UNIVERSITY OF CALIFORNIA
SANTA CRUZ

APPLICATIONS OF HIGH THROUGHPUT SEQUENCING FOR IMMUNOLOGY AND CLINICAL DIAGNOSTICS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Hyunsung John Kim

June 2014

The Dissertation of Hyunsung John Kim
is approved:

Professor Nader Pourmand, Chair

Professor David Haussler

Professor Phil Berman

Professor Martha Zuniga

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Applications of High Throughput Sequencing for Immunology and Clinical Diagnostics

by

Hyunsung John Kim

High throughput sequencing methods have fundamentally shifted the manner in which biological experiments are performed. In this dissertation, conventional and novel high throughput sequencing and bioinformatics methods are applied to immunology and diagnostics.

In order to study rare subsets of cells, an RNA sequencing method was first optimized for use with minimal levels of RNA and cellular input. The optimized RNA sequencing method was then applied to study the transcriptional differences between subpopulations of T follicular helper cells, which are integral to the adaptive immune response to pathogenic invasion.

In some cases, pathogens have long lasting effects either by integration of viral DNA into the host genome or by immune evasion. Paired-end and mate-pair sequencing are applied to identify the integration of DNA from high risk strains of human papilloma virus, an event that acts as a precursor to the cervical carcinogenesis. Some bacterial pathogens are able to escape the adaptive immune response and antibiotics are necessary to clear infections. However, evolved resistances can nullify the therapeutic benefit of antibiotic treatment. Whole genome sequencing was able to identify the genetic causes of antibiotic-resistance in both clinical isolates and directed evolution studies.

In most cases, the adaptive immune system is able to clear pathogenic invasions without the help of antibiotics. A key player in both adaptive immunity and tissue transplantation is the human leukocyte antigen (HLA). HLA molecules are responsible for surface display of healthy and pathogenic peptides. Knowledge of an individual's HLA types is imperative for successful tissue transplantation and is useful for diagnosis of autoimmune diseases such as type I diabetes, systemic lupus erythematosus and ankylosing spondylitis. Because balancing selection has generated thousands of HLA alleles in the population, identification of an individuals HLA alleles typically requires

specialized molecular assays. A novel method is presented that can predict HLA types directly from RNA-seq data without the need for specialized molecular assays.

HLA molecules that display abnormal peptides are recognized by T-cells via their characteristic T-cell receptor. Unlike most protein coding genes, the peptide sequence of the T-cell receptor is not encoded directly in the genome. Instead, a somatic recombination process generates receptors with the ability to bind a wide range of different peptides displayed by HLA molecules. The UCSC Immunobrowser was developed to explore, compare and analyze high throughput T-cell receptor sequencing experiments using interactive visualizations. The public web-based tool can serve as a repository for T-cell receptor sequencing experiments, track blood cancers, identify potential causes of autoimmunity and search the expanse of published literature for studies that have observed similar sequences.

Together, these applications highlight the utility of high throughput sequencing and bioinformatics methods for the study of immunology and the translation of relevant findings to clinical diagnostics.

To my parents, who through great faith and sacrifice showed me a righteous path.

And to Nicole, who held me up during the darkest of times.

Acknowledgments

I would like to thank everyone who took their donated their own time to help me complete this work. Foremost, I would like to thank my adviser and friend, Nader Pourmand, without whose support and guidance, this work would be impossible. I also thank the members of my thesis committee, David Haussler, Martha Zuniga and Phil Berman whom taught me by example how to stay positive, work hard and never give up. To my undergraduate advisor, Kevin Karplus, who showed me what it meant to be a great scientist and teacher. To my high school biology teacher, Tim Krieger, who showed me how magical biology was. To all my collaborators and fellow graduate students with whom I learned and explored the edge of knowledge.

This work was supported by a grant from the NASA, PSOC and HIPC.

Chapter 1

Introduction

Science, technology and medicine are intertwined. Often, cutting edge technology drives medicine, as in the development of the x-ray. Other times, medicine furthers science, as in the discovery of the vaccine. And in many cases, science and medicine are pursued simultaneously, as was the case with the sequencing of the human genome.

The human genome project was a breakthrough. Although the project was expensive with a price tag of $2.7 billion dollars, the initial draft genome held tremendous promise. For the first time, humanity had unravelled its own blueprint. Within the seemingly endless string of adenine, cytosine, guanine and thymines held not only the instructions to make a man, but also the cure to genetic disease. However, the single genome alone was not enough to identify, let alone, cure any genetic disease. After all, treatments based on one's genome would require treatments personalized to their unique genetic code.

