HIGH-THROUGHPUT PAIRING OF ANTIGEN RECEPTOR CHAINS

by

Georgia Mewis

BASc., University of Waterloo, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October, 2015

Abstract

The specificity of antigen recognition by a T cell or B cell is determined by its unique T cell receptor (TCR) or B cell receptor (BCR), each consisting of two, paired polypeptide chains (alpha and beta, or heavy and light, respectively). An immense diversity of receptors is created during T cell and B cell development through a process of gene recombination. Previously, this diversity has been studied by extracting RNA from large numbers of cells, amplifying the alpha and beta chain (or heavy and light chain) transcripts, and then deep sequencing. However, through this process, information on correct chain pairing is lost. In this thesis, I present a high-throughput approach for maintaining paired-chain information in next-generation sequencing libraries. Briefly, a bulk cell population is divided into a number of sub-populations, and TCR or BCR transcripts are independently amplified; chains are considered paired if they co-occur in more sub-populations than expected by random chance. Fundamental to this approach is a reliable, sensitive library preparation chemistry in which a sub-population specific index can be incorporated. Such a chemistry was validated on primary human $CD8^+$ T cells. This approach for antigen receptor chain pairing will enable in-depth studies of immune dynamics, tracking of disease progression, and personalized immunotherapeutics.

Preface

This dissertation does not contain any previously published material. All work was performed by the author, Georgia Mewis, apart from the following contributions: Kathleen Lisaingo assisted with PBMC isolation, Marketa Ricicova performed T-cell enrichment and activation, and Michael VanInsberghe wrote the simulation code, and MiTCR analysis pipeline. Michael VanInsberghe also assisted with primer design. Methods were approved by the UBC BC Cancer Agency Research Ethics Board (Certificate #H09-00581, PAA: #H09-00581-A010).

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to sincerely thank my supervisor, Dr. Carl Hansen, for his inspiration and support. I would also like to thank my committee members, Dr. Rob Holt and Dr. Brad Nelson.

I am very grateful for my amazing lab mates. Thanks to Kevin for getting me started on transcriptome sequencing, Marketa and Kathleen for teaching me how to keep cells alive, and Hans for his optimism and creative solutions. Particular thanks to Mike, for his healthy doses of realism, his invaluable insights into all-things-science, and for providing chocolate and laughs during long TRIzol sessions. Thanks also to Michelle Moksa, for generously sharing both her equipment and expertise.

Also, big thanks to my friends and family: to Esther for all the pep talks over chai tea, to my Mom for keeping me from "lying down in the snow", and to my Dad for teaching me to always "seize the carp".

Lastly, thanks to my incredible husband Keith; I couldn't have done this without you.

Chapter 1: Background

1.1 The Adaptive Immune Response

B cells and T cells are two important classes of lymphocytes involved in the adaptive immune response. A common feature of mature B and T cells is that they display on their surface a unique antigen recognition molecule, a B cell receptor (BCR) or T cell receptor (TCR), respectively, that is the product of combinatorial gene rearrangement. Upon infection, naive lymphocytes with receptors specific to the foreign challenge are activated, resulting in proliferation of the lymphocytes, and differentiation into effector and memory cell subsets. Effector cells execute a multi-faceted, short-lived attack against the foreign challenge, while memory cells allow for a rapid, enhanced response upon re-exposure.

B cells make up the arm of the adaptive immune system responsible for combatting extracellular pathogens. They perform this role through the synthesis and secretion of antibodies, molecules that are identical in structure to the BCR, but are not membrane-bound. Antibodies can neutralize pathogens by direct inhibition of molecular interactions, or can coat extracellular pathogens (opsonization), marking them for phagocytosis or for destruction by other leukocytes (e.g. natural killer (NK) cells).

T cells serve a complementary role in the adaptive immune system in that they are capable of responding to antigens that are within host cells, and therefore out of reach of antibodies. TCRs recognize non-self peptides mounted on major histocompatibility complexes (MHCs), highly polymorphic surface proteins present on all cell types (Figure 1.1 a and b). These non-self peptides may be degradation products of phagocytosed pathogens, or may be the result of genomic mutation. A number of $CD4^+$ T cell subsets exist, defined by the production of different cytokines. The Th1 subset produces interferon-γ, which promotes activation of macrophages. The Th2 subset produces IL-4, which stimulates production of IgE antibodies and IL-5, resulting in elimination of helminthic parasites by eosinophils. The Th17 subset secretes IL-17 and Il-22, which recruit neutrophils and monocytes to the site of infection [1]. Subsets are defined in a similar way for $CD8^+$ T cells, although the functional differences between subsets is not yet as clear as for $CD4^+$ T cells. Apart from their classic role of directly killing infected or cancerous host cells, $CD8^+$ T cells have been show to effect CD 4^+ T cell development, the Tc1 subset promoting development of Th1 cells, and the Tc2 subset promoting development of Th2 cells [2]. There are also regulatory subsets of both $CD4^+$ T cells and $CD8^+$ T cells that block activation of self-reactive T cells. $CD4^+$ and $CD8^+$ T cell subsets have been reviewed in [3-5].

The TCR is a heterodimer consisting of two polypeptide chains termed alpha (α) and beta (β) (γ and δ in a small subset of T cells). The BCR is a Y-shaped molecule made up of two identical, disulphide-

linked heterodimers, each with a light chain (κ or λ), and a heavy chain (μ, δ, γ, ε, or α). The incredible diversity in TCRs and BCRs is generated during a process called somatic recombination. The antigen receptor genome loci contain a number of variable (V), joining (J), and constant (C) region sequences, with the Ig heavy-chain and TCR beta-chain loci also containing diversity (D) gene segments. During lymphocyte maturation in the thymus (T cells) or bone marrow (B cells), one of each V, (D) and J gene segments are recombined to form a V-(D)-J exon. This gene is transcribed and the V-(D)-J exon is spliced to one of the C-region exons (Figure 1.1 c). The number of possible V(D)J combinations is about ~10^6 for BCRs, and about ~3×10^6 for alpha-beta TCRs. Deletion of germline-encoded bases and addition of nontemplated nucleotides at the recombination junctions increases the potential diversity to about ~10^{11} for BCRs and ~10^{16} for alpha-beta TCRs. The hypervariable region encompassing the recombination junctions is called the complementarity-determining region 3 (CDR3) and is the primary determinant of antigen specificity [1].

Only a fraction of the potential TCR and BCR repertoire (10^6- 10^7 clones [1]) is expressed in each individual. The identity and abundance of clones that make up this fraction is non-random; biases in the somatic recombination machinery, positive and negative selection, and exogenous antigen exposure all help to shape an individual's immune repertoire. Exogenous antigen exposure and subsequent clonal expansion not only alters frequencies of clones in the repertoire, but also creates increased diversity in the B cell subset, as clonal expansion is accompanied by somatic hypermutation in BCR variable regions. During this process, known as affinity maturation, a single B cell gives rise to a "clade" of closely related cells that make different but closely related antibodies sharing the same gene recombination and junctional structure.

In these ways, an individual's immune repertoire is imprinted with a history of past immune challenges, and an instruction of future immune responses, making profiling of the repertoire of profound interest to both researchers and medical professionals. Direct applications of immune repertoire analysis include: tracking of disease progression, therapeutic antibody discovery, vaccine development, and personalized immunotherapeutics such as adoptive T cell therapy, in which a person's own T cells are genetically engineered to express TCRs recognizing tumour-specific antigens.

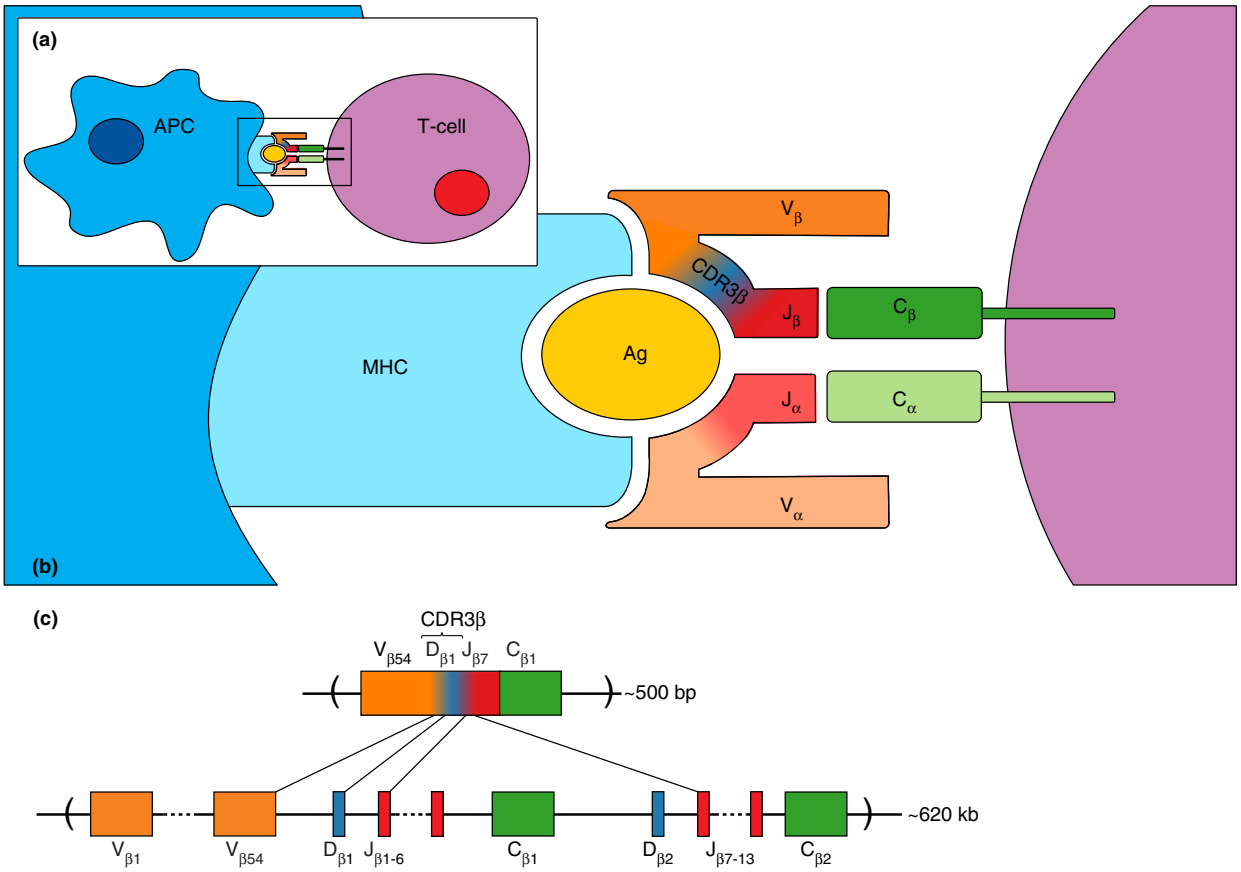


Figure 1.1: TCR-antigen-MHC interaction and TCR gene recombination.

(a) T-cell encountering an antigen-presenting cell (APC). (b) APC presenting a peptide antigen (Ag) on its major histocompatibility complex (MHC), and T-cell engaging the peptide-MHC complex through its T-cell receptor (TCR). The TCR is composed of two chains, α and β, each with a constant region (C) anchored to the cell membrane, and joining (J), diversity (beta chain only), and variable (V) regions. The VJ (for TCR-α) or VDJ (for TCR-β) junction makes up the complementarity determining region 3 (CDR3). (c) Somatic recombination of the TCR-β locus (reproduced with permission from **[6]**)

1.2 Repertoire Profiling

TCRs and BCRs can be directly probed on the surface of living cells using FACS and a panel of fluorochrome-conjugated monoclonal antibodies specific for the variable regions. This approach is simple and fast, and allows for a rough determination of the clonal structure of the repertoire. At the same time, cells can be easily sorted based on surface markers, enabling simultaneous diversity profiling of a number of different lymphocyte compartments. The output, however, is low resolution; no information is provided on CDR3 composition or J-gene usage. In addition, important lymphocyte populations can be missed entirely because specific antibodies are unavailable for some variable regions. For example, a

commercial kit for profiling TCR beta chains (IOTest betamark kit from Beckman Coulter), consisting of a panel of antibodies directed against 24 different TCR Vβ regions, claims to cover only 70% of the repertoire.

More commonly, the TCR and BCR are studied at the nucleic acid-level i.e. the messenger RNA (mRNA) encoding the TCR or BCR, or the antigen receptor locus on the genomic DNA (gDNA). The vast number of tools available for analysis of nucleic acids make repertoire profiling at this level appealing. In the early days, CDR3s were PCR amplified from mRNA, and run on an electrophoretic gel. The visualized CDR3 length distribution, or spectratype, was then used as a measure of repertoire diversity [7]. This method was largely replaced as sequencing became readily available.

Sanger sequencing of TCR and BCR genes has allowed for many important insights into the adaptive immune system, and is still extensively used today [8-11]. Each Sanger sequencing reaction is limited to a single sequence. In the case of TCR or BCR genes this necessitates starting from single cells or clones and separately amplifying, purifying and sequencing heavy and light, or alpha and beta chains. Because the alpha chain locus does not show genotypic allelic exclusion, further cloning is often required to identify multiple alpha chains from a single cell. This can become incredibly labour-intensive; for example, Han *et al.* found multiple alpha chains in 58% of single cells analyzed (in 14% of these, both alpha chains were productive) [12].

Because of its limited throughput, Sanger sequencing allows for examination of only a small sliver of the potential antigen receptor repertoire. The development of next-generation sequencers, capable of simultaneous analysis of millions of different sequences, has made in-depth- or even complete-repertoire profiling of an individual a possibility. In 2011, next generation sequencing was used to exhaustively sequence the peripheral TCR beta repertoire of a healthy male; 1.06 million distinct TCR beta nucleotide sequences were reported [13]

The incredible throughput comes at a cost, however. Next-generation sequencing of bulk populations offers no information about which transcripts originated from the same cell, and therefore which alpha and beta or heavy and light chains pair to form a functional receptor. Only rough predictions of pairings can be made by matching chains with similar read frequencies, or, to account for PCR biases, similar cDNA counts [14]. Two main strategies exist for tracking the alpha and beta or heavy and light transcripts that come from a single cell: barcoding transcripts with a cell-specific unique sequence (Indexed PCR), or physically attaching the transcripts from a single cell before pooling (Fusion PCR). A third approach, presented in this thesis and also independently pursued by Adaptive Biotechnologies [15], is based on statistical inference of chain pairing.

I first describe general considerations and molecular strategies for generation of repertoire sequencing libraries, and then review published reports of antigen-receptor pairing.

1.2.1 Choice of Template

gDNA and mRNA can both be used to generate libraries for next-generation sequencers, the better choice depending on application. In situations where accurate clonotype frequencies are required, gDNA is generally preferred, as there is only one copy of each rearranged antigen receptor locus per cell (i.e. in the absence of PCR bias, the number of sequencing reads is directly proportional to the abundance of a clone). Transcriptional differences between cells obscure this relationship when using mRNA as template. This is particularly true for unsorted B cell populations; plasma cells can have up to 100× more transcripts than naive B cells [16]. The single copy per cell can make detection more difficult, however. Libraries prepared from gDNA often require more PCR cycles, in which errors and bias can be introduced.

