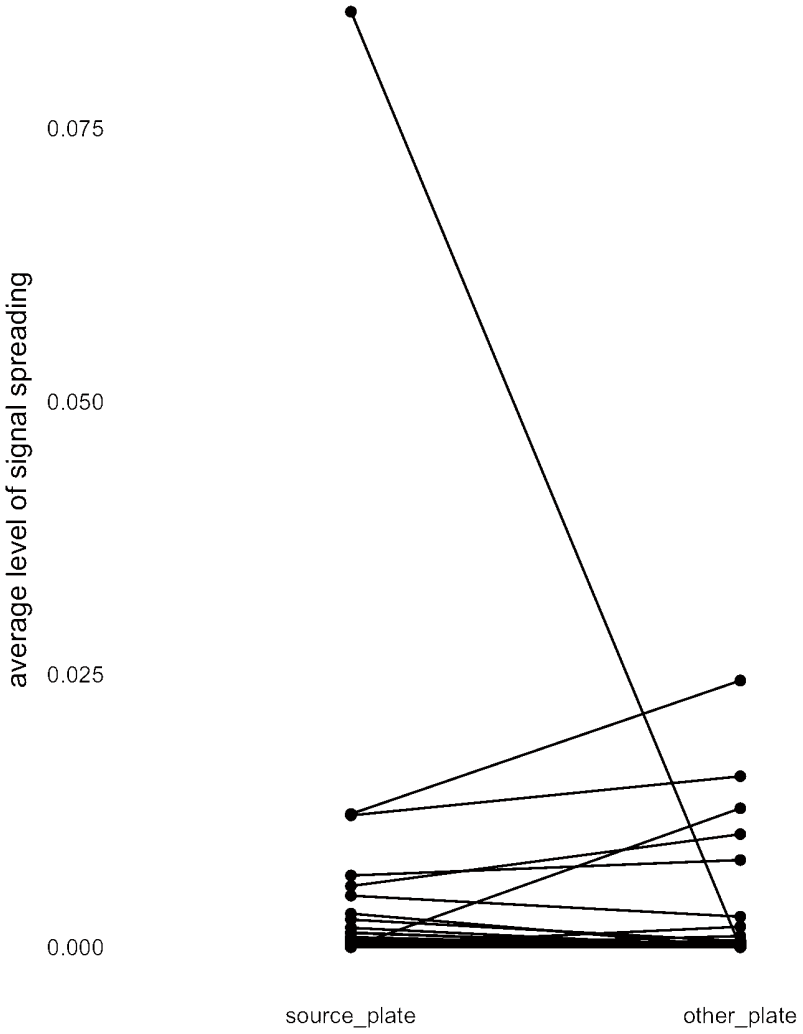


Fig 2. Similar average level of signal spreading to wells of the source plate compared with that to wells in the other plate in a batch sequenced on the same lane. Values on the Y axis are the average relative expression of each TCR sequence in wells in the source plate versus the average relative expression of this TCR sequence in wells in the other plate of the same batch. The wells with the highest expression of each TCR sequence (i.e. the origin wells) was excluded. P-value = 0.24 in paired samples T test.

https://doi.org/10.1371/journal.pone.0208484.g002

reads from this well, the full length TCR sequences could not be reconstructed for this particular well. To successfully reconstruct a full-length TCR sequence, TraCeR requires enough reads to cover an entire TCR chain. In a source cell with few total reads, the coverage might be insufficient to cover the entire receptor chain, in particular the part covering the non-germline encoded and recombined V(D)J which would lead to a failure to assemble a full-length TCR chain.

B cell receptors reconstructed by BraCeR further verifies index switching

BCR sequences expressed by plasma cells were reconstructed by BraCeR(10) and used as a marker for further verification. Since there were no plasma cells in Plate 2 of Batch 2, it was