It is no wonder then, that Archon Genomics sponsored an XPRIZE of $10 million to sequence a genome for less than $10,000. Affordably sequencing genomes was necessary in order to apply the human genome to medicine. High throughput sequencing (HTS) or Next Generation sequencing (NGS) methods were created to meet this goal. In 2008, James Watson was the first individual to have his genome sequenced using the first generation of HTS, 454 [169]. The genome cost less than $1 million– a 3000-fold cost reduction just five years after the release of the initial human genome.

The cost of sequencing has declined so dramatically that the $10 million XPRIZE was cancelled. Today a human genome can be sequenced for less than $1,000

using HTS. Despite the incredible reduction in cost, we still have yet to make the leap towards truly personalized medicine. This disjunction between affordability and application is caused not only by the complexity of the underlying biology, but also by the complexity of bioinformatic analysis. Many diseases are not completely genetic or have complex genetic elements. When many genomes of diseased individuals are sequenced, genetic variants are only capable of predicting small changes in risk of developing disease.

Other features such as faulty gene expression, epigenetics and abnormal representation of immune cells may account for larger portions of disease risk and etiology. As such, many alternative HTS assays have been created to study different features of nucleic acids including exome sequencing, RNA-seq, Chip-Seq and adaptive immune repertoire sequencing. With these assays, HTS can be used to study not only genetics, but also epigenetics, transcriptomics and immunomics.

Although sequencing costs are low, bioinformatic analysis is still costly. Even when a disease is known to be simply mendelian, the analytic cost of identifying causative variants is high ($1000-$16,000) in comparison to sequencing cost [19]. Analytic expenses are associated not only with large computational requirements but also with high development costs of specialized algorithms. For example, in order to perform a genome-wide association study (GWAS) using HTS genome sequencing, algorithms for short read mapping, SNP calling and GWAS enrichment methods must be developed. Each algorithm requires domain specific knowledge, long development times and extensive testing and benchmarking. Likewise, specialized algorithms must be developed in order to study specialized HTS assays such as RNA-Seq or immune repertoire sequencing. To make matters worse, different methods for similar sequencing assays carry their own method-specific biases that need to be accounted for.

This dissertation covers many aspects of HTS relating to diagnostics and study of the immune system. Chapters 2 and 3 utilize common bioinformatics methods for the analysis of RNA and DNA sequencing, beginning with the technical characterization of an RNA-seq method and following with the application of that method to study rare subsets of T cells. I then utilize HTS to study pathogenic challenges to human health including precancerous integration of viral DNA into the human genome and antibiotic

resistance from both clinical isolates and directed evolution studies. Chapter 4 describes a novel bioinformatic method for HLA typing from RNA-seq data useful for tissue transplantation and disease diagnosis. Chapter 5 describes visual data exploration and analytical methods for a specialized targeted HTS assay useful for studying the adaptive immune response to immune challenge.

Chapter 2

Pushing the Boundaries of RNA-Seq

2.1 Minimizing Total RNA Input

2.1.1 Introduction

When HTS is applied to sequencing cDNA, it is a powerful tool for the study of the transcriptome. Dubbed RNA-seq in 2008 by Mortazavi et al [104], it serves as a successor to the microarray. In contrast to microarrays, RNA-seq can uncover novel transcripts and provide a digital read out of expression levels with a vastly wider dynamic range [163, 103]. Since its emergence, RNA-seq assays have become the gold standard method for quantifying expression levels, detecting fusion genes and for general study of the transcriptome. RNA-seq is currently being applied to increasingly challenging studies, such as the sequencing of RNA from a single cell or even subcellular compartments [176, 147, 1, 174].

In 2010, however, RNA-seq assays typically required large quantities of input material. Protocols for generating sequencing libraries recommended starting with microgram quantities of total RNA. Although such quantities of RNA can be acquired from in vitro cultured cells or animal studies, the burdensome requirement greatly limited the types of studies to which RNA-seq could be applied. Any experiment for which input material is rare or limited was not feasible [118]. Thus, RNA-seq studies of rare cell populations such as stem cells, circulating tumor cells, samples from needle biopsies, developmental biology, forensics and immunology were limited by a lack of characterization.

Attempts to limit the quantities of input RNA were also hampered by ribosomal RNA and tRNAs that behave as contaminants by accounting for 60-90% of the RNA within a cell [170]. Poly-A enrichment and ribosomal depletion are normally recommended to prevent these highly abundant RNA sequences from dominating a sequencing experiment [158, 94].