1.2.2 Methods for Molecular Amplification

mRNA templates must first be reverse transcribed, using a C-region primer (mRNA 3'-end), to produce complementary DNA (cDNA). cDNA or gDNA is then commonly amplified by multiplexed PCR with C-region reverse primers and a set of V-region forward primers. Han *et al.*, for example, use a set of 74 TCR variable region primers (38 alpha, 36 beta) to amplify TCR sequences from cDNA [12]. Adaptive Biotechnologies, a biotechnology company started out of the Fred Hutchinson Cancer Research Center, offers a service and kit (immunoSEQ) for profiling both BCRs and TCRs that is based on this multiplexed strategy.

Problems arise from having such a large set of primers, however. Because each primer has a different amplification efficiency, multiplexed PCR libraries can be substantially biased (i.e. sequencing read frequency may not be proportional to the frequency of that clone in the starting repertoire). Extensive reaction optimization and complex strategies are thus required to reduce bias [17]. It is also common to get mispriming, which generates unwanted side products.

The use of large multiplexed primer sets can be avoided by adding a universal PCR priming site to the 3' end of the cDNA. This is accomplished by ligation, 3'-end poly(A) tailing [18], or, most commonly, template switching. Template switching, also referred to as SMART (**S**witching **M**echanism **a**t 5' End of **R**NA **T**emplate), relies on the terminal-transferase activity of a reverse transcriptase derived from Moloney Murine Leukemia Virus (M-MLV). When the M-MLV reverse transcriptase reaches the 5'-end of the mRNA, it adds a number of non-templated bases, primarily cytosine. Then, when an oligonucleotide with a number of guanine bases at its 3'-end (called the "template switching oligo") is

added to the reaction, annealing of the complementary bases occurs, and the M-MLV reverse transcriptase "switches" template and begins copying the template switching oligo (TSO) sequence. PCR can then be performed using a single forward primer bearing the same sequence as the TSO.

There is a substantial body of work focused on the design of the TSO. Isomeric nucleotide bases have been incorporated at the 5'-end to prevent TSO concatamerization [19], and blocks such as a C3 (propyl) spacer have been added at the 3'-end to prevent cDNA synthesis primed by the TSO [20]. Picelli *et al.* observed a two-fold increase in yield by substituting the last of three guanylates in the TSO for a locked nucleic acid (LNA) guanylate [21, 22]. This LNA TSO was one of the main optimizations introduced in a recent whole-transcriptome profiling protocol dubbed Smart-seq2 [21, 22].

1.2.3 Strategies for Retaining Paired-chain Information

1.2.3.1 Indexed PCR

Paired-chain information can be retained in next-generation sequencing libraries by attaching a unique sequence to the antigen receptor transcripts from each single cell. An index is generally incorporated into the final sequencing construct with either the forward or reverse PCR primer. *N* cells can be uniquely identified using *N* indexes, with single-end indexing, or $2 * \sqrt{N}$ indexes, with dual-end indexing. For simplicity, a unique index is often used for each row and each column of microwell plate. For example, to dual-index a 384 well plate, 24 column indexes, and 16 row indexes are required.

Busse *et al.* used dual-indexed PCR in an attempt to pair heavy and light chains from 1152 FACS-sorted single murine B cells [23]. 454 sequencing of the libraries revealed low efficiency of amplification; only 404 reactions contained both heavy and light chain reads. In addition, their library construction strategy used is not cost-effective for analysis of human Igs, as 14 variable region forward primers for each index must be synthesized (human: 7× Igλ-V, 3× Igκ-V, 4× IgH-V [24]; mouse: 2× Igλ-V, 1× Igκ-V, 1× IgH-V).

Han *et al.* reported an improved dual-indexing strategy for pairing of alpha and beta or heavy and light chains [12]. Universal sequences were added to the end of each amplicon in the second PCR, so that in the third PCR only a single forward indexing primer, and a single reverse indexing primer were required. The flexibility of this strategy allowed them to amplify and index not only the TCR genes, but also a number of genes characteristic of different T cell subsets (i.e. genes encoding cytokines and transcription factors). Paired sequences were identified in 58%-82% wells containing FACS-sorted single $CD4^+$ T cells.

Indexed PCR products, in contrast with fusion PCR products, are simple to sequence and are compatible with Ig gene cloning and expression, allowing for direct assessment of antibody specificity.

Indexed PCR has a number of drawbacks, however. First, care is required when designing index sequences to limit primer-primer interactions, and, for Illumina platforms, to ensure adequate colour balancing (i.e. for proper image registration there must be a balance of the four fluorescently labeled nucleotides in each cycle). Second, reads can be assigned to the wrong cell as a result of cross-contamination of indexes (during oligonucleotide synthesis and purificiation, or during liquid handling) or because of errors introduced into index sequences (during oligonucleotide synthesis, amplification or sequencing). Finally, the throughput of indexed approaches is limited due to the physical requirement of adding a unique indexed primer to each reaction.

1.2.3.2 Fusion PCR

A second strategy for retaining paired-chain information involves physically attaching the alpha and beta (or heavy and light) transcripts from each single cell. While fusion of transcripts can be accomplished by a number of different molecular approaches, published reports of alpha-beta and heavy-light pairing by fusion exclusively use a multiplexed, overlap-extension PCR; the set of variable region forward PCR primers are designed to have one of two complementary "tails", which hybridize and extend during PCR.

In one report from the Georgiou group, single B cells were first distributed across a polydimethylsiloxane (PDMS) microwell plate (125 pl/well), then lysed, and mRNA was captured on poly-(dT) magnetic beads [25]. Beads were then recovered and emulsified with RT/linkage PCR mix. The final ~850-bp linked VH:VL DNA product was sequenced on an Illumina MiSeq with 2 × 250-bp reads, allowing for complete CDR3 coverage, but only partial coverage of the variable region. For this reason, somatic variants of clonally related B cells could not be distinguished. This method was used to profile three human B-cell subpopulations: i. 61000 peripheral IgG^+ B cells from a healthy donor, ii. 400 peripheral Tetanus Toxoid-specific plasmablasts, isolated 7 days after Tetanus Toxoid immunization, and iii. 8000 peripheral memory B cells, isolated 14 days after influenza immunization. From these samples only, i. 2716, ii. 86, and iii. 240 heavy-light pairs were identified, respectively, demonstrating low efficiency.

Turchaninova *et al.* describe an approach in which single cells are isolated in emulsion droplets instead of PDMS microwells, eliminating the need for microfabrication [26]. Heat lysis, reverse transcription, and overlap-extension are performed in the same droplet, and then the emulsion is broken to carry out a pooled, nested PCR (Figure 1.2). In initial experiments, Turchaninova *et al.* observed the predominant products to be random overlap extensions of chains originating from different droplets. By introducing an oligonucleotide to block 3'- ends of free alpha and beta chain PCR products, before nested PCR, they were able to eliminate random overlap extension and extract only native pairings. DeKosky *et*

al. did not perform the controls required to identify this random overlap extension issue, and so their pairing accuracy may be low; the false discovery rate they report is based only on mispairing of a spiked in control cell line [25].

With the PCR suppression technique, Turchaninova *et al.* extracted ~700 alpha-beta pairings from one million starting PBMCs. This low efficiency restricts the method to pairing of highly represented TCRs; a clone at a frequency of 1000 cells per million PBMCs (~0.2% of T cells) was at the limit of detection. The method was also restricted to TCRβV7-family.

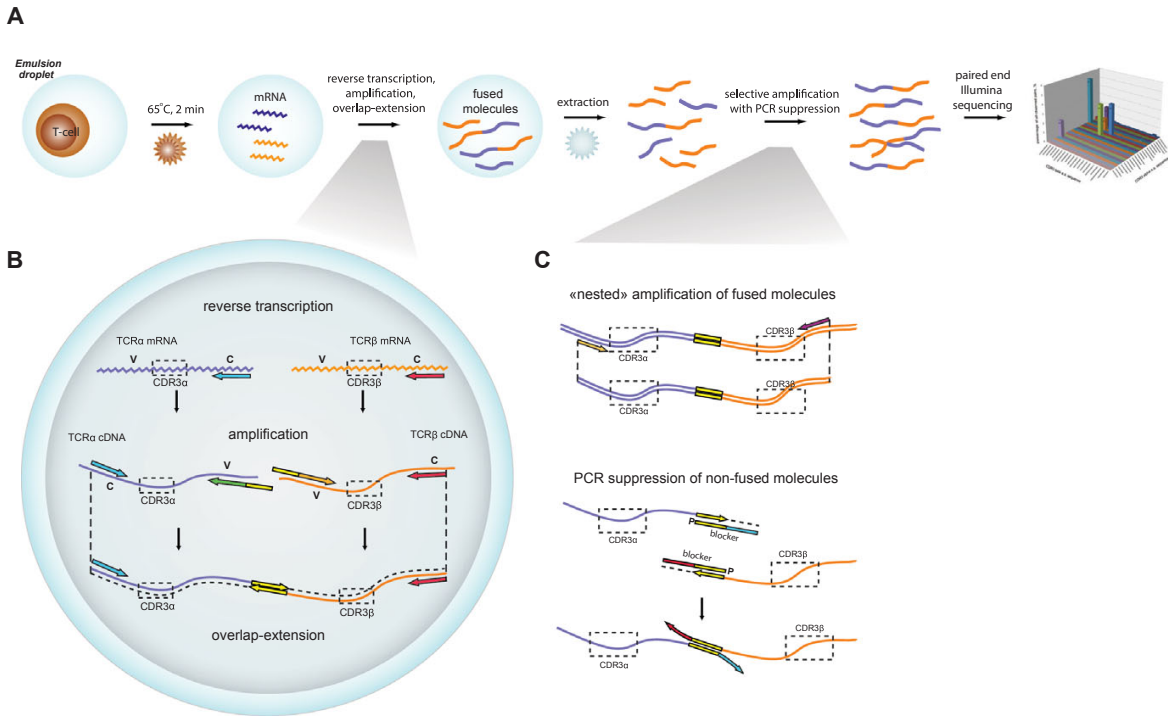


Figure 1.2: Fusion PCR approach for pairing of TCR alpha and beta transcripts

(a) Flowchart of approach. Single T-cells are isolated in emulsion droplets, heat lysis is performed, and mRNA encoding TCRα and TCRβ chains is reverse transcribed. cDNA is amplified by overlap-extension PCR. The emulsion is then broken, and a nested PCR is performed (b) Schematic of reverse transcription and overlap-extension PCR. (c) Use of blocking oligonucleotide to suppress amplification of non-fused molecules (reproduced with permission from **[26]**)

Chapter 2: Design

2.1 Design Goal

All previously described approaches for antigen receptor chain pairing are single-cell methods, and as such, are fundamentally limited in throughput, are high-cost, or require complex equipment and are technically demanding. The goal of this research was to develop a method for pairing TCR and BCR chains that is both high-throughput and accessible, requiring only standard lab equipment and commercially available reagents. Our microwell plate-based method relies on the inherent clonality of the sample, that is, the presence of multiple cells that are progeny of a common B cell or T cell ancestor, and hence have the same rearranged genome loci. This redundancy allows for pair determination by statistical analysis of the co-occurrence of any given pair of chains across a number of sub-pools.

2.2 Workflow

The proposed workflow is as follows:

1. Cells are distributed, by pipette or FACS, into wells of a microwell plate at an occupancy of hundreds to thousands of cells per well
2. mRNA encoding TCR or BCR chains is reverse transcribed and PCR amplified (a well-specific index is added during this step)
3. Products are pooled, purified and sequenced
4. The occurrence patterns of alpha/light and beta/heavy chains across the plate are compared, with significant co-occurrence indicating a putative pair

Figure 2.1 illustrates the concept of pairing by co-occurrence, with a dot indicating the presence of a hypothetical chain in a given well. For example, beta chain B-0001 appears in wells 2, 6, 12, 18 and 19. Because alpha chain A-0007 appears in these same wells, and only these wells, it is likely that B-0001 and A-0007 are paired. In our current implementation, step 4 involves constructing a contingency table for each alpha-beta (or heavy-light) combination and applying the Fisher Exact Test to compute a p-value.

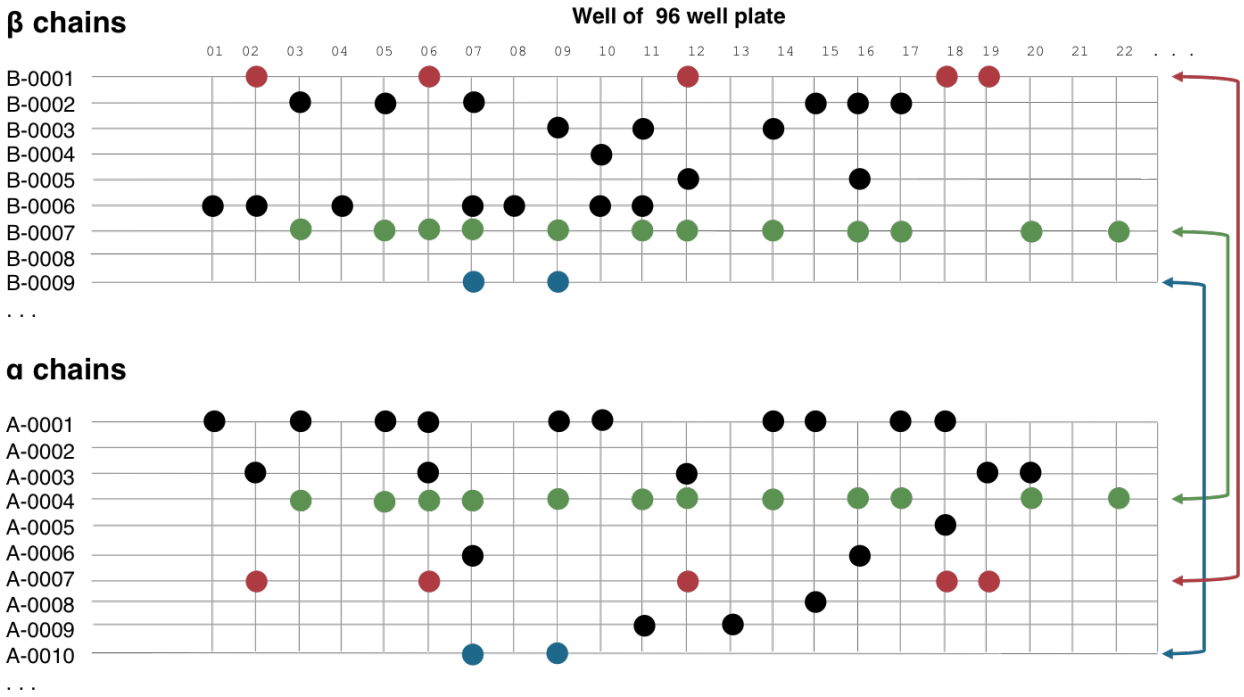


Figure 2.1: Concept of chain pairing by co-occurrence

Each row corresponds to a different, hypothetical chain. Each column corresponds to a specific well of a 96 well plate. A dot represents the presence of a specific chain in a specific well. B chain B-0001 and α chain A-0007 occur in all of the same wells (red dots), so it is likely these chains originated from the same clone, and pair to form a functional TCR. The same is true for B-0007 and A-0004 (green dots), and B-0009 and A-0010 (blue dots).