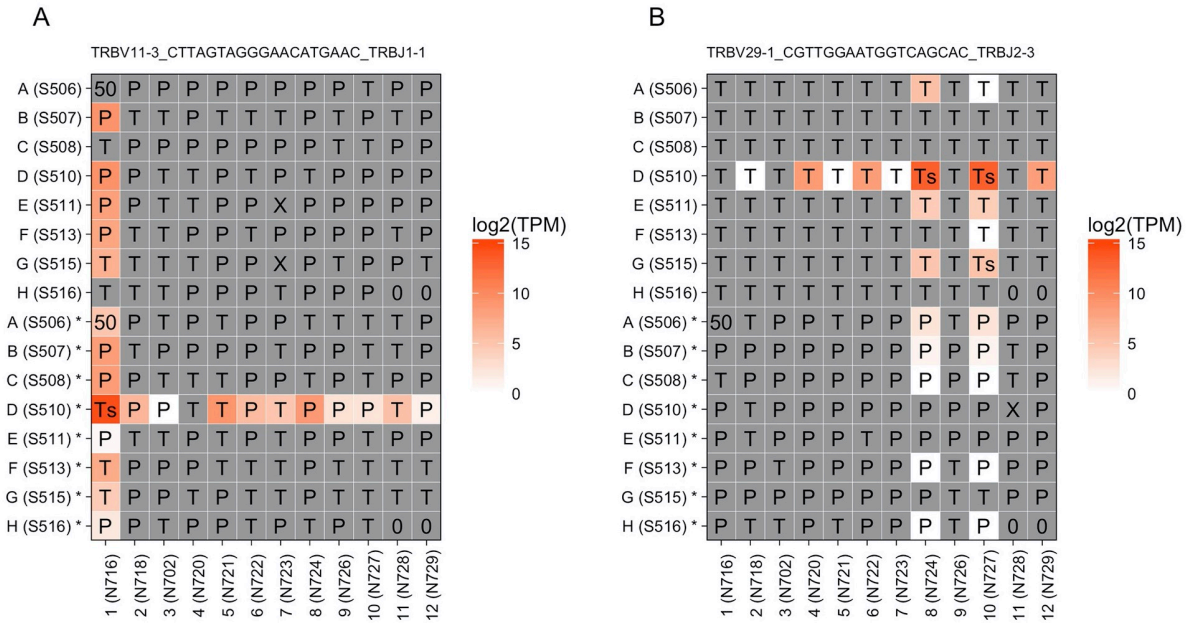


Fig 3. Identical TCR in multiple wells was caused by either index switching or in vivo clone expansion. (A) Expression of TraCeR-reconstructed TCR sequence `TRBV11-3_CTTAGTAGGGAACATGAAC_TRBJ1-1` in Plate 4 and Plate 5 shows one unique source well. (B) Expression of TCR "`TRBV29-1_CGTTGGAATGGTCAGCAC_TRBJ2-3`" in Plate 2 and Plate 3 shows two independent sources of this TCR due to clone expansion in vivo. Two plates of the same batch are bound by common column indices. Upper 8 rows labeled with only letters represent Plate 2 or Plate 4; bottom 8 rows labeled with letters plus a star are from Plate 3 or Plate 5. The indices used are given in the brackets. Cell type is labeled in the corresponding well, T for T cell, P for plasma cell, 0 for empty, 50 for mixture of multiple cells and X for unknown type.

https://doi.org/10.1371/journal.pone.0208484.g003

applied only to Batch 3 of our experiment. To our surprise, 521 unique BCR sequences were observed in Batch 3, in which there were only 108 plasma cells. Upon closer inspection, 322 of the BCR sequences (64%) were only detected in one cell with such a low expression that the median expression for these BCR sequences was 13.2 TPM (IQR 4.4–114.5). We speculated that aside from real cases of lowly expressed BCRs, there were considerable misassemblies by BraCeR. The high expression level of BCR genes in plasma cells combined with index switching has led to the high number of lowly expressed 'artificial' BCR sequences misassembled by index-switched reads from several different source wells. We therefore decided to not include any BCR markers in the quantification of index switching. However, after elimination of the BCR sequences that were detected in one well only, the results for 41 out of 199 BCR sequences showed clear cross patterns in 2 merged plates (Fig 6), confirming that indeed index switching indeed was a major problem also for BCR sequences as well.

3.9% of reads were affected by index switching with no preference for specific indices

After quality filtering, we had found in total 48 markers (15 TR_V gene and 33 TCR sequences) in 55 wells that showed unambiguous signs of spread of signal caused by index switching. This relatively large number of different markers allowed us to quantify and compare index switching between batches and between different indices. For every identified marker, we quantified the level of signal spreading by summarizing its relative expression in its recipient wells. The median percentage of incorrectly detected markers was estimated to be 3.9% (IQR: 1.7%-7.3%). For each sequencing platform, the mean level of signal spreading and