We characterized an alternative cDNA synthesis method, Nugen Ovation, that does not require separate enrichment steps and is capable of generating cDNA from small quantities of RNA. Nugen Ovation utilizes a single primer isothermal amplification (SPIA) protocol [79]. SPIA is a highly sensitive RNA amplification method that utilizes a combination of poly-T chimeric primers and not-so-random hexamers for double stranded cDNA synthesis. Hexamers in not-so-random amplification methods are depleted for sequences found in rRNA sequences [5]. cDNA is then amplified linearly using an RNase H mediated isothermal method. A strand-displacing polymerase extends from a RNA primer. Primers are then degraded by RNase H only when an RNA-DNA duplex is formed. Following degradation of a primer, another free RNA primer is allowed to bind. The process creates a large number of single stranded DNA products. These single stranded products are then converted to double stranded DNA through the use of random nonamer priming.

Because Nugen Ovation potentially enabled transcriptomic studies where input RNA was limited, we assessed the viability of using the cDNA synthesis method for low input studies. We benchmarked the method against RiboMinus, an RNA depletion method, and TruSeq, a polyA enrichment method developed by Illumina. I tested sequencing libraries made from decreasing quantities of RNA for reproducibility, sensitivity and other factors that affect transcriptomic sequencing experiments. In addition, I compared two major sequencing platforms, Illumina and SOLiD, for their effect on RNA-sequencing experiments.

2.1.2 Methods

Detailed methods for sequencing library preparation are available in Tariq et al, 2011 [148].

Table 2.1: Samples Used for Sequencer Comparison

Radiation Exposure (Gy)	Replicates	Input (ng)	cDNA Synthesis Method	Sequencer
0	4	100	Nugen Ovation	lllumina GAIIx
2	4	100	Nugen Ovation	lllumina GAIIx
0	4	100	Nugen Ovation	SOLiD
2	4	100	Nugen Ovation	SOLiD

Table 2.2: Samples Used for Enrichment and cDNA Synthesis Comparison

Input	cDNA synthesis method	Enrichment	Replicates	Other
500 ng	Nugen Ovation	None	1	
100 ng	Nugen Ovation	None	2	
100 ng	Nugen Ovation	None	2	DNase Treated
50 ng	Nugen Ovation	None	1	
50 ng	Nugen Ovation	None	1	Sheared cDNA
50 ng	Nugen Ovation	None	1	Mixed: Sheared + Unsheared cDNA
10 ng	Nugen Ovation	None	1	
500 pg	Nugen Ovation	None	1	
50 pg	Nugen Ovation	None	1	
4 μg	TruSeq	Poly-A Enrichment	2	One replicate excluded due to sample preparation issues
4 μg	TruSeq	rRNA Depletion	2	

2.1.2.1 Samples

Eight Male Balb/C mice were acquired from Harlan Laboratories. Following a two day rest, four mice were exposed to 2 Gy dose of charged particle radiation from a proton source at a dose rate of 1 Gev/45s (2Gy). Four mice were kept unexposed as

controls (0Gy). Following dosage, mice were sacrificed by cervical decapitation. Testis tissue were removed by dissection and immediately flash frozen in liquid nitrogen. All biological replicates of 0Gy and 2Gy mice were used for comparisons between SOLiD and Illumina sequencing platforms. RNA from a single 0Gy mouse was used for comparisons between Nugen, RiboMinus and TruSeq.

2.1.2.2 cDNA synthesis using RiboMinus and Poly-A enrichment

Two replicates of 4 μg of total RNA was enriched using either Invitrogen's RiboMinus Eukaryote kit or Illumina's TruSeq Low-Throughput RNA sample protocol according to manufacturer's specifications. Enriched RNA was converted to cDNA following Illumina's TruSeq RNA sample preparation kit according to manufacturer's protocol.

2.1.2.3 cDNA Synthesis using Nugen Ovation

Total RNA without enrichment was used as the input sample for the Nugen Ovation Kit. For 0Gy and 2Gy samples utilized in sequencing platform comparison, 100 ng of starting total RNA was used. Our study used 500 ng, 100 ng, 50 ng, 10 ng, 500 pg and 50 pg as input material. The 100 ng sample, two technical replicates were generated using DNase treatment in the RecoverAll Total Nucleic Acid Isolation Kit (Applied Biosystems/Ambion). cDNA was synthesized following manufacturer's specification using Nugen Ovation RNA-seq Preparation kit (Nugen).