2.3 Cell Distribution

Intuitively, to enable pairing, the same TCR/BCR must occur in more than one well, but not every well, of the plate. The number of occurrences of a particular TCR/BCR is determined by the frequency of the clone in the starting sample, the extent of expansion performed before splitting the population, and the number of cells per well. Thus, by simply changing the extent of expansion, or the number of cells per well, the method can be tuned to pair clones in a frequency range of interest.

To explore the effect of the number of cells per well (N_{well}) in more detail, a simple simulation was performed. The input was an exhaustive set of TCR beta chain sequences from a healthy human male ([13], male 1). N_{well} sequences from the pool of 494 796 distinct nucleotide sequences were sampled into each of 96 "wells", with sampling probabilities equal to sequencing read frequencies. This sampling procedure was repeated 200 times, and a clone was considered "paired" if its beta chain occurred in between 5 and 91 wells in at least 190 repetitions (95 % of repetitions). The bounds of 5 and 91

correspond to p-values of 1.636×10^{-8} by the Fisher Exact test, in the ideal case where all chains present are detected.

Density vs. chain frequency for the entire repertoire is shown in Figure 2.2. The inset illustrates the fraction (pink) of the repertoire that was paired using an N_{well} of a. 1000, b. 2000, and c. 5000. The mean number of pairs detected using each N_{well} is shown in Figure 2.3. Using fewer cells per well yields fewer pairs, but allows for pairing of higher frequency clones. By using a set of plates, each with different N_{well}, a large range of clonal frequencies can be investigated.

Theoretically, the N_{well} can be increased to allow for pairing of even the lowest frequency clones. In practice, however, the upper limit of N_{well} is determined by sample availability and sequencing depth (e.g. with 10000 cells in each of 96 wells, each alpha and beta chain will be represented by only ~13 reads using an Illumina MiSeq V3 kit). Instead of using large N_{well}, a fraction of low frequency clones can be paired by polyclonal expansion of the starting pool; by expanding we create additional "copies" of all clones ($2^{\#\text{ of divisions}}$), so that, upon splitting of the population, the low-frequency clones appear in enough wells to statistically extract the pairings.

While generating lists of pairings from an entire repertoire is very valuable, most interest lies in extracting paired sequences from functional or reactive clones. This can be accomplished by performing an antigen specific activation before (i) or after (ii) splitting among a number of wells; a clone appearing at higher frequency (i. % of wells, ii. % of sequencing reads in a well) than expected, given the frequency of that clone in the original population, is classified as reactive. The occurrence of this clone across multiple wells allows for extraction of paired-chain information.

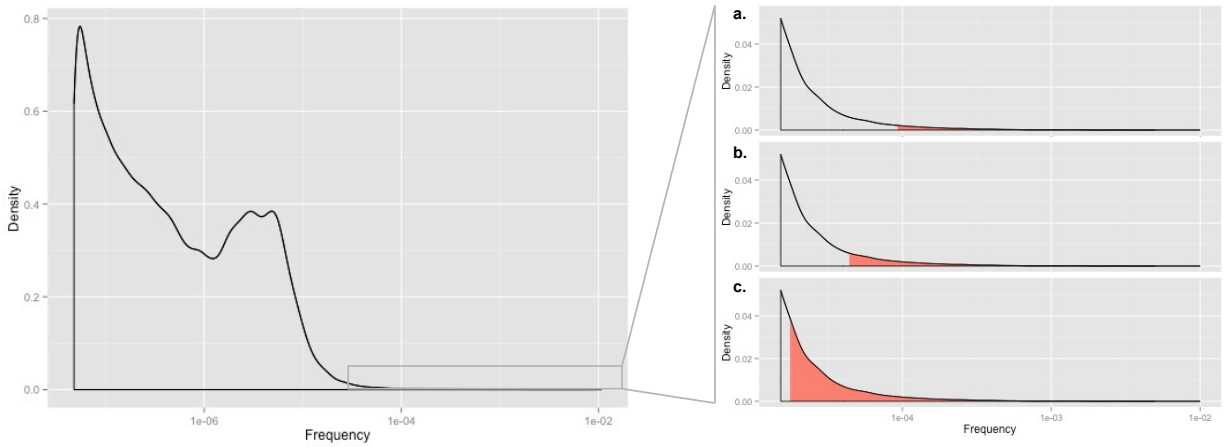


Figure 2.2: Density vs. chain frequency for a healthy male β-chain repertoire ([13]male 1)

Pink shading indicates the chains paired with (a) 1000, (b) 2000, (c) 5000 cells in each of 96 wells (by simulation). With more cells per well, a greater fraction of lower frequency clones are paired, although all three ranges contain very little of the repertoire density.

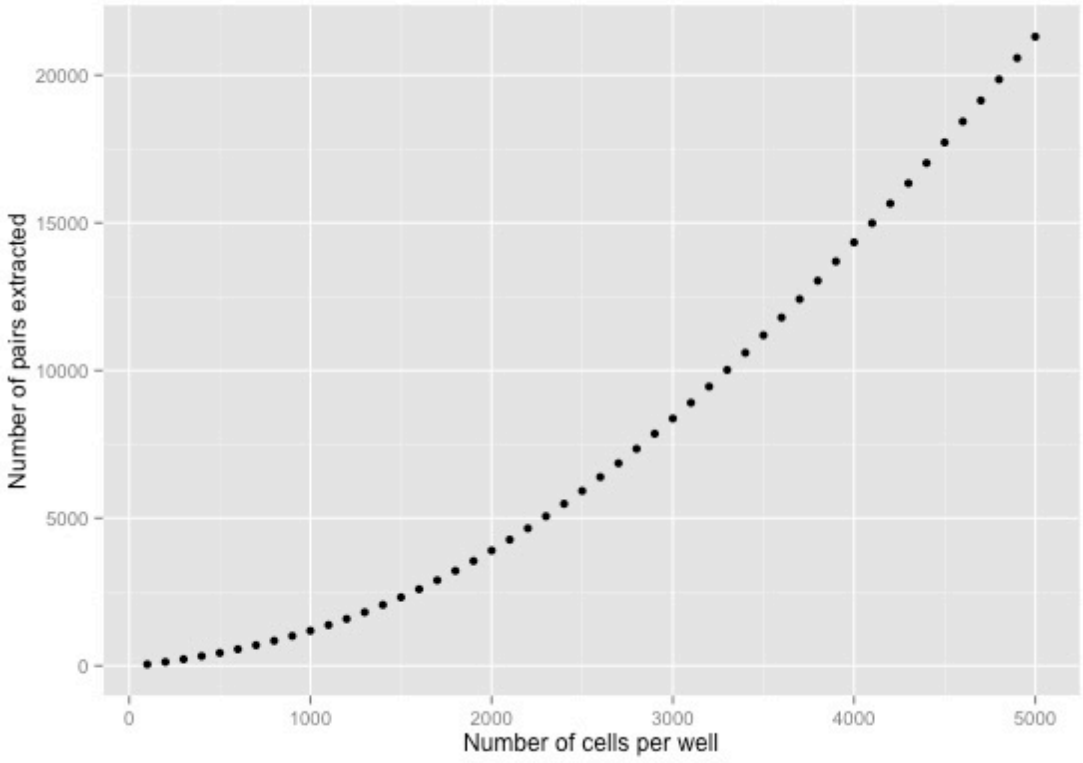


Figure 2.3: Number of pairs extracted vs. number of cells per well (by simulation)

2.4 Library Preparation

A template switching and dual indexed PCR approach was chosen for amplification of TCR/BCR transcripts. Figure 2.4 is a schematic of the approach. Reverse transcription is performed using an M-MLV-derived reverse transcriptase from Clontech (SMARTScribe), and a set of constant-region primers. A universal priming site with a sequence complementary to the template switching oligo is appended. The forward PCR primer consists of the template switching oligo sequence, a column-specific 6-bp index, and the Illumina P5 sequencing adapter. The reverse PCR primers are generated in each PCR reaction from a set of gene-specific forward primers with universal tail, and a single reverse primer consisting of the complement of the universal tail, a row-specific 6-bp index, and the Illumina P7 sequencing adapter. This is hereafter referred to as the primer-extension strategy. The gene-specific segments of the full-length reverse PCR primers are designed to anneal at regions "stepped-in" from the reverse transcription priming sites, to increase specificity. The inclusion of P5 and P7 sequencing adapters in the PCR primers enables direct sequencing of the final construct on an Illumina MiSeq with 2 × 300-bp reads.

Because only one moderate length reverse primer is required per reaction, instead of a number of longer (i.e. more expensive to synthesize) reverse primers, one for each gene, this strategy can be easily applied to study any gene pairing of interest. In addition, since our pairing method requires only that each well (i.e. ~1000 cells) be uniquely indexed, the cost of indexing oligonucleotides is not prohibitive as it can be for pairing approaches relying on indexing of single-cells. Finally, cross-contamination of indexes is not as problematic with this method due to the statistical nature of the pairing.

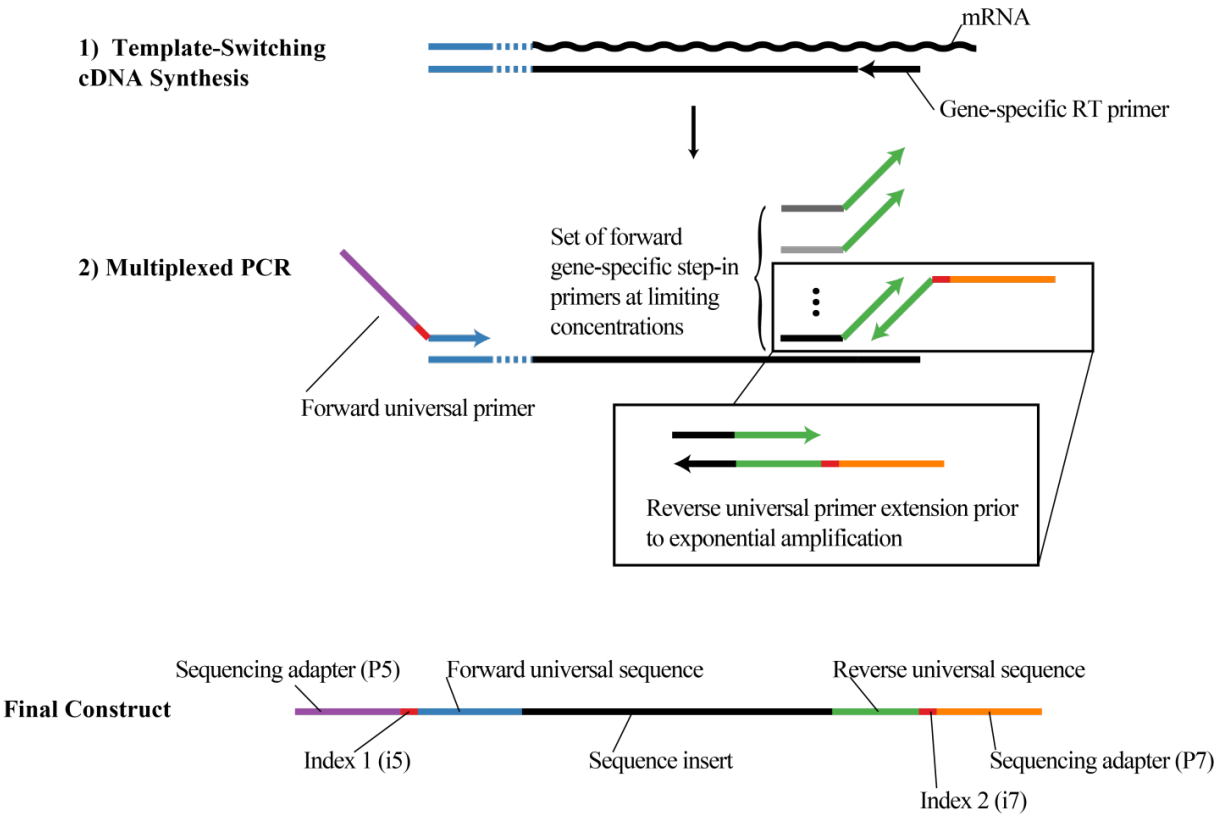


Figure 2.4: Library preparation schematic

mRNA is reverse transcribed using gene specific primers and an M-MLV reverse transcriptase. A universal primer site is added by template switching. cDNA is then amplified using a universal forward primer and a set of reverse primers which are generated in the reaction by primer extension.

Chapter 3: Demonstration

3.1 Preliminary Validation

All method validation was performed using primers specific for T cell receptor alpha and beta chains (see Methods for sequences). Total RNA, purified from the Jurkat leukaemic T cell line (Ambion), was used as template in three initial experiments. Resulting PCR products were visualized on 2 % agarose gels (E-gel EX, Invitrogen), with TrackIt 100-bp DNA ladder (Invitrogen).

First, to limit complexity, the protocol was performed using PCR primers lacking indexes and flow cell adapters. Alpha chain and beta chain PCRs were carried out separately. Bands of the expected size (~750 bp for alpha and ~550 bp for beta) were observed for 1000 ng, 300 ng, and 30 ng RNA inputs (Figure 3.1). Sanger sequencing of the bands confirmed alignment to Jurkat alpha and beta chain sequences.

Next, the protocol was performed using a full-length forward primer and full-length reverse primers. This enabled testing of the PCR without concern for side products or inefficiencies resulting from the primer-extension strategy. Again, alpha chain and beta chain PCRs were carried out separately. The PCR products were ~100 bp longer than in the previous experiment, as expected (Figure 3.2).

Finally, the full primer-extension protocol was performed on 300 ng Jurkat RNA, using 500 nM reverse primer, and gene specific forward primer at concentrations from 0.78 nM to 500 nM. Alpha and beta chain transcripts were successfully amplified in the 6.25 nM- 50 nM primer concentration range (Figure 3.3). Because of the appearance of a ~150-bp side product at lower primer concentrations, 25 nM was chosen for all future experiments.

To explore effects of cell lysate, the protocol was performed on Jurkat cells. RT was carried out on 0 (PBS control), 30, 300, or 3000 Jurkat cells. 1 µl or 4 µl of RT product was then used as template for a 25 µl PCR. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and visualized on an Agilent 2100 Bioanalyzer, using the Agilent High Sensitivity DNA Kit. The Bioanalyzer profile for 300 cells, with 4 µl into PCR is given in Figure 3.4 a. While there is a substantial amount of 160-bp side product, alpha (843 bp) and beta (652 bp) peaks are also clearly visible. Figure 3.4 b compares different conditions in the region of interest. As expected, we observed more PCR product using 300 cells than 30 cells, and, for both 300 and 30 cell inputs, more PCR product using 4 µl RT product than 1 µl RT product. For 3000 cells, we did not observe these same trends, however. With 1 µl in PCR, we observed less product than with 300 cells/4 µl, and with 4 µl, less still, close to the level of 300 cells/1 µl. If the sole issue was cell lysate inhibiting RT, we would expect less PCR product overall,

but still four times as much PCR product from the reaction with 4 µl RT product over that with 1 µl. Instead, it appears that the lysate is inhibiting PCR.

For the first attempt of the full protocol (below), only 300 primary human cells were used per reaction to avoid lysate inhibition.