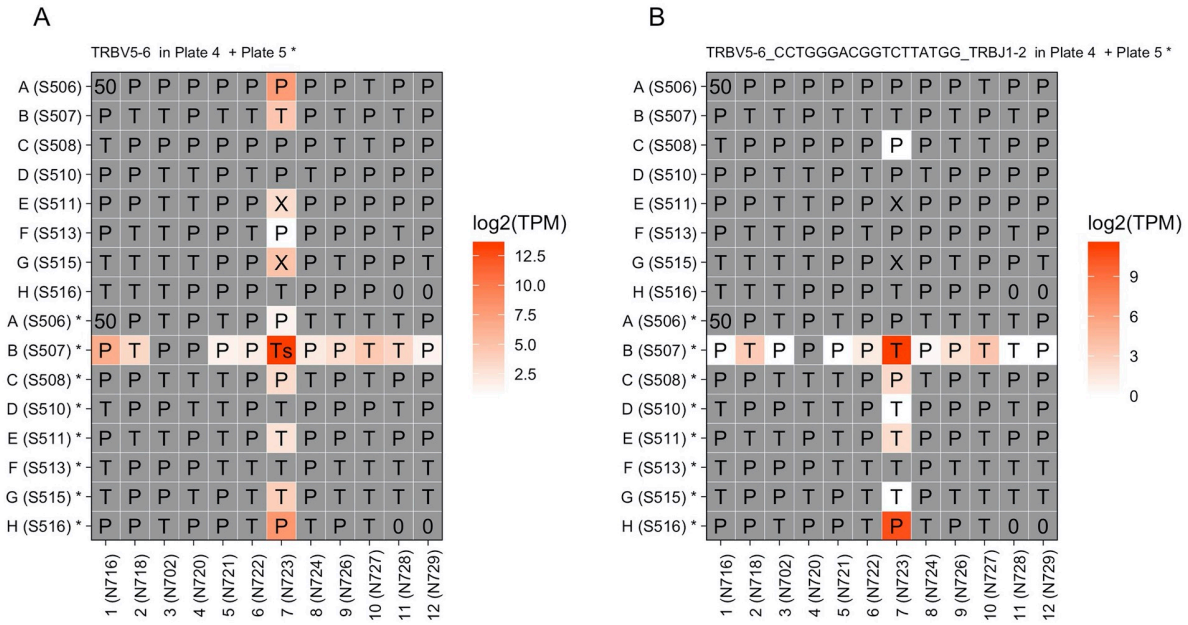


Fig 4. TR_V gene segments are more readily detected than full-length TCR sequences. (A). Expression of TR_V gene segment *TRBV5-6* in Plate 4 and Plate 5. (B). Expression of TCR sequence TRBV5-6_CCTGGGACGGTCTTATGG_TRBJ1-2 in Plate 4 and Plate 5. Two plates are bound by common column indices. Upper 8 rows labeled with only letters represent Plate 4; bottom 8 rows labeled with letter plus star are from Plate 5. The indices used are given in the brackets. Cell type is labeled in the corresponding well, Ts for source T cell, T for T cell, P for plasma cell, 0 for empty, 50 for mixture of multiple cells and X for unknown type.

https://doi.org/10.1371/journal.pone.0208484.g004

systematic variation across all identified markers were assessed. We found no significant difference in sig. ' spre ' ig levels between Illumina HiSseq 3000 and HiSseq 4000 platforms (Fig 7), (p = 0 , Stu nt's t-test).

We wonde l if c ain indices were more prone for switching than others. To test this, for every row-ind and umn-index, we summarized the relative expression level of a marker for every row-index and column-index whenever the marker was detected in wells corresponding to the row-index or column-index. After subtracting those relative expressions in the source-well, we adjusted the values by dividing them by the corresponding frequencies that the row-index or column-index was not used as source. As shown in Fig 8, we did not observe any consistent proneness of index switching to any particular row- or column-index in the three batches of our experiments.

Discussion

This study utilized the unique expression of antigen receptors by T cells and plasma cells to quantify the level of index switching. The large majority of index switching events are limited to one out of two indices and thus lead to spread-of-signal in a clear cross pattern. After confirming that there are strong cross-pattern signals in our data set, we focused the subsequent analyses on markers that allowed accurate estimation of the degree of signal spread along the cross pattern. In order to reduce noise during quantification, we used stringent criteria to select markers that appear to be clearly affected by index switching instead of other confounding factors. More specifically, by assuming that the expression levels for each marker are independent, a marker was eliminated from index switching quantification if it had any of the following three features. 1) A marker was detected in less than three recipient wells. 2) A