2.1.2.4 Sequencing Library Preparation

Libraries were generated following manufacturer's specifications in the Illumina TruSeq RNA sample preparation kit for all samples. All cDNA generated from enrichment-based methods were used as input for library preparation. 0.5-1 μg of cDNA generated by Nugen Ovaction was used as input. In addition, three libraries were constructed to test the effects shearing cDNA on final RNA-seq libraries. Non-sheared cDNA (fragment size-100-550 bp), sheared (100-360 bp) and equal mixtures by mass of sheared and non-sheared cDNA were used to generate libraries from cDNA generated

from 50 ng of starting total RNA. cDNA was sheared with a Covaris S2200 ultrasonicator.

2.1.2.5 Sequencing

0Gy and 2Gy samples were pooled and sequenced in a single quadrant of a SOLiD sequencer or two lanes of an Illumina GAIIx following manufacturer's specifications. Nugen, Poly-A and RiboMinus samples were pooled and sequenced across four lanes of an Illumina GAIIx following manufacturer's specifications.

2.1.2.6 Short Read Mapping

All short reads were initially aligned to mouse ribosomal sequences acquired from GenBank. Ribosomal subunits included 5s, 5.8s, 12s, 16s and 28s subunits [11]. Alignments were performed using Bowtie with default parameters [82]. Non-ribosomal reads were then mapped to the genome using the spliced aligner, TopHat, using the mm9 mouse genomic reference sequence and RefSeq genes [152]. Both the mm9 reference sequence and the RefSeq annotation were downloaded from the UCSC Table browser [71]. 0Gy and 2Gy samples were mapped as single ends and SOLiD reads were mapped using the colorspace option. Dilution and enrichment samples were mapped as paired-ends, where insert size was determined from BioAnalyzer traces of sequencing libraries. Only uniquely mappable reads were considered for downstream analysis.

2.1.2.7 Random Sampling

For serial dilution and enrichment samples, reads were randomly sampled to remove confounding effects of sequencing depth. An equal number of reads were sampled for all samples. The sample with the minimum number of uniquely mappable reads was used to set the number of reads to sample. Random lines of a SAM alignment were chosen from a uniform distribution until the requisite number of reads were met. All subsequent data analysis utilized randomly sampled reads unless explicitly stated otherwise.

2.1.2.8 Gene Abundance Estimation

RefSeq isoforms belonging to the same gene were collapsed into a single gene cluster by creating a single entry containing all known exons for the gene in bed12 format. Aligned reads were divided into six separate subclasses. Exonic or Intronic reads mapped wholly within a single exon or intron, respectively. Exon-Exon junctional reads spanned multiple exons. Exon-Intron junctional reads spanned the boundary of an exon and an intron. Intergenic reads mapped outside of gene clusters. Ribosomal reads were generated in the previously mapping-exclusion step described in the mapping methods subsection. Read counts for gene clusters were determined by counting the number of reads that landed within exonic or exon-exon junctional regions. Reproducibility of gene cluster abundance estimation was assessed using Spearman's correlation between gene counts.

2.1.2.9 Differential Expression and GC content bias

Differential expression analysis was performed to identify sequencer specific biases. Counts for gene clusters were used as input to DESeq [4]. Comparisons were made between 0Gy samples sequenced on Illumina versus 0Gy samples sequenced on SOLiD. 2Gy samples were similarly tested for differential expression. Gene clusters with a false discovery rate ¡ 0.01 were considered to be differentially expressed. GC content of all gene clusters were calculated within exonic regions only. Densities of up-regulated, down-regulated genes and all genes based on GC content were plotted using R.

2.1.2.10 Sensitivity of Exon Detection

A set of non-overlapping exons was created by collapsing all exons represented in the RefSeq annotation. In instances where exons overlapped, a single exon representing the union overlapping exons was reported. An exon was considered to be detected if the exon had at least 1x coverage.

2.1.2.11 Evenness of Coverage

Coefficient of variation (cv) was used to assess the evenness of coverage across transcripts. Low cv values correspond to even coverage. RefSeq transcripts were sorted

by average coverage over the transcripts. For the top 50% of transcripts, the cv of coverage at each position along the transcript was calculated. The average cv value was reported for each sample.

2.1.2.12 Library Complexity

As another proxy to measure evenness of coverage across a transcript, we examined the number of unique single-end and paired-end starting positions. Increased rates of unique start positions indicates more even coverage across transcripts.

2.1.2.13 Extreme End Bias

Extreme end bias was measured multiple ways. All transcripts with at least 5x coverage were considered. For percentile biases, all transcripts were binned into percentiles and the average coverage over each percentile was calculated. For nucleotide level biases, all transcripts greater than 2000bp in length with 5x overall coverage were considered. The percentage of transcripts with coverage at each position from the 5' or 3' end was reported.