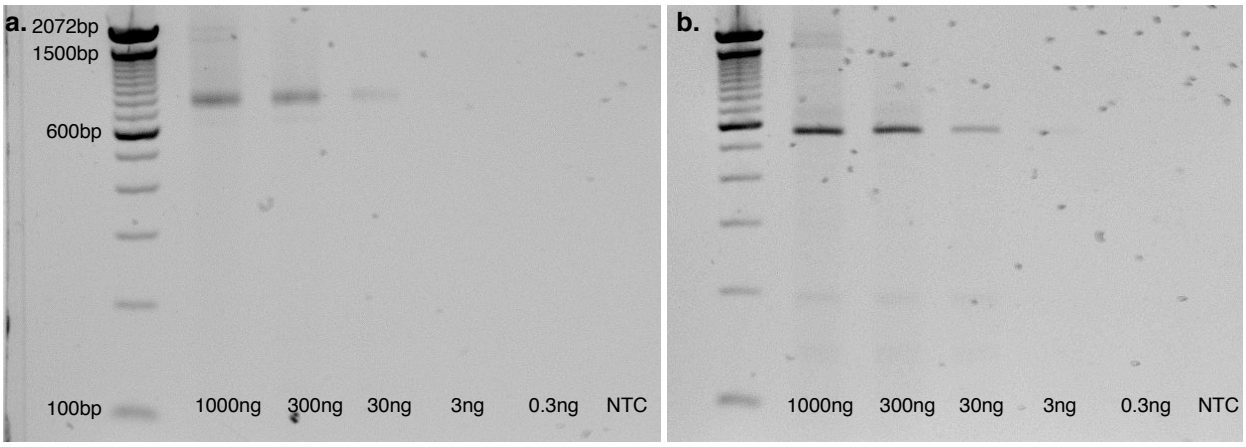


Figure 3.1: Preliminary validation experiment 1

0.3 ng-1000 ng Jurkat RNA, or water (NTC) input, PCR performed using primers lacking indexes and flow-cell adapters (a) PCR with reverse primer specific to α-chain constant region. Bands at expected size (~750 bp) for 1000 ng, 300 ng, and 30 ng inputs (b) PCR with reverse primer specific to β-chain constant region. Bands at expected size (~550 bp) for 1000 ng, 300 ng, and 30 ng inputs.

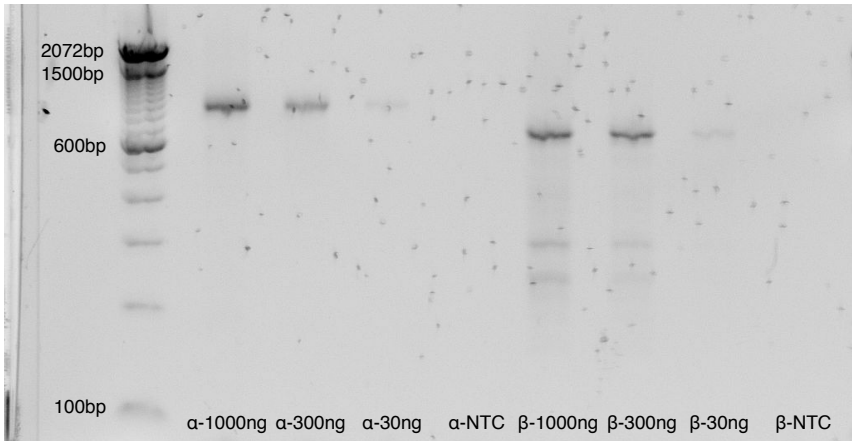


Figure 3.2: Preliminary validation experiment 2

30 ng-1000 ng Jurkat RNA or water (NTC) input, PCR performed using full-length forward primer (TCRF-GSCX) and full-length reverse primers (TCRA-GSCZ or TCRB-GSCZ). Bands at ~850 bp for α-chain PCR, and ~650 bp for β-chain PCR, ~100 bp longer than in preliminary validation experiment 1, as expected.

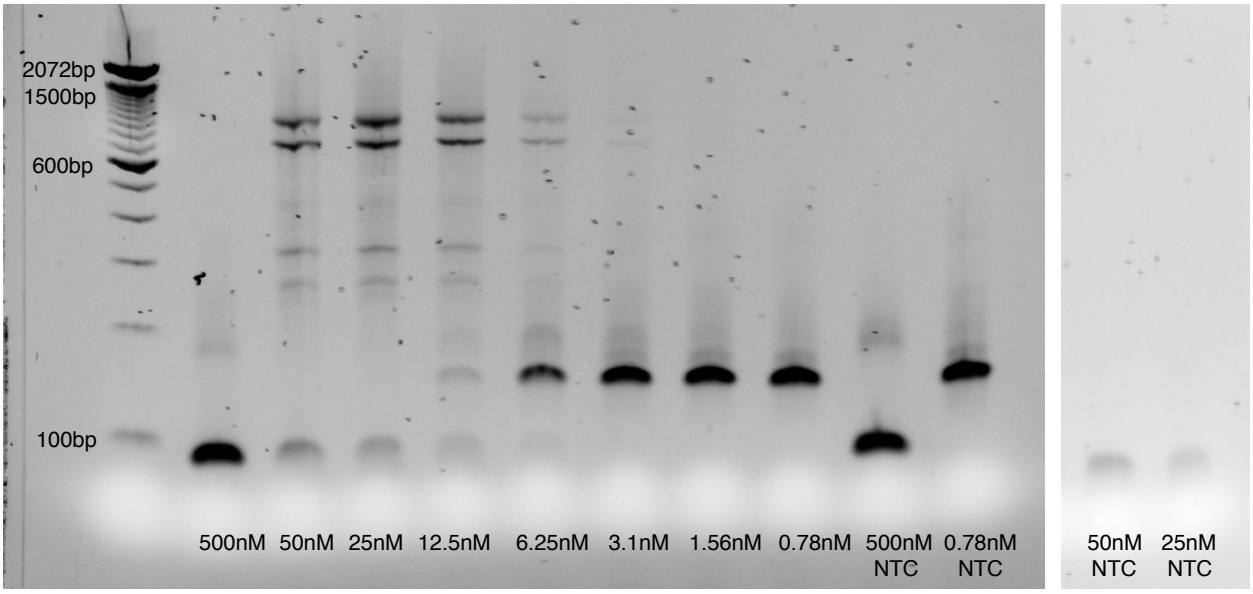


Figure 3.3: Preliminary validation experiment 3

300 ng Jurkat RNA or water (NTC) input, PCR performed using 500 nM full-length forward primer (TCRF-GSCX), 500 nM reverse primer (TruSeq-GSCY), and gene specific forward primers (Alpha SI-UT and Beta SI-UT) at concentrations ranging from 0.78 nM to 500 nM. Bands at ~850 bp and ~650 bp indicate success of the primer-extension strategy with Alpha SI-UT and Beta SI-UT concentrations between 6.25 nM and 50 nM. A decrease in Alpha SI-UT and Beta SI-UT primer concentration from 12.5 nM is accompanied by a decrease in specific product, and an increase in ~150 bp side product. 25 nM Alpha SI-UT and Beta SI-UT was used in all subsequent experiments.

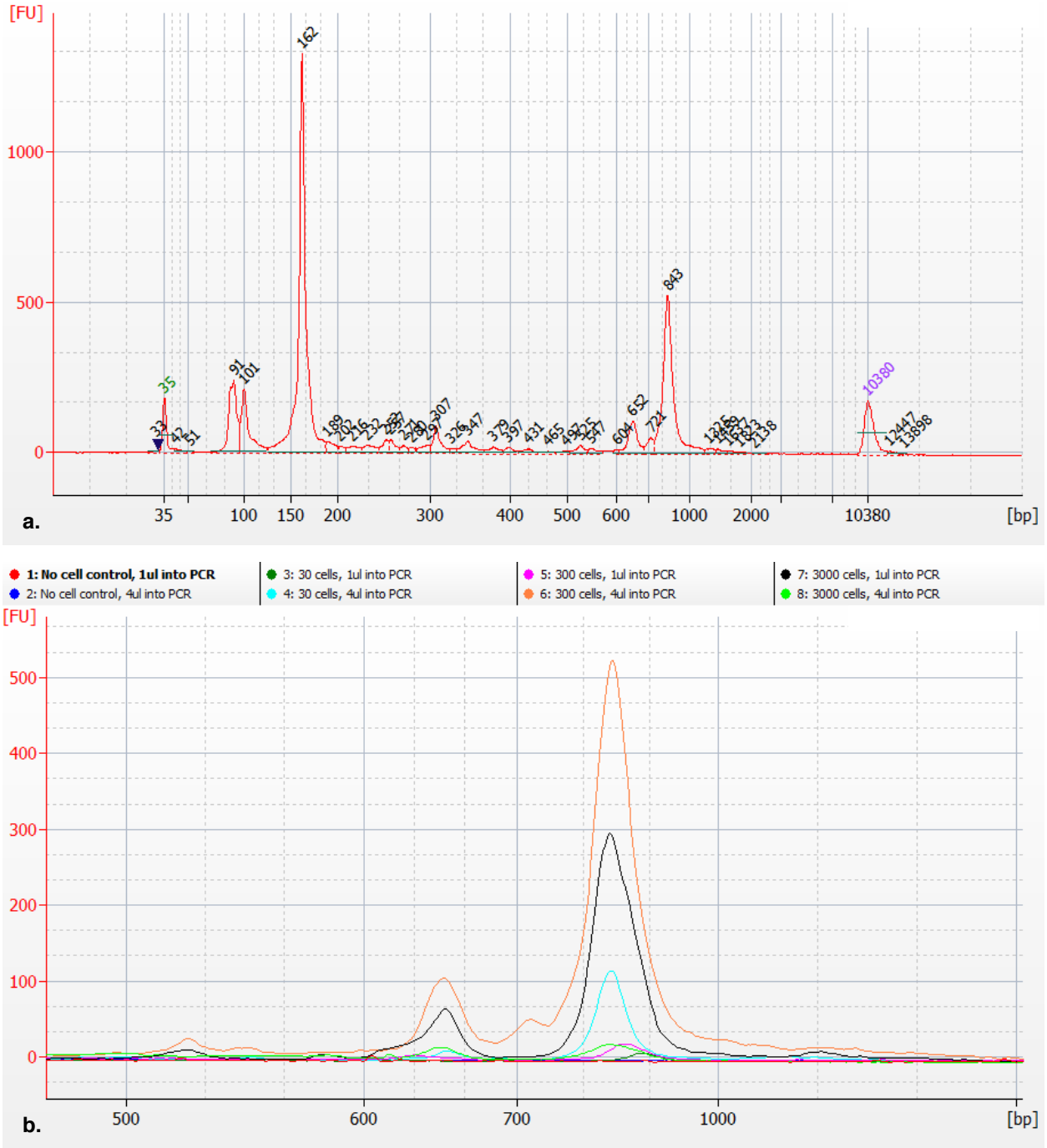


Figure 3.4: Preliminary validation experiment 4

(a) 300 Jurkat cells input, 4 µl RT product into PCR (b) 0-3000 Jurkat cells input, 1 µl or 4 µl of RT product into PCR. See main text.

3.2 Full Protocol- Attempt 1

Given the success of my library preparation method with purified Jurkat RNA and Jurkat cells, I proceeded to test it on a full plate of primary human T cells. 10000 $CD8^+$ T cells, isolated from the blood of a healthy female, were activated using CD3/CD28 Dynabeads. Microscope images of the cells and control four days after the start of activation are given in Figure 3.5. After seven days of activation (~150000 cells), the activation beads were removed and FACS was used to deposit 300 cells into each of 96 wells containing 2.5 µl of lysis mix. This plate was stored at -80 °C until use.

Heat lysis and reverse transcription were performed on the above samples. RT product from one well was first used to determine the optimum dilution into PCR; as with Jurkat cells, 4 µl of RT product in a 25 µl PCR reaction yielded sufficient product (Figure 3.6). Based on this result, the full plate of PCR was therefore performed using 4 µl RT product. PCR products were then pooled and purified. The final library (Figure 3.7), spiked with 10.6 % Phi X, was sequenced on an Illumina MiSeq with 2 × 300-bp reads.

Of 19.6 million total reads, 17.5 million (89.3%) were identified as alpha chain sequences, and 2.0 million (10.3%) were identified as beta chain sequences. Altogether, 996 unique alpha chains and 539 unique beta chains were extracted. From the previously described simulation, with 300 cells per well, we would expect approximately 23000 unique beta chains in the 96-well plate (mean from 200 repetitions). Although a better estimate could be made with bulk repertoire sequencing data from the same individual, the sensitivity is clearly very low.

Because the alpha chain does not show allelic exclusion, it is not surprising that we observe more unique alpha chains than beta chains. Allelic exclusion cannot explain an almost 9× difference in the number of alpha and beta reads, however; this is likely due to differences in RT and PCR efficiencies.

The statistical chain-pairing pipeline was run, with a chain considered "present" in a well if it appeared at more than 10 reads. The computed p-values are shown in Figure 3.8. Only a single p-value was below 1×10^{-4}. For this putative pair (p-value= 7.46×10^{-6}), alpha occurs without beta once, beta occurs without alpha once, and they occur together four times. The failure of the method to extract pairs was due to inadequate sensitivity, and the resulting inability to detect the same chain in enough wells.

Because final yield was more than sufficient for sequencing, but library diversity was low, we suspected reverse transcription to be the sensitivity-limiting step. The following section describes reverse transcription protocol optimizations, specifically, effects of oligonucleotide design, buffer composition, and sample preparation.