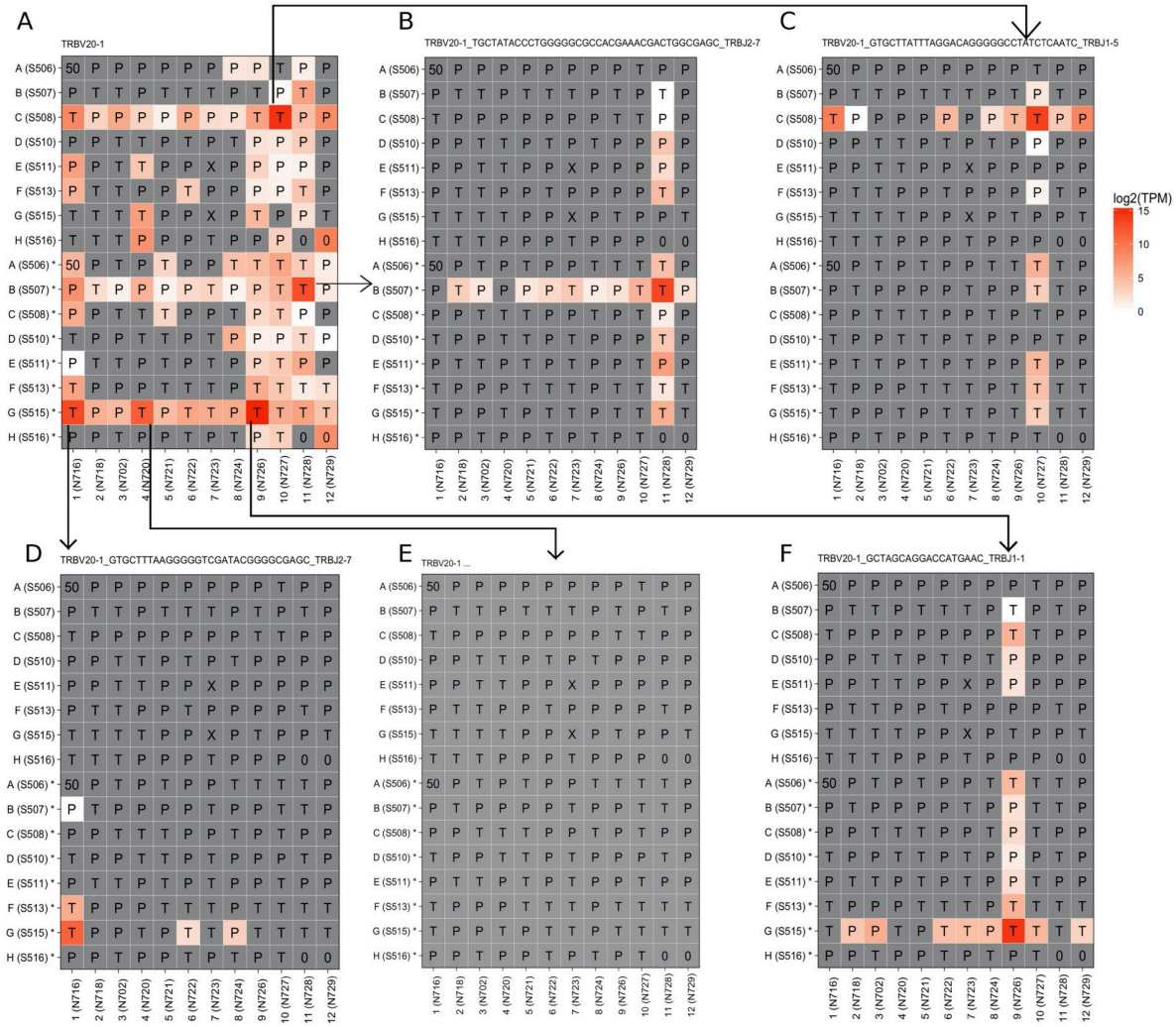


Fig 5. TR_V gene segment TRBV20-1 was used by multiple different TCR sequences in Plate 4 and Plate 5. (A) Expression of TR_V gene segment *TRBV20-1*. (B) Expression of TCR sequence "TRBV20-1_TGCTATACCCTGGGGGCGCCACGAAACGACTGGCGAGC_TRBJ2-7". (C) Expression of TCR sequence "TRBV20-1_GTGCTTATTTAGGACAGGGGGCCTATCTCAATC_TRBJ1-5". (D) Expression of TCR sequence "TRBV20-1_GTGCTATACCCTGGGGGCGCCACGAAACGACTGGCGAGC_TRBJ2-7". (E) Potentially undetected TCR sequence that used *TRBV20-1*. (F) Expression of TCR sequence "TRBV20-1_GCTAGCAGGACCATGAAC_TRBJ1-1".

https://doi.org/10.1371/journal.pone.0208484.g005

marker was detected in more than three wells outside of the cross pattern. 3) More than two sources were identified for the same marker in a batch.