2.1.3 Results

Of the different cDNA enrichment and synthesis methods, TruSeq with Poly-A enrichment performed the best in terms of uniquely mapping rate, followed by Nugen then by RiboMinus. Poly-A enriched TruSeq samples had $8.7\text{x}10^7$ out of $1.3\text{x}10^8$ reads map (64.0%). Nugen samples had a total of $1.1\text{x}10^7$ out of $1.7\text{x}10^7$ reads map (60.6%). RiboMinus had significantly fewer mappable reads with $4.7\text{x}10^7$ out of $1.0\text{x}10^8$ reads (46.7%). Of reads that mapped, 69.9% mapped to exons for Poly-A enriched TruSeq Libraries versus just 30.2% and 36.6% for Nugen and Ribominus respectively. Both Nugen and RiboMinus methods had a large number of reads mapping to intergenic regions. This result has been corroborated in RNA-seq, microarray and EST sequencing studies involving whole transcriptome versus poly-A enriched samples [23, 93].

In terms of removal of contaminating ribosomal reads, Poly-A enrichment performed the best, with Nugen and Ribominus following (Figure 2.1). Poly-A enriched

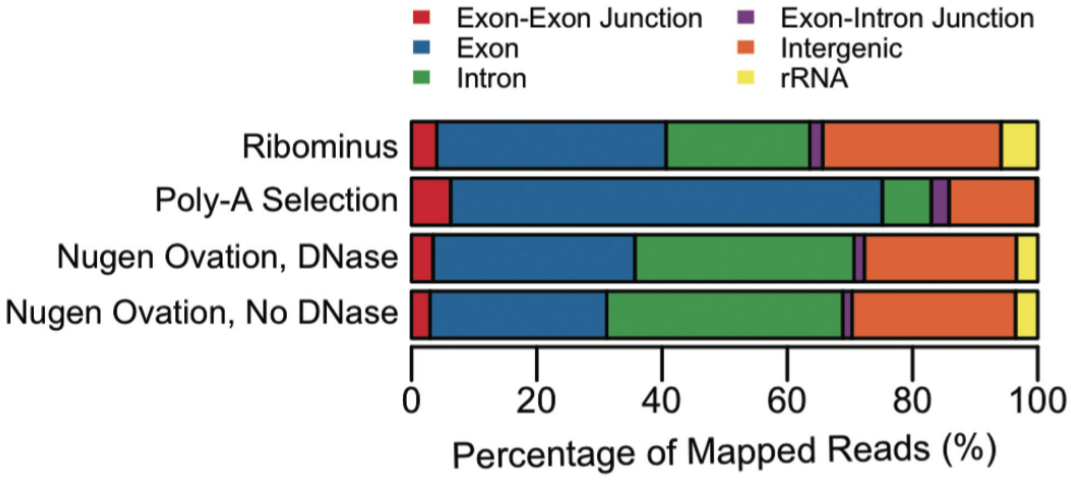

Figure 2.1: Comparison of average mapping statistics between Ribominus, Poly-A selection using Illumina TruSeq and Nugen Ovaction with or without DNase treatment.

samples had less than 1% of all mapped reads mapping to ribosomal sequences. Nugen and RiboMinus had 3.5% and 5.8% respectively, indicating that Nugen is suitable method for removing ribosomal contamination.

Sequence mapping rates were higher for 0Gy and 2Gy samples sequenced on the Illumina GAIIx as compared to SOLiD. Of the $1.7\text{x}10^8$ reads generated for all samples on SOLiD, $9.3\text{x}10^7$ (54.6%). For Illumina GAIIx, $6.15\text{x}10^7$ reads were generated, of which $3.9\text{x}10^7$ mapped (62.6%).

Library complexity was measured by the number of unique start sites found within a sample. When assayed as single ends, Nugen had the largest percentage of reads with unique start sites at 79%. Ribominus and Poly-A enriched samples followed with 59% and 37% respectively. However, when start and end positions of a fragment were considered, the vast majority of all libraries had unique start and end sites (¿ 99%). Although this could be caused by extreme end biases, it is likely that the increase in unique single end start sites was a result of increased intergenic reads within Nugen and Ribominus samples.

We found that Poly-A enriched and RiboMinus samples had similar evenness of coverage across transcripts with cv values of 1.94 and 2.07, respectively (Figure 2.3). However, Nugen samples had a drastically larger cv value of 3.54 indicating that coverage