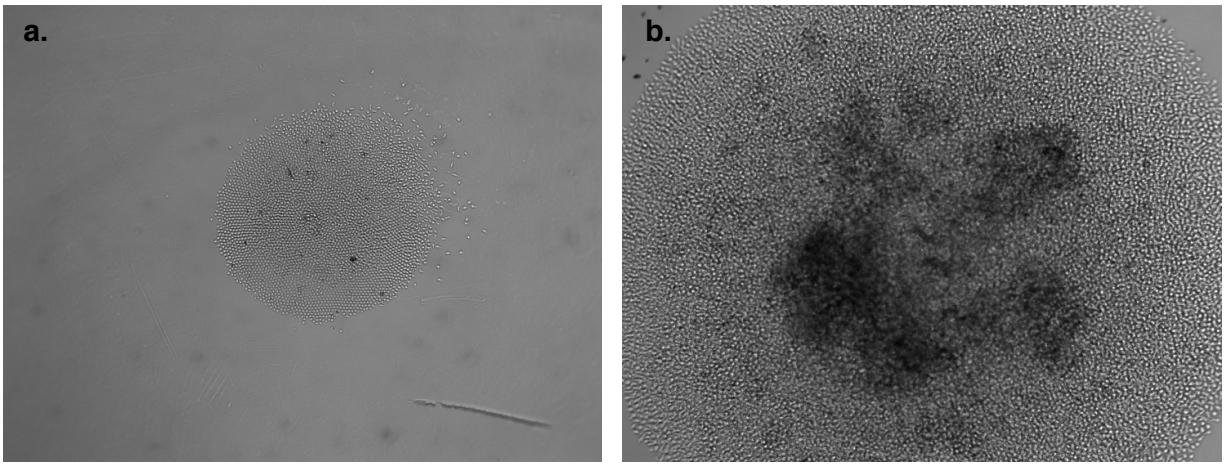
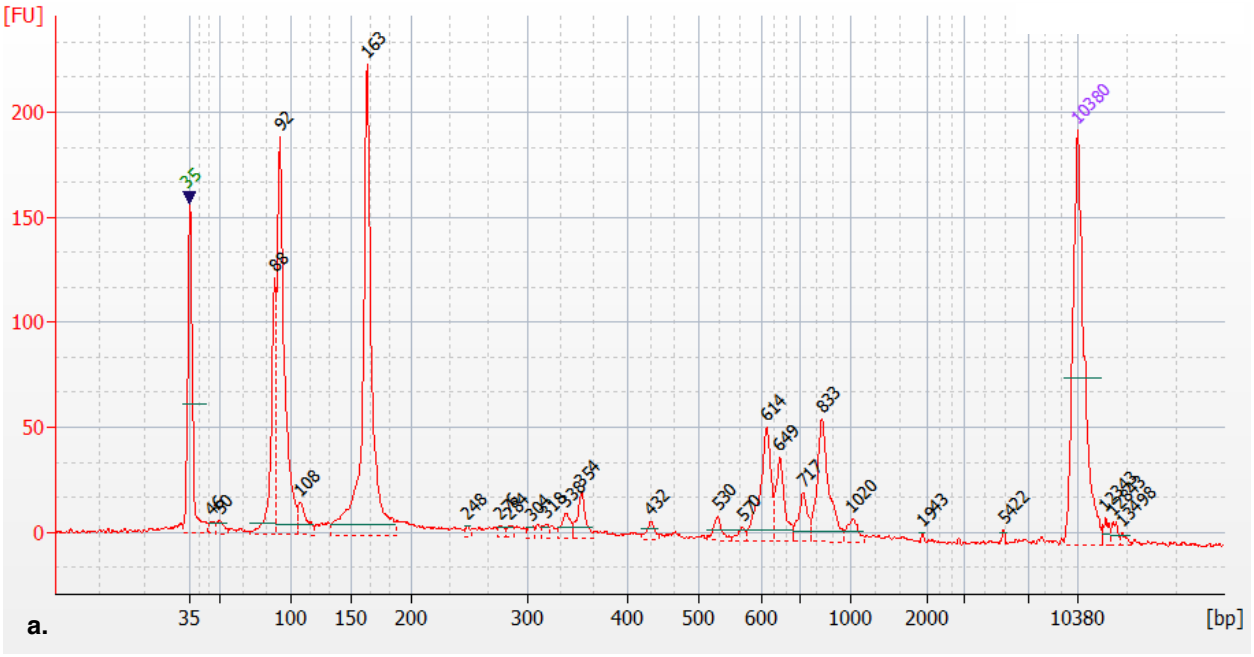


Figure 3.5: Optical microscope image of ~10000 $CD8^+$ T cells

(a) after four days of culture (b) after four days of culture in the presence of Dynabeads Human T-Activator CD3/CD28. Dark areas are clusters of beads and activated cells.



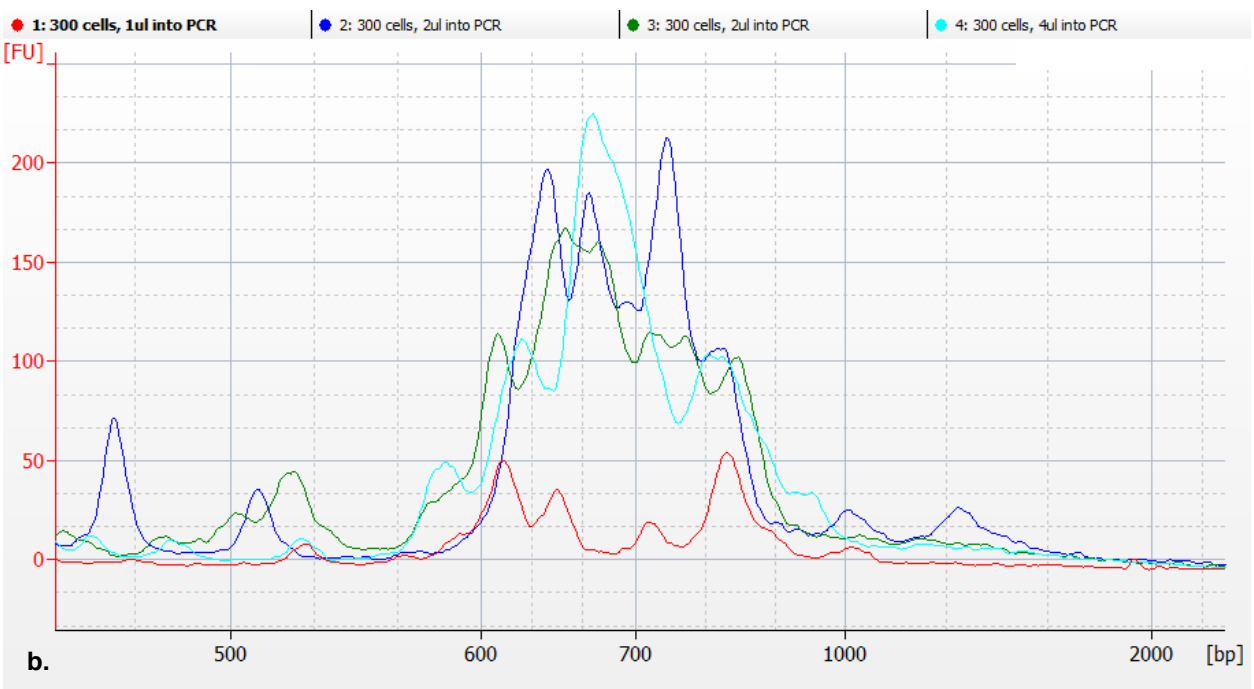


Figure 3.6: Full protocol attempt 1

PCR performed on RT product from a single well containing 300 primary human $CD8^+$ T cells (a) 1 µl RT product into PCR (b) 1 µl- 4 µl RT product into PCR. 4 µl resulted in good yield of specific product (~600 bp to ~900 bp) and so was used for PCR of the remaining wells.

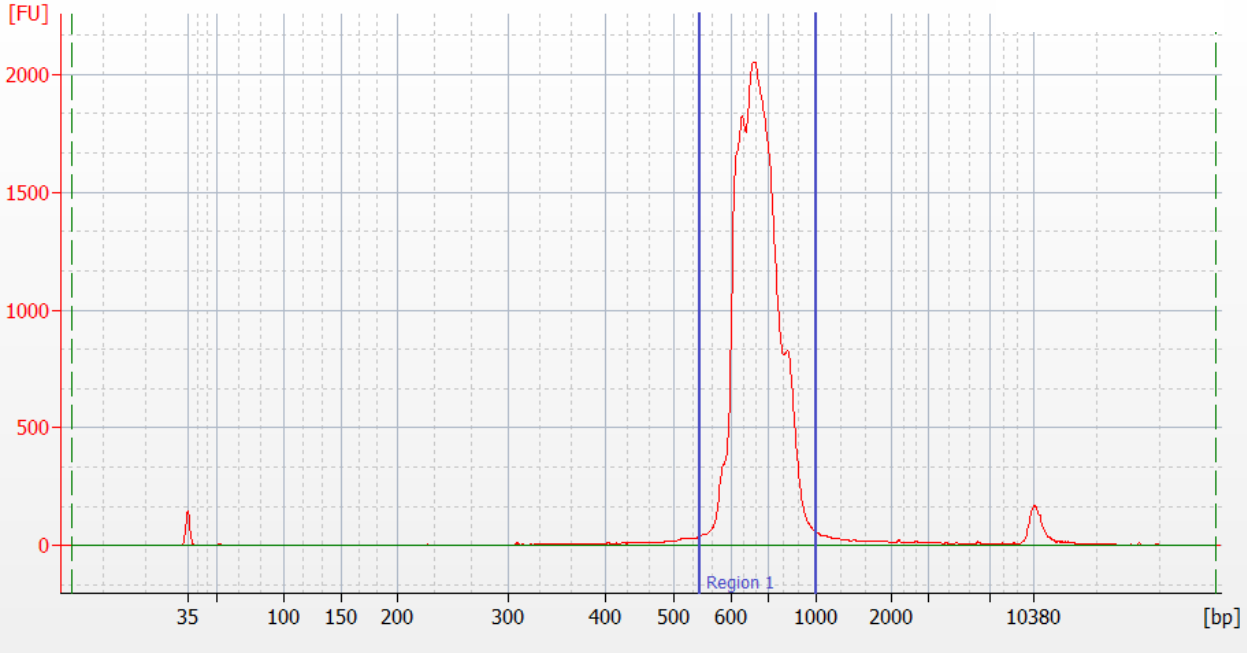


Figure 3.7: Full protocol attempt 1, final library

Region 1: 545-1000 bp, average size= 689 bp

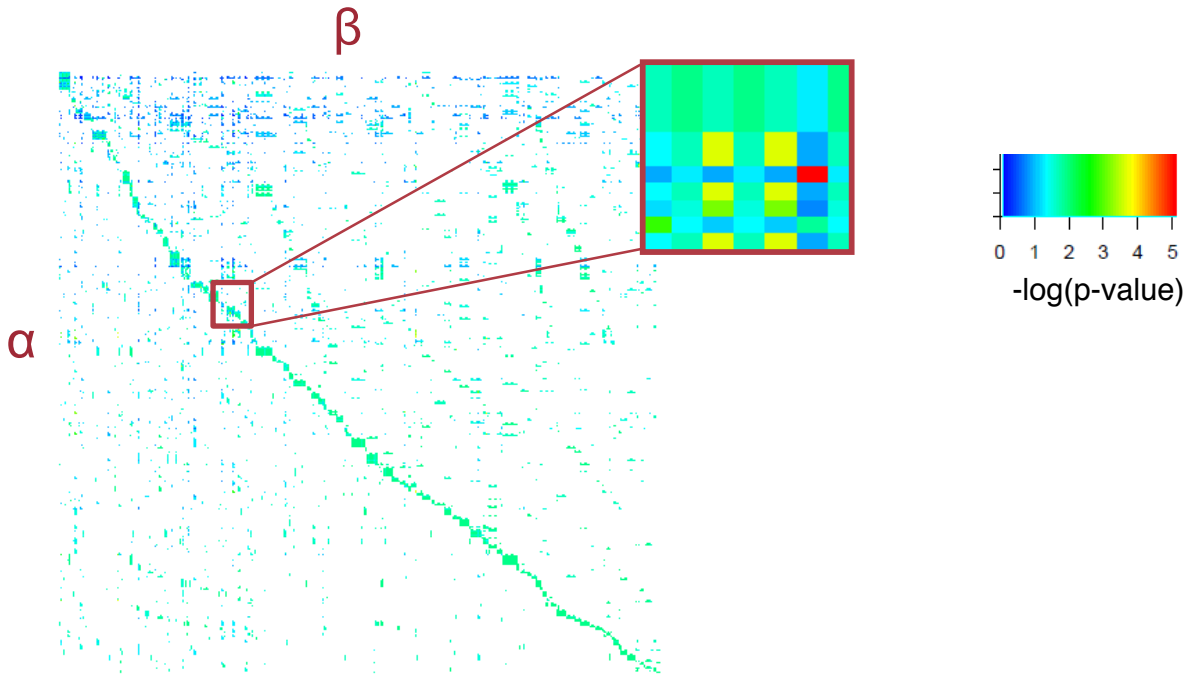


Figure 3.8: Full protocol attempt 1, heatmap of p-values

Each row corresponds to a unique alpha chain, and each column corresponds to a unique beta chain. Inset shows the one significant p-value extracted (p-value= 7.46×10^{-6}). The green diagonal is an artifact of the alpha and beta chain ordering, and corresponds to a single co-occurrence of an alpha and beta chain (p-value= 0.0104).

3.3 Optimizations

3.3.1 Reverse Transcription Primers

I first tested blocked RT primers, with the hypothesis that I could increase yield of desired product by limiting the generation of side products. Alpha and beta RT primers containing 5' biotin blocks were compared to standard RT primers (beta C gene primer C6 and alpha C gene primer TCRA-5). For both Jurkat RNA and Jurkat cell inputs, substantially more PCR product was produced using blocked RT primers (Figure 3.9), and so blocked RT primers were used in all subsequent experiments.

3.3.2 Template Switching Oligonucleotide

I then investigated increasing sensitivity through improved design of the TSO. TCR-TSv2, the TSO used in all previous experiments, contains only three RNA bases, and a single block, a C3 on the 3' end. I designed a new TSO construct (TCR-TSv3) that incorporated a number of features reported to increase yield. TCR-TSv3 is an all RNA oligo [27], apart from a single locked nucleic acid (LNA) guanylate [21, 22], blocked on the 5'-end with biotin, and on the 3'-end with C3 [20]. The library preparation protocol was performed on Jurkat cells and Jurkat RNA, using either TCR-TSv3 or TCR-TSv2. While TCR-TSv3 gives slightly more product with Jurkat RNA, TCR-TSv2 yields more product with Jurkat cell inputs (Figure 3.10). Because TCR-TSv3 offered no significant improvement, I continued to use the simpler TCR-TSv2.

3.3.3 Reverse Transcription Buffer

I next evaluated the effect of varying buffer composition in the reverse transcription reaction. The presence of betaine in reverse transcription reactions has been shown to increase cDNA yield, both by reducing RNA secondary structure and by stabilizing proteins (i.e. a higher melt temperature can be used without denaturing the reverse transcriptase) [21, 22]. The addition of betaine requires adjustment of the magnesium chloride concentration [21, 22]. Picelli *et al.* found 1 M Betaine and 9 mM $MgCl_2$ to be optimal for single-cell whole transcriptome analysis [21, 22].

All previous experiments were performed with 0.5 M betaine, and 15 mM $MgCl_2$. For optimization, the protocol was performed on 2 ng Jurkat RNA inputs, with betaine concentrations ranging from 0 M-1 M, $MgCl_2$ concentrations ranging from 6 mM to 18 mM, and full length reverse PCR primers (no primer-extension). Because SMARTScribe 5x Buffer contains 30 mM MgCl2, no additional $MgCl_2$ was added to achieve a 6 mM final $MgCl_2$ concentration.

The most successful condition, 1 M Betaine, with 12 mM $MgCl_2$, resulted in a drastic improvement in yield over previous RT conditions (Figure 3.11), and so was used in all future experiments.

3.3.4 RNA Purification

Finally, I investigated if RNA purification could be used to enhance the sensitivity and specificity of our protocol. By using only a fraction of the total cDNA in each PCR reaction, we are imposing an upper bound on sensitivity. Instead of diluting out PCR-inhibiting cell lysate, we can eliminate it entirely by purifying RNA from the cells in each well. Not only does RNA purification allow more RT product to be used in the PCR, it also allows for higher numbers of cells per well, thereby increasing the achievable throughput to levels limited only by sequencing depth.

RNA purification was first attempted on replicates of 1000 Jurkat cells/tube using TRIzol Reagent (Invitrogen). RNA was precipitated using isopropanol, resuspended in 4 µl of 0.2 % Triton X-100, 2 U/µl RNAse inhibitor, 0.1 % Tween 20 (lysis mix minus primers and dNTPs), and then heated at 72 °C for 3 min. Samples were visualized on an Agilent 2100 Bioanalyzer, using the Agilent RNA 6000 Pico Kit. RNA Integrity Numbers (RIN) were close to 10, indicating high quality RNA (Replicate 1: Concentration=953 pg/µl, RIN=9.7; Replicate 2: Concentration=833 pg/µl, RIN=10; Replicate 3: Concentration=995 pg/µl, RIN=9.9). For comparison, RINs for replicates of 1000 pg/µl Jurkat RNA purchased from Ambion were 9.9 and 9.7.