Some markers were found infrequently and in general the expression levels of these markers were very low, even in the presumptive source well with the highest expression level. This is probably due to low number of transcripts amplified in the source well during library preparation or possible misassembly during reconstruction of the TCR marker. In both cases, the information of index switching we could get by using these markers would be unreliable, as the relative expressions of the marker would be confounded either by the detection efficiency or the rate of misassembly. Therefore, we imposed the first filter to exclude those markers. However, we might introduce a slight selection bias in our estimation of index switching rates by omitting these infrequently observed markers.

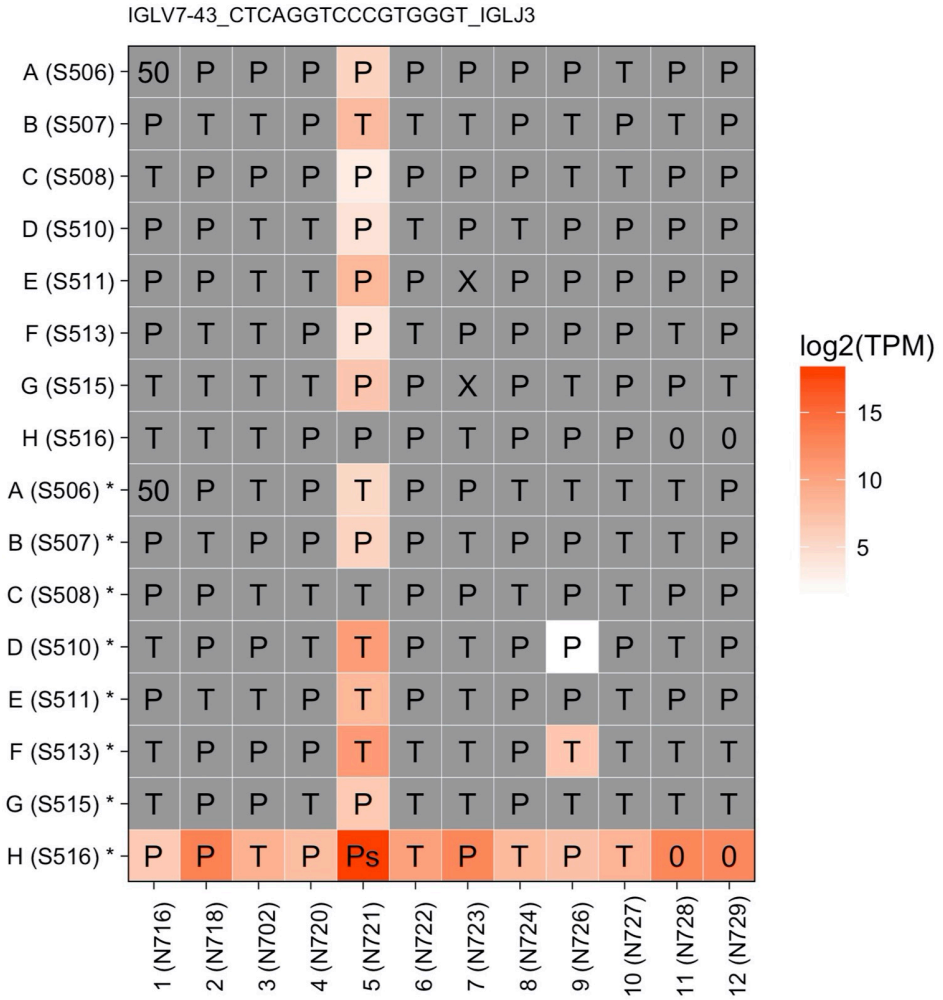


Fig 6. BCR reconstructed by BraCeR further verifies index switching. Expression of BraCeR-reconstructed BCR sequence "IGLV7-43_CTCAGGTCCCGTGGGT_IGLJ3" in Plate 4 and Plate 5. Two plates are bound by common column indices. Upper 8 rows labeled with only letters represent Plate 4; bottom 8 rows labeled with letter plus star are from Plate 5. The indices used are given in the brackets. Cell type is labeled in the corresponding well, T for T cell, Ps for source plasma cell, P for plasma cell, 0 for empty, 50 for mixture of multiple cells and X for unknown type.

https://doi.org/10.1371/journal.pone.0208484.g006

There could be several causes for a marker to be detected in other wells than the source well and its cross. The most common scenario is that the marker is not specific enough, i.e. a V gene segment was used by multiple TCRs, or biological multiplication due to clonal expansion. Most of the cells that substantially expressed the same marker would be identified as sources, but we may fail to identify low-expression markers as the spread of signal could be too low to be detected. Failing to identify such a source would inflate the estimation of index switching rates since some of the signals we detected within the crosses could represent either additional unidentified sources or spread of signal from those. We therefore employed the second filter (excluding markers detected in more than three wells outside of the cross pattern) to eliminate these 'unspecific' markers from the quantification process. In addition, other causes such as sample contamination, dual index switching and misassembly, could also lead to signals outside the cross pattern. Though it is impossible to measure the effect of those potential causes in