Next, the effect of RNA purification on the full library preparation protocol was examined using primary human cells. $CD8^+$ T cells were isolated from peripheral blood of a healthy male, collected 22 days after live attenuated influenza immunization. 10000 cells were activated using CD3/CD28 Dynabeads, and total RNA was isolated from the remaining cells using *mir*Vana™ miRNA Isolation Kit (Ambion).

After 6 days of activation, TRIzol purification was performed on 10, 100, 500, and 1000 cells. Reverse transcription was carried out on these samples, as well as parallel 10, 100, 500, and 1000 cells samples that were not TRIzol purified, and 0.01 ng, 0.1 ng, 1 ng, and 10 ng amounts of primary cell RNA. 5 µl, 2.5 µl, or 1 µl of RT product was used as template for PCR reactions. For all cell samples, PCR amplification was performed using full-length reverse PCR primers. For purified RNA samples, both the full-length reverse PCR primers and the primer-extension strategy were used.

Little or no product was observed on the Bionanalyzer for non-purified cell samples with 5 µl in PCR, and so these samples were not carried through further analysis. PCR products were subject to two rounds of bead purification and then select samples were run on the Bioanalyzer to ensure adequate

removal of short side products. DNA concentrations were determined with a Qubit fluorometer, using the Qubit dsDNA HS Assay Kit (Table 1). Trizol purification not only saved 1000 cell samples from lysate inhibition; it dramatically increased yield for all cell occupancies and dilutions.

Different sample volumes, given in Table 1, were combined to create a normalized pool. The final library, spiked with 7 % Phi X, was sequenced on an Illumina MiSeq with 2 × 300-bp reads. The percentage of reads mapping to each sample is given in Table 1. The low percentages for samples that were not included in the pool indicate negligible cross-contamination of indexes.

Sequencing data was visualized in two ways. First, the chains identified in each sample were numbered by decreasing frequency, and frequency versus chain number was plotted. Across different samples, a certain chain number does not refer to the same chain. Second, plots were generated to show correlation between reaction A and reaction B; a dot represents a single chain, and x and y coordinates are, respectively, the frequency of that chain in reaction B, and the frequency of same chain in reaction A. Relevant plots are shown for the alpha chain in Figure 3.12 and for the beta chain in Figure 3.13.

With 2.5 µl in PCR, many more unique chains were extracted using TRIzol purification (Figure 3.12 and Figure 3.13, a and b). For example, with the 1000-cell sample, the number of chains present at above 10 sequencing reads increased from 94 alpha and 23 beta to 629 alpha and 422 beta with TRIzol purification. Among TRIzol purified samples, those using 5 µl of RT product in PCR had the most chains (e.g. 749 alpha and 591 beta chains for 1000 cell sample) (Figure 3.12 and Figure 3.13, c). The fact that more beta chains were extracted from a single 1000-cell reaction, than from the entire plate of 300 cells/well in the first attempt, highlights the remarkable improvement in sensitivity achieved by optimization. Correlation between 1000 cell, TRIzol purified 2.5 µl and 5 µl reactions was strong above single-cell frequency (10^{-3}), demonstrating the reproducibility of the PCR (Figure 3.12 and Figure 3.13, d). It is intriguing that, for 2.5 µl in PCR with TRIzol purification (Figure 3.12 and Figure 3.13, a), plateaus in frequency appear at roughly single-cell frequencies i.e. at log(1/1000)=-3 for the 1000 cell sample, log(1/500)=-2.7 for the 500 cell sample, and log(1/100)=-2 for the 100 cell sample. The primer-extension strategy proved to be as efficient as the full-length reverse primers in extracting chains from purified RNA (Figure 3.12 and Figure 3.13, e and f). Overall, these results demonstrate that TRIzol purification greatly improves sensitivity.

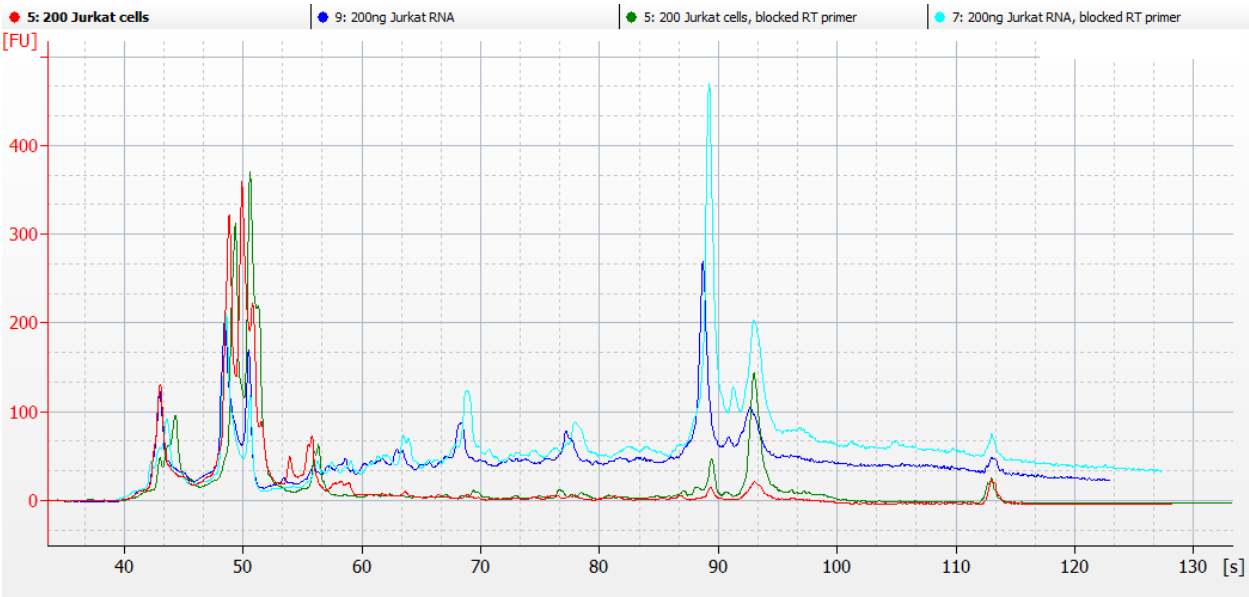


Figure 3.9: Optimization 1: reverse transcription primers

200 ng Jurkat RNA or 200 Jurkat cells input, TCR-TSv2 template switching oligo, comparison of 5'-blocked beta C gene primer C6 and alpha C gene primer TCRA-5 with unblocked beta C gene primer C6 and alpha C gene primer TCRA-5

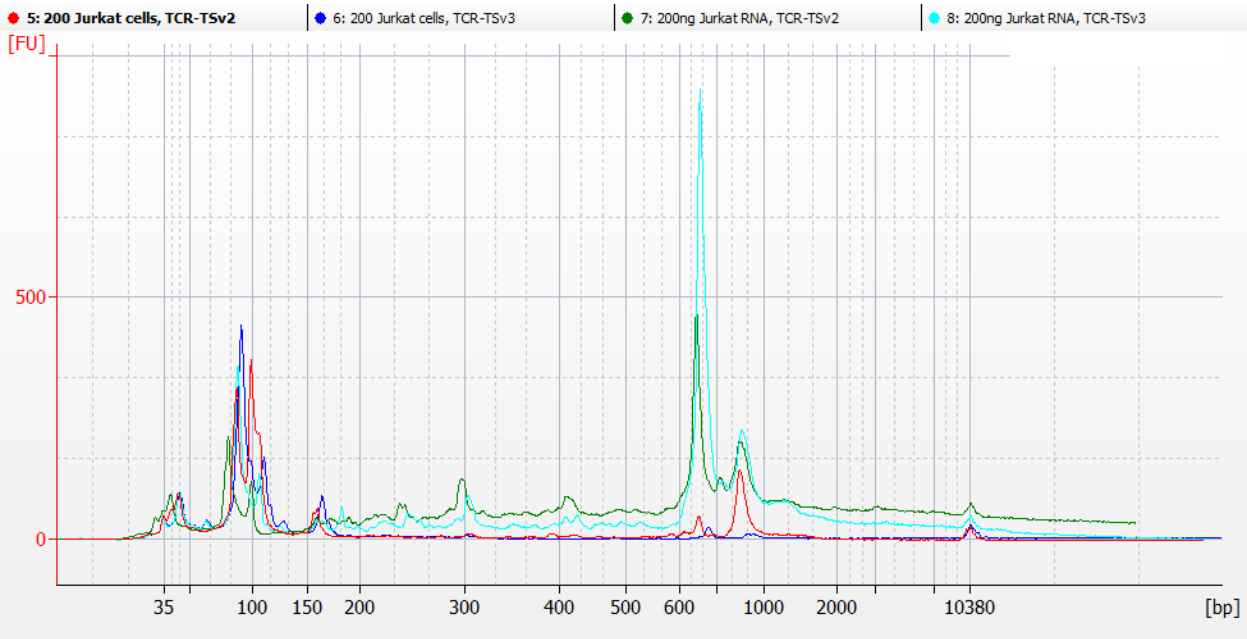


Figure 3.10: Optimization 2: template switching oligonucleotide

200 ng Jurkat RNA or 200 Jurkat cells input, 5'-blocked beta C gene primer C6 and alpha C gene primer TCRA-5, comparison of TCR-TSv2 and TCR-TSv3 template switching oligos (sequences in Methods)

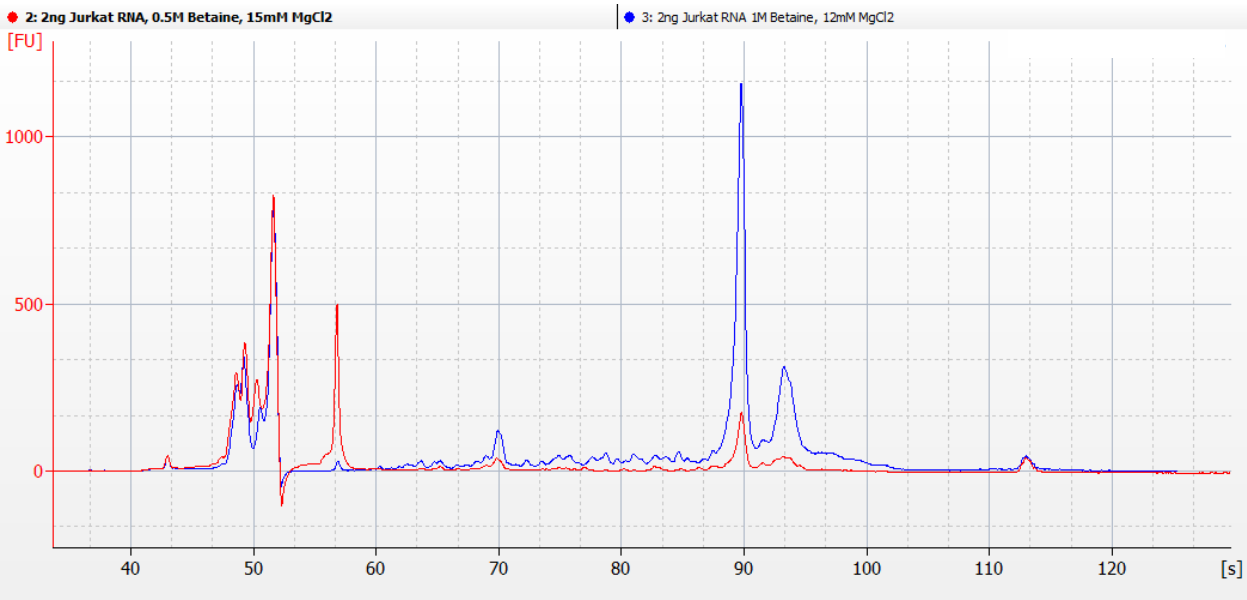


Figure 3.11: Optimization 3: reverse transcription buffer

2 ng Jurkat RNA input, 5'-blocked beta C gene primer C6 and alpha C gene primer TCRA-5, TCR-TSv2 template switching oligo, comparison of pre- (0.5 M betaine, 15 mM $MgCl_2$) and post-optimization (1 M betaine, 12 mM $MgCl_2$) concentrations in reverse transcription reaction

Table 1: Optimization 4: RNA Purification- concentrations, volumes into pool, percent sequencing reads

Input Amount	RNA Purification	Volume into PCR	PCR Primer Strategy	Concentration (µg/ml)	Volume into pool (µl)	% reads
10ng	n/a	5 µl	full length primers	2.26	2.21	2.7618
1ng	n/a	5 µl	full length primers	0.723	6.92	2.6078
0.1ng	n/a	5 µl	full length primers	0.083	0.00	0.0031
0.01ng	n/a	5 µl	full length primers	0	0.00	0
10ng	n/a	2.5 µl	full length primers	2.69	1.86	4.1915
1ng	n/a	2.5 µl	full length primers	0.633	7.90	3.1555
0.1ng	n/a	2.5 µl	full length primers	0.061	0.00	0.0047
0.01ng	n/a	2.5 µl	full length primers	0	0.00	0.0001
10ng	n/a	1 µl	full length primers	1.2	4.17	4.8592
1ng	n/a	1 µl	full length primers	0.208	24.06	2.6932
0.1ng	n/a	1 µl	full length primers	0	0.00	0.004
0.01ng	n/a	1 µl	full length primers	0	0.00	0.0001
1000	TRIzol	5 µl	full length primers	7.01	0.71	2.2034
500	TRIzol	5 µl	full length primers	3.45	1.45	2.9843
100	TRIzol	5 µl	full length primers	3.1	1.61	2.57
10	TRIzol	5 µl	full length primers	0.763	6.56	0.0792
1000	n/a	2.5 µl	full length primers	1.24	4.04	2.5677
500	n/a	2.5 µl	full length primers	1.56	3.21	3.2882
100	n/a	2.5 µl	full length primers	0.663	7.55	3.3537
10	n/a	2.5 µl	full length primers	0.145	0.00	0.0011
1000	TRIzol	2.5 µl	full length primers	9.74	0.51	3.4235
500	TRIzol	2.5 µl	full length primers	4.22	1.19	4.811
100	TRIzol	2.5 µl	full length primers	2.67	1.87	3.9478
10	TRIzol	2.5 µl	full length primers	0.345	14.50	0.1152
1000	n/a	1 µl	full length primers	0.111	0.00	0.0053
500	n/a	1 µl	full length primers	0.584	8.57	3.7053
100	n/a	1 µl	full length primers	0.359	13.94	3.1712
10	n/a	1 µl	full length primers	0	0.00	0.001
1000	TRIzol	1 µl	full length primers	5.84	0.86	5.0834
500	TRIzol	1 µl	full length primers	1.79	2.80	5.271
100	TRIzol	1 µl	full length primers	1.12	4.47	4.0818
10	TRIzol	1 µl	full length primers	0	0.00	0.0002
10ng	n/a	5 µl	primer-extension	4.92	1.02	1.6347
1ng	n/a	5 µl	primer-extension	3.04	1.65	1.5044
0.1ng	n/a	5 µl	primer-extension	0.393	12.73	0.7984
0.01ng	n/a	5 µl	primer-extension	0.16	0.00	0.0006
10ng	n/a	2.5 µl	primer-extension	5.08	0.99	3.2709
1ng	n/a	2.5 µl	primer-extension	1.94	2.58	2.3295
0.1ng	n/a	2.5 µl	primer-extension	0.377	13.27	0.7816
0.01ng	n/a	2.5 µl	primer-extension	0.214	23.38	0.0859
10ng	n/a	1 µl	primer-extension	2.06	2.43	0.0008
1ng	n/a	1 µl	primer-extension	0.416	12.03	0.0007
0.1ng	n/a	1 µl	primer-extension	0.153	0.00	0.0001
0.01ng	n/a	1 µl	primer-extension	0.142	0.00	0

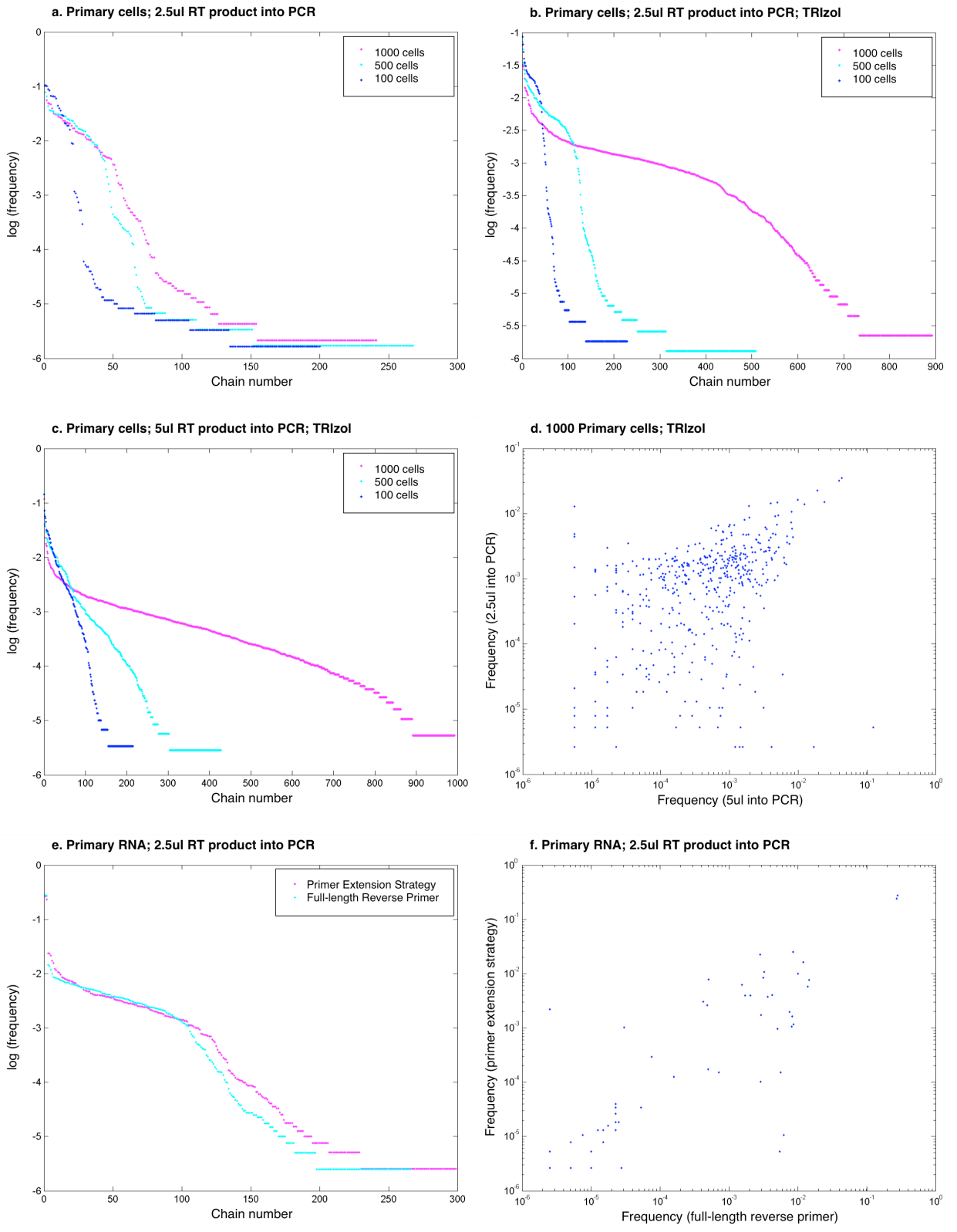


Figure 3.12: Optimization 4: RNA purification, α-chain sequencing data

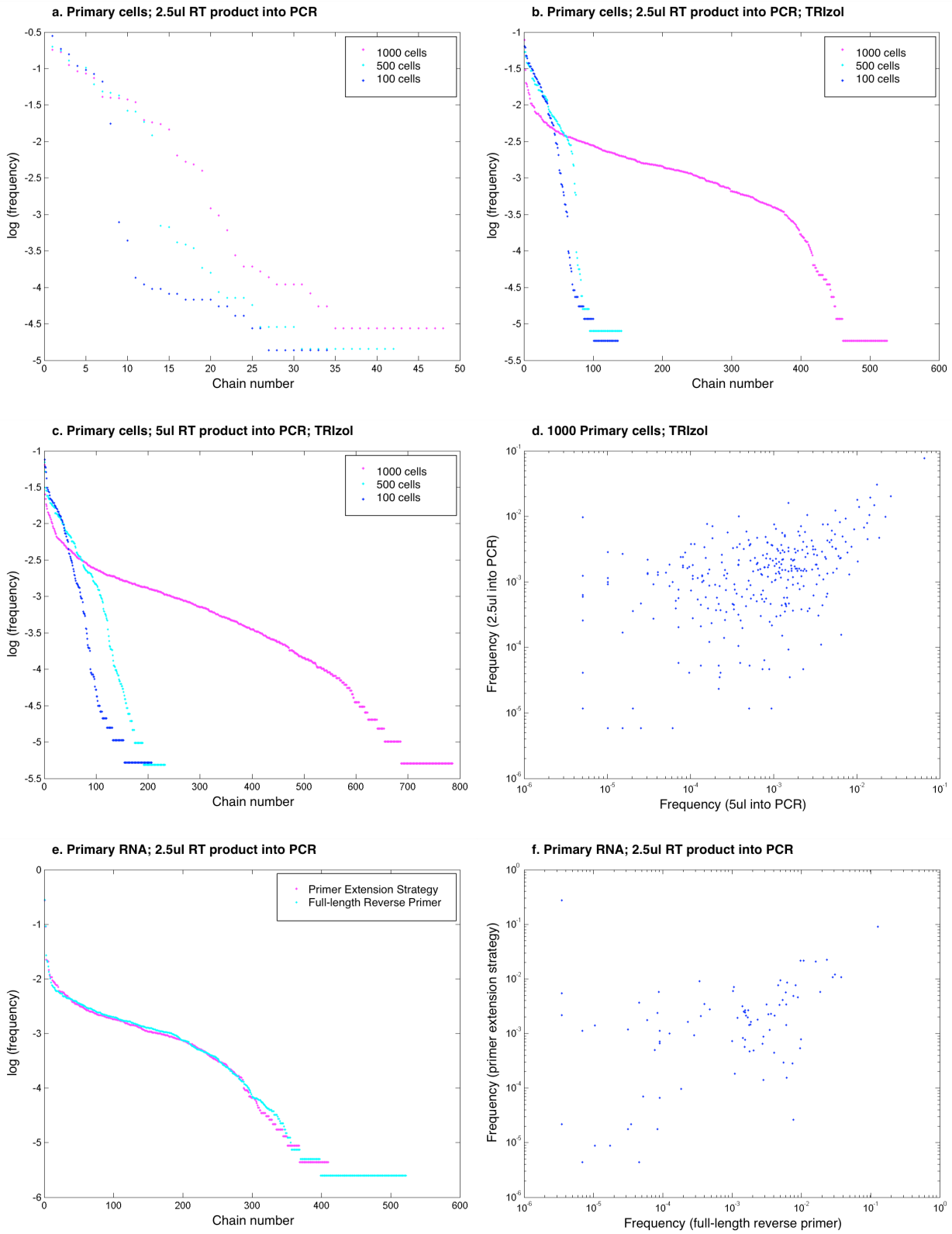


Figure 3.13: Optimization 4: RNA purification, β-chain sequencing data

3.4 Full Protocol- Attempt 2

Together, the optimizations described above drastically improved sensitivity of the library preparation protocol. The full protocol was therefore attempted a second time on a plate of primary human T cells. $CD8^+$ T cells were isolated from previously frozen PBMCs. TRIzol purification was performed on 60 replicates of 1000 cells, 12 replicates of 500 cells, and 12 replicates of 100 cells. Reverse transcription was carried out using optimized conditions: blocked RT primers, TCR-TSv2, 1 M betaine, and 12 mM $MgCl_2$. All 10 µl of RT product was used in each 50 µl PCR reaction. Products were pooled without normalization, and purified. The final library (Figure 3.14), spiked with 8.4 % Phi X, was sequenced on an Illumina MiSeq with 2 × 300-bp reads.

Of 15173052 reads from the 1000 cell samples, 11781698 were identified as alpha chain sequences, and 3294722 as beta chain sequences. 2344 unique alpha chains, and 2831 unique beta chains were extracted, only 6.6 % of the unique beta chains estimated by simulation. Because of this low sensitivity, not a single p-value below 1×10^{-4} was computed with the chain-pairing pipeline. Including 100 and 500 cell samples in the analysis did not improve pairing capacity. This failure is likely due to sample quality; this was the first experiment in which frozen cells were used.

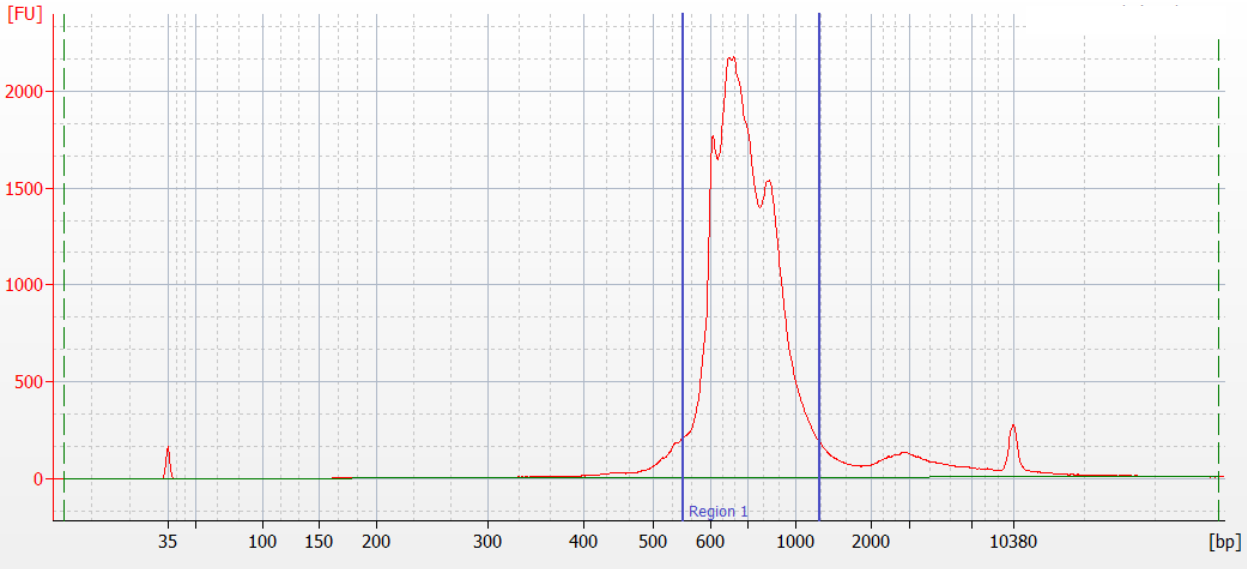


Figure 3.14: Full protocol attempt 2, final library

Region 1: 552-1314 bp, average size= 736 bp

3.5 Methods

3.5.1 Jurkat Culture

Jurkat culture was maintained between 1×10^5 and 1×10^6 cells/mL in RPMI supplemented with 10 % FBS, and 1 × L-GlutaMAX (Gibco).

3.5.2 Isolation of Human PBMCs

Blood was collected in Vacutainer tubes containing 22.0 g/L trisodium citrate, 8.0 g/L citric acid and 24.5 g/L dextrose (BD). PBMCs were isolated by density gradient centrifugation with Lymphoprep medium (STEMCELL Technologies) and SepMate-50 tubes (STEMCELL Technologies). If required, PBMCs were frozen in 10 % DMSO and 40 % FBS, and stored at -150 °C.

3.5.3 T cell Enrichment and Activation

$CD8^+$ T cells were isolated using EasySep magnetic particles (STEMCELL Technologies). Polyclonal activation of $CD8^+$ T cells was performed using Dynabeads Human T-Activator CD3/CD28 (Gibco) in RPMI with 10 % FBS, 1 × Glutamax, 1 mM sodium pyruvate, 25 mM HEPES, 50 µM 2-mercaptoethanol, 1 × penstrep, and 30 U/mL IL-2.

3.5.4 Reverse Transcription

Following TRIzol purification, RNA pellets were resuspended in 3 µl of 2 U/µl RNasin Plus RNase Inhibitor (Promega), 0.2 % Triton X 100, 0.1 % Tween 20, 1.67 µM each beta C gene primer C6 and alpha C gene primer TCRA-5, and 3.33 mM Advantage UltraPure PCR Deoxynucleotide Mix (Clontech) (to give final primer and dNTP concentrations in RT of 0.5 µM and 1 mM, respectively). Denaturation/primer annealing was performed at 72 °C for 3 min, and then samples were immediately transferred to ice. 7 µl of RT mix was added, to give a 10 µl RT reaction containing 1 x SMARTScribe First-Strand Buffer (Clontech), 10 U/µl SMARTScribe Reverse Transcriptase (Clontech), 2.5 mM DTT (Clontech), 0.5 M betaine (Sigma-Aldrich), 9 mM magnesium chloride, 1 µM TCR-TSv2, 1 U/µl RNasin Plus RNase Inhibitor (Promega), and 0.1 % Tween 20. Optimized betaine and magnesium chloride concentrations were 1 M, and 6 mM, respectively. RT was performed at 42 °C for 90 min, and reverse transcriptase denatured by holding at 70 °C for 15 min.

beta C gene primer C6 **:**
`CACGTGGTCGGGGWAGAAGC`

alpha C gene primer TCRA-5:
`CATTTGTTTGAGAATCAAAATCGGTGA`

TCR-TSv2 :
`CGC/ideoxyU/CCAAACCC/ideoxyU/ACGCAAACArGrGrG/3SpC3/`

TCR-TSv3 (Exiqon):
/5Biosg/rCrGrCrUrCrCrArArArCrCrCrUrArCrGrCrArArArCrArGrG+G/3SpC3/

3.5.5 Indexed PCR

Various volumes of RT product were used in 25 µl PCR reactions containing 1× Phusion HF Buffer (Thermo Scientific), 0.02 U/µl Phusion Hot Start II DNA Polymerase (Thermo Scientific), 0.25 mM Advantage UltraPure PCR Deoxynucleotide Mix (Clontech), 3 % DMSO (Thermo Scientific), 0.1 % Tween 20, 0.5 µM indexed forward Primer (TCRF-GSCX), 25 nM each Alpha SI-UT and Beta SI-UT, and 0.5 µM indexed reverse primer (TruSeq-GSCY). In some validation experiments full-length reverse primers (0.5 µM each TCRA-GSCZ and TCRB-GSCZ) were used instead of Alpha SI-UT, Beta SI-UT, and TruSeq-GSCY. Cycling conditions were: 98 °C 30 s; 98 °C 15 s, 70 °C 30 s, 72 °C 30 s × 35 cycles; 72 °C 5 min.