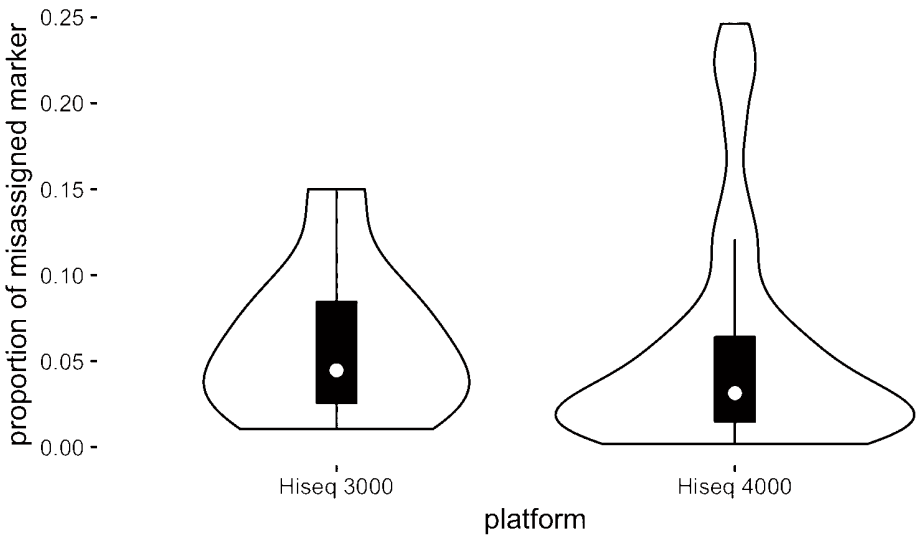


Fig 7. Similar signal spreading levels on Illumina HiSeq 3000 and HiSeq 4000. White spot in the center represents median value in each group. The median proportion of misassigned markers on HiSeq 3000 was 0.044. Median proportion of misassigned markers on HiSeq 4000 was 0.038.

https://doi.org/10.1371/journal.pone.0208484.g007

our experiment setting, it would be alleviated by the second filter. However, despite the second filter for marker selection, and that the result of the paired samples t-test did not indicate contamination to such an extent that it would confound the index switching effect, we cannot completely rule out the possibility of contamination, which might slightly inflate our estimated level of index switching. Similarly, unidentified additional source wells within the cross could also inflate our estimate.

Although unlikely, it is possible that index switching occurs at both ends of a read. We were unable to measure it in this study, since with such a low frequency it might not be distinguishable from those caused by sample contamination. But it would slightly deflate our estimation if there were any.

The tools that we used to reconstruct full-length antigen receptor sequence, i.e. TraCeR and BraCeR, are designed for single cell RNA-seq experiments. They are internally using Trinity [12] to assemble the receptor derived reads into a full-length receptor chain. The likelihood of misassemblies will increase as a large number of reads from different source wells are incorrectly assigned to a cell because of index switching. Compared to TCR transcripts that amounted to around 3% of all reads from T cells, a very large proportion of reads (60%-70%) from plasma cells is derived from the BCR genes due to the cell's biological function as dedicated immunoglobulin factories. Thus, due to the extremely high expression level of BCR genes in the source wells, we observed several cases in which a BCR marker was also detected in wells outside of the cross pattern, such as D*9 and F*9 in Fig 6. However, in these cases, we were unable to ascertain whether this spread was caused by dual index switching, contamination or misassemblies.

There is a theoretical possibility that one index could convert to another as a result of PCR or sequencing errors and thus resulting in a cross pattern that is indistinguishable from that caused by index switching. However, due to the highly distinct indices that were used in our experiments, we believe it is negligible. It is especially true for this particular study, where the measurements of expressions were based on either full-length antigen receptor chains or V gene segments which are approximately twice as long as a read (150 bp). Thus, in this setting,

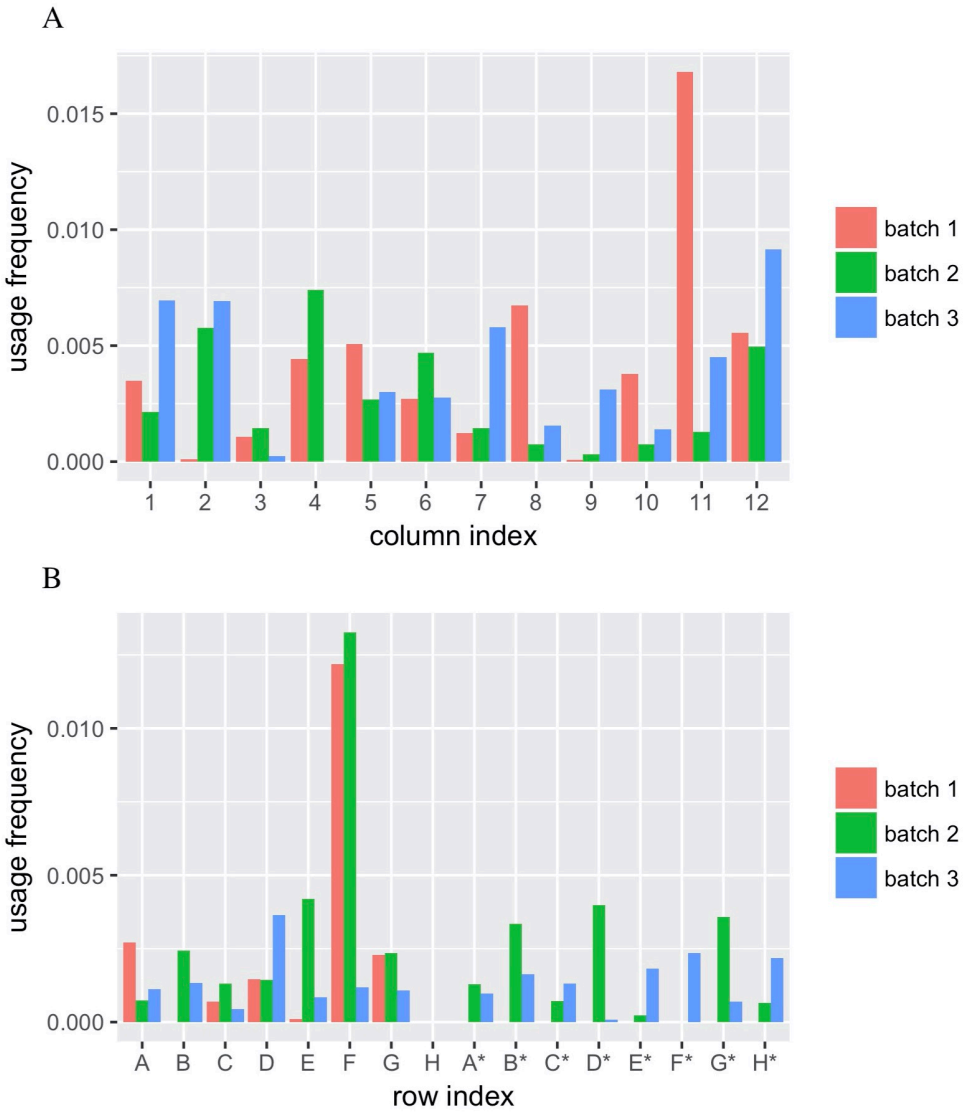


Fig 8. No proneness of index switching to any particular row- or column-index. (A) Estimated probability for every column-index to be used in index switching in three batches. (B) Estimated probability for every row-index to be used in index switching in three batches.

https://doi.org/10.1371/journal.pone.0208484.g008

it seems improbable that a marker would be mis-assigned due to sequencing or PCR error since it is unlikely that the exact same error would occur in a sufficient number of reads covering the marker.

It has been shown that the rate of index switching is dependent on the presence of free adaptors, usually primers with molecular size <100 bases [2]. In our three sequencing batches where indices were incorporated by PCR, the presence of low molecular-size primers was either negligible or undetectable. Compared to other published studies estimating the rate of index switching, which used a single or a handful markers, we have quantified the independent spread-of-signal of 48 cell-unique antigen receptor markers from 55 different wells. Due to the inherent variability in free adaptor amount, expression level, and how successfully signal spread is detected in the recipient wells, it is not surprising to find a relatively large variation in

the estimated rate of index switch from each of these 55 wells. The median rate of 3.9% found in this study is in general agreement with the previously reported rates, ranging from 0.47% to up to 10% [1]. Thus, the index switching rate we have measured in this work is a reasonable estimate of what one could expect on the HiSeq 3000 and HiSeq 4000 platforms when libraries are prepared according to 'best practices'.