TCRF-GSCX:
AATGATACGGCGACCACCGAGATCTACACXXXXXXAAGCCTGTAATACGCTCCAAACCCTACGCAAA*C*A

Alpha SI-UT:
/5Phos/AATCCAGTGACAAGTCTGTCTGCCTAGATCGGAAGAGCACACGTCTG

Beta SI-UT:
/5Phos/GTTTGAGCCATCAGAAGCAGAGAAGATCGGAAGAGCACACGTCTG

TruSeq-GSCY:
CAAGCAGAAGACGGCATACGAGATYYYYYYGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T

TCRA-GSCZ:
CAAGCAGAAGACGGCATACGAGATZZZZZZGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCAGACAGACT
TGTCACTGGA*T*T

TCRB-GSCZ:
CAAGCAGAAGACGGCATACGAGATZZZZZZGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTCTGCTTCTGA
TGGCTCAA*A*C

3.5.6 Sequencing

In preparation for sequencing, libraries were 1:1 bead purified with 20 % PEG SeraMag beads (made in-house), pooled, ethanol precipitated, and size-selected on a 1 % agarose gel. Gel extraction was performed using the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel). Libraries were quantified by qPCR, using the Kapa Illumina Quantification Kit with custom standards. Sequencing was performed on an Illumina MiSeq with a 600-cycle MiSeq Reagent Kit v3. A custom read-1 sequencing primer was spiked into cartridge reservoir #12.

3.5.7 Sequencing Data Analysis

Indexed reads were demultiplexed using the MiSeq Reporter software. Using the FastX Barcode Splitter [28], alpha and beta reads were split into separate files by the following sequences (5 mismatches

permitted): alpha-AGGCAGACAGACTTGTCACTGGA, beta-TCTCTGCTTCTGATGGCTCAAAC. MiTCR [29] was then run on each set of reads (e.g. 192 sets for a 96 well plate), for partial correction of sequencing and PCR errors, and extraction of the CDR3 and variable, joining, constant, and diversity regions. For pairing, unique alpha and beta chains (unique CDR3 amino acid sequences) were extracted from MiTCR output files (custom perl script), and Fisher's exact test was performed on each alpha-beta combination using R [30].

Chapter 4: Conclusion

Next-generation sequencing has enabled inexpensive, in-depth profiling of immune repertoires. With bulk inputs, however, reads cannot be mapped back to their cell of origin, preventing pairing of TCR alpha and beta, or BCR heavy and light chains. Current methods for antigen receptor pairing require single-cell partitioning, and as such, are often low-throughput and costly. One higher-throughput approach, in which single cells are isolated in emulsion droplets and transcripts are fused, requires specialized equipment, is hard to properly control, and results in products that are difficult to sequence due to their long length.

By simply partitioning a bulk sample, and incorporating a tag that allows reads to be mapped back to these partitions, chain-pairings can be easily extracted using statistics. This is the basis for the method presented in this thesis. This method allows for comparable throughput to emulsion approaches and, because of the statistically nature, is internally controlled. Simulation demonstrated the method's potential; with 5000 cells per partition, and 96 partitions, over 20000 pairs were extracted. Also, in contrast to any previous reports, this method allows for selection of the clonal frequency range to be paired, by adjusting number of cells per well or extent of expansion, and for identification and pairing of functional clones.

The individual amplification and indexing reactions can be carried out in any number of different containers. We have chosen to use a 96-well microwell plate, a format that is standard in research laboratories around the world. To reduce cost, and further encourage adoption of the protocol, the following molecular biology features were chosen: i. *template switching*- by adding a universal priming site to cDNA, only a single forward PCR primer is required per index, and ii. *primer-extension strategy*- by synthesizing full-length reverse primers in the PCR reaction, only a single moderate-length reverse PCR primer is required per index. There is a minor drawback, however, in using template switching and a universal primer instead of a set of variable region forward primers; because of the 5' untranslated region on the mRNA, template switching results in a longer construct, which can not be completely sequenced with 2 × 300-bp reads.

The main focus of this thesis was optimization of the library preparation protocol. Successful amplification and sequencing of alpha and beta chain transcripts from primary human $CD8^+$ T cells was demonstrated. With 1000 input cells, 749 alpha and 591 beta chains were extracted using only half of the RT product in the PCR. This sensitivity is likely sufficient for statistical determination of pairings. Due to time constraints, only two attempts of the full chain-pairing pipeline could be performed. Unfortunately,

we did not observe the same sensitivity that we had in optimization experiments, and so no statistically significant pairs could be extracted. It was likely the sample quality that limited sensitivity in this case; $CD8^+$ T cells were isolated from previously-frozen PBMCs and immediately lysed, whereas in previous experiments $CD8^+$ T cells were isolated from fresh PBMCs and were activated for at least six days before use. Although cell viability was high after thawing, mRNA quality was not assessed. It is possible that mRNA was partially degraded in the freeze-thaw process. An activation step returns thawed cells to full health and likely results in higher transcript counts, as cells are preparing for division. In the future, select samples should be run on Agilent 2100 Bioanalyzer, using the Agilent RNA 6000 Pico Kit, to insure adequate RNA quality. Before the full protocol is attempted again, however, it is recommended that a higher-throughput RNA purification be validated to enable faster experiment iteration and ease debugging. TRIzol purification and isopropanol precipitation of 96 samples requires two days of tedious work.

If, after other issues are resolved, the sensitivity proves inadequate for pairing, further optimization can be performed; primer designs, primer concentrations, primer annealing temperatures, and the number of PCR cycles can be adjusted. Polyclonal activation could also be performed after partitioning to increase the number of starting transcripts.

While all validation in this work was performed on T cells, the method can be easily be applied to B cells by the substitution of only the small set of reverse transcription primers and constant-region forward PCR primers, and the analysis software (MiGEC [31], or MiXCR [32] in place of MiTCR [29]). In fact, pairing of plasma cells may prove possible even with a low-sensitivity protocol, as plasma cells have been shown to have up to 100× more transcripts than naive B cells [16].

The chain-pairing method I have described in this thesis will, with further refinement, enable immunological exploration not possible with single-chain repertoire analysis. First, it will allow us to revisit basic questions and challenge dogma that was established by sequencing of only a fraction of the potential repertoire. For example, Howie *et al.*, using a similar chain-pairing method, found that 2.8 % of TCR alpha sequences paired with two productive beta sequences [15]. This may indicate productive rearrangement of both beta genome loci, and an exception to beta genotypic allelic exclusion, putting into question the validity of employing beta chain sequencing to assess TCR repertoire diversity. Second, with paired-chain sequencing we can identify functional receptors, which is critical for design of personalized therapies. For example, by applying our method to tumor infiltrating lymphocytes we could identify both alpha and beta chains from neo-antigen reactive T cells. Adoptive T cell transfer could then be performed using T cells engineered to express these receptors [33, 34].

References

[1] A. Abbas, A. Lichtman and S. Pillai, *Basic Immunology: Functions and Disorders of the Immune System*. Philadelphia, PA: Elsevier Saunders, 2014.

[2] M. Vukmanovic-Stejic, B. Vyas, P. Gorak-Stolinska, A. Noble and D. M. Kemeny, "Human Tc1 and Tc2/Tc0 CD8 T-cell clones display distinct cell surface and functional phenotypes," *Blood,* vol. 95, pp. 231-240, 2000.

[3] J. A. Bluestone, C. R. Mackay, J. J. O'Shea and B. Stockinger, "The functional plasticity of T cell subsets," *Nat Rev Immunol,* vol. 9, pp. 811-816, 2009.

[4] S. M. Kaech and W. Cui, "Transcriptional control of effector and memory CD8+ T cell differentiation," *Nat Rev Immunol,* vol. 12, pp. 749-761, 2012.

[5] S. L. Swain, K. K. McKinstry and T. M. Strutt, "Expanding roles for CD4+ T cells in immunity to viruses," *Nat Rev Immunol,* vol. 12, pp. 136-148, 2012.

[6] D. J. Woodsworth, M. Castellarin and R. A. Holt, "Sequence analysis of T-cell repertoires in health and disease," *Genome Med,* vol. 5, 2013.

[7] J. Gorski, M. Yassai, X. Zhu, B. Kissela, C. Keever and N. Flomenberg, "Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status," *The Journal of Immunology,* vol. 152, pp. 5109-5119, MAY, 1994.

[8] G. C. Wang, P. Dash, J. A. McCullers, P. C. Doherty and P. G. Thomas, "T Cell Receptor αβ Diversity Inversely Correlates with Pathogen-Specific Antibody Levels in Human Cytomegalovirus Infection," vol. 4, pp. 128ra42, APR, 2012.

[9] A. Eugster, A. Lindner, A. Heninger, C. Wilhelm, S. Dietz, M. Catani, A. Ziegler and E. Bonifacio, "Measuring T cell receptor and T cell gene expression diversity in antigen-responsive human CD4(+) T cells," *J. Immunol. Methods,* vol. 400, pp. 13-22, DEC 31, 2013.

[10] E. Kobayashi, E. Mizukoshi, H. Kishi, T. Ozawa, H. Hamana, T. Nagai, H. Nakagawa, A. Jin, S. Kaneko and A. Muraguchi, "A new cloning and expression system yields and validates TCRs from blood lymphocytes of patients with cancer within 10 days," *Nat. Med.,* vol. 19, pp. 1542-1546, 2013.

[11] T. Cukalac, W. Kan, P. Dash, J. Guan, K. M. Quinn, S. Gras, P. G. Thomas and N. La Gruta L., "Paired TCRαβ analysis of virus-specific CD8+ T cells exposes diversity in a previously defined 'narrow' repertoire," *Immunol. Cell Biol.,* APR, 2015.

[12] A. Han, J. Glanville, L. Hansmann and M. M. Davis, "Linking T-cell receptor sequence to functional phenotype at the single-cell level," *Nat. Biotechnol.,* vol. 32, pp. 684-692, JUL, 2014.

[13] R. L. Warren, D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb and R. A. Holt, "Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen

selection and a directly measured repertoire size of at least 1 million clonotypes," *Genome Res.,* vol. 21, pp. 790-797, MAY, 2011.

[14] E. S. Egorov, E. M. Merzlyak, A. A. Shelenkov, O. V. Britanova, G. V. Sharonov, D. B. Staroverov, D. A. Bolotin, A. N. Davydov, E. Barsova, Y. B. Lebedev, M. Shugay and D. M. Chudakov, "Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers," *The Journal of Immunology,* MAY, 2015.

[15] B. Howie, A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson and H. S. Robins, "High-throughput pairing of T cell receptor α and β sequences," *Science Translational Medicine,* vol. 7, pp. 301ra131, AUG, 2015.

[16] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann and S. R. Quake, "The promise and challenge of high-throughput sequencing of the antibody repertoire," *Nat Biotech,* vol. 32, pp. 158-168, 2014.

[17] C. S. Carlson, R. O. Emerson, A. M. Sherwood, C. Desmarais, M. Chung, J. M. Parsons, M. S. Steen, M. A. LaMadrid-Herrmannsfeldt, D. W. Williamson, R. J. Livingston, D. Wu, B. L. Wood, M. J. Rieder and H. Robins, "Using synthetic templates to design an unbiased multiplex PCR assay," *Nature Communications,* vol. 4, pp. 2680, OCT, 2013.

[18] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao and M. A. Surani, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods,* vol. 6, pp. 377-382, MAY, 2009.

[19] J. Kapteyn, R. He, E. T. McDowell and D. R. Gang, "Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples," *BMC Genomics,* vol. 11, pp. 413, JUL 2, 2010.

[20] F. L. Pinto and P. Lindblad, "A guide for in-house design of template-switch-based 5′ rapid amplification of cDNA ends systems," *Anal. Biochem.,* vol. 397, pp. 227-232, FEB, 2010.

[21] S. Picelli, A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg and R. Sandberg, "Smart-seq2 for sensitive full-length transcriptome profiling in single cells," *Nature Methods,* vol. 10, pp. 1096-1098, NOV, 2013.

[22] S. Picelli, O. R. Faridani, A. K. Bjorklund, G. Winberg, S. Sagasser and R. Sandberg, "Full-length RNA-seq from single cells using Smart-seq2," *Nature Protocols,* vol. 9, pp. 171-181, JAN, 2014.

[23] C. E. Busse, I. Czogiel, P. Braun, P. F. Arndt and H. Wardemann, "Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes," *Eur. J. Immunol.,* vol. 44, pp. 597-603, 2014.

[24] K. Smith, L. Garman, J. Wrammert, N. Zheng, J. D. Capra, R. Ahmed and P. C. Wilson, "Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen," *Nat. Protocols,* vol. 4, pp. 372-384, 2009.

[25] B. J. DeKosky, G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Doerner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson,

A. D. Ellington and G. Georgiou, "High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire," *Nat. Biotechnol.,* vol. 31, pp. 166-169, FEB, 2013.

[26] M. A. Turchaninova, O. V. Britanova, D. A. Bolotin, M. Shugay, E. V. Putintseva, D. B. Staroverov, G. Sharonov, D. Shcherbo, I. V. Zvyagin, I. Z. Mamedov, C. Linnemann, T. N. Schumacher and D. M. Chudakov, "Pairing of T-cell receptor chains via emulsion PCR," *Eur. J. Immunol.,* vol. 43, pp. 2507-2515, SEP, 2013.

[27] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nat Meth,* vol. 11, pp. 163-166, 2014.

[28] Hannon Lab, "FASTX-Toolkit: FASTQ/A short reads pre-processing tools," 2015.

[29] D. A. Bolotin, M. Shugay, I. Z. Mamedov, E. V. Putintseva, M. A. Turchaninova, I. V. Zvyagin, O. V. Britanova and D. M. Chudakov, "MiTCR: software for T-cell receptor sequencing data analysis," *Nat Meth,* vol. 10, pp. 813-814, 2013.

[30] R Core Team. R: A language and environment for statistical computing. 2015.

[31] M. Shugay, O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, LukyanovSergey, T. N. Schumacher and D. M. Chudakov, "Towards error-free profiling of immune repertoires," *Nat Meth,* vol. 11, pp. 653-655, 2014.

[32] D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva and D. M. Chudakov, "MiXCR: software for comprehensive adaptive immunity profiling," *Nat Meth,* vol. 12, pp. 380-381, 2015.

[33] S. D. Martin, G. Coukos, R. A. Holt and B. H. Nelson, "Targeting the undruggable: Immunotherapy meets personalized oncology in the genomic era," *Annals of Oncology,* SEPT, 2015.

[34] M. Kalos and C. H. June, "Adoptive T cell Transfer for Cancer Immunotherapy in the Era of Synthetic Biology," *Immunity,* vol. 39, JULY, 2013.