Conclusions

In this study, we estimated the level of read misassignment due to index switching in three single-cell transcriptomics libraries sequenced on the Illumina HiSeq 3000 or HiSeq 4000 sequencing platforms. We took advantage of the unique antigen receptor genes expressed by T cells and B cells and could thus quantify the level of signal spreading from 48 different gene markers in 55 different wells. The median percentage of incorrectly assigned markers was estimated to be 3.9% (IQR 1.7%-7.3%). There was no significant difference in index switching level between Illumina HiSeq 3000 and HiSeq 4000 sequencing platforms. Furthermore, we did not detect any consistent pattern of certain indices to be more prone for switching than others, suggesting that index switching is a stochastic process. Our study confirms that index switching is a problem that affects all samples run in multiplexed libraries on Illumina HiSeq 3000 and HiSeq 4000 platforms, and suggests that immune cell receptor information can be utilized to quantify the level of index switching for a given setup of single cell experiments.

Supporting information

S1 Table. Diagram showing the usage of indices and cell type in all wells in the 3 batches.
(DOCX)

S2 Table. Number of common TR_V detected. Number on diagonal position is number of unique TCRs detected in each plate. Others are numbers of common TCRs in two corresponding plates. Samples in Plate 2 and Plate 3 were indexed with the same set of column indices and sequenced on the same lane. Similarly, samples in plate 4 and plate 5 were indexed with the same set of column indices and sequenced on the same lane.
(DOCX)

S1 Fig. Expression of all markers used to quantify index switching, including T cell receptor V-gene segment and TraCeR-reconstructed T cell receptor, in 3 batches. Cell type is labeled in the corresponding well, T for T cell, P for plasma cell, 0 for empty, 50 for mixture of multiple cells and X for unknown type. Ts for T cell identified as a source of signal spreading.
(PDF)

S2 Fig. BioAnalyzer traces of sequencing-ready libraries show negligible amount of small-molecular primer-dimers. After Tagmentation, adaptors including indices were incorporated by PCR. The 96 samples from each plate were then pooled and free primers were removed by two rounds of purification with AMPure XP beads. The pooled libraries were quality-checked on the BioAnalyzer by the High-sensitivity DNA kit before kept frozen at -20C until sequencing.
(PDF)

Author Contributions

Conceptualization: Geir Kjetil Sandve, Shuo-Wang Qiao.

Data curation: Ying Yao, Shuo-Wang Qiao.

Formal analysis: Ying Yao.

Funding acquisition: Shuo-Wang Qiao.

Investigation: Asima Zia, Łukasz Wyrożemski.

Methodology: Geir Kjetil Sandve, Shuo-Wang Qiao.

Project administration: Shuo-Wang Qiao.

Resources: Asima Zia, Łukasz Wyrożemski, Shuo-Wang Qiao.

Software: Ying Yao, Ida Lindeman.

Supervision: Geir Kjetil Sandve, Shuo-Wang Qiao.

Visualization: Ying Yao.

Writing – original draft: Ying Yao.

Writing – review & editing: Ida Lindeman, Geir Kjetil Sandve, Shuo-Wang Qiao.

References

1. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. bioRxiv. 2017; https://doi.org/10.1101/125724
2. Effects of Index Misassignment on Multiplexing and Downstream Analysis [Internet]. Illumina, Inc.; https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf
3. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012; 40: e3. https://doi.org/10.1093/nar/gkr771 PMID: 22021376
4. Matsuda F, Ishii K, Bourvagnet P, Kuma K i, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 1998; 188: 2151–2162. PMID: 9841928
5. Li A, Rue M, Zhou J, Wang H, Goldwasser MA, Neuberg D, et al. Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: implications for the mechanisms of VDJ recombination and for pathogenesis. Blood. 2004; 103: 4602–4609. https://doi.org/10.1182/blood-2003-11-3857 PMID: 15010366
6. Parham P. The Immune System. 4th ed. Garland Science; 2014.
7. van der Valk T, Vezzi F, Ormestad M, Dalen L, Guschanski K. Low rate of index hopping on the Illumina HiSeq X platform. bioRxiv. 2017; 179028. https://doi.org/10.1101/179028
8. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols. 2014; 9: 171–181. https://doi.org/10.1038/nprot.2014.006 PMID: 24385147
9. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016; 34: 525–527. https://doi.org/10.1038/nbt.3519 PMID: 27043002
10. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Meth. 2016; 13: 329–332. https://doi.org/10.1038/nmeth.3800 PMID: 26950746
11. Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao S-W, et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nature Methods. 2018; 15: 563–565. https://doi.org/10.1038/s41592-018-0082-3 PMID: 30065371
12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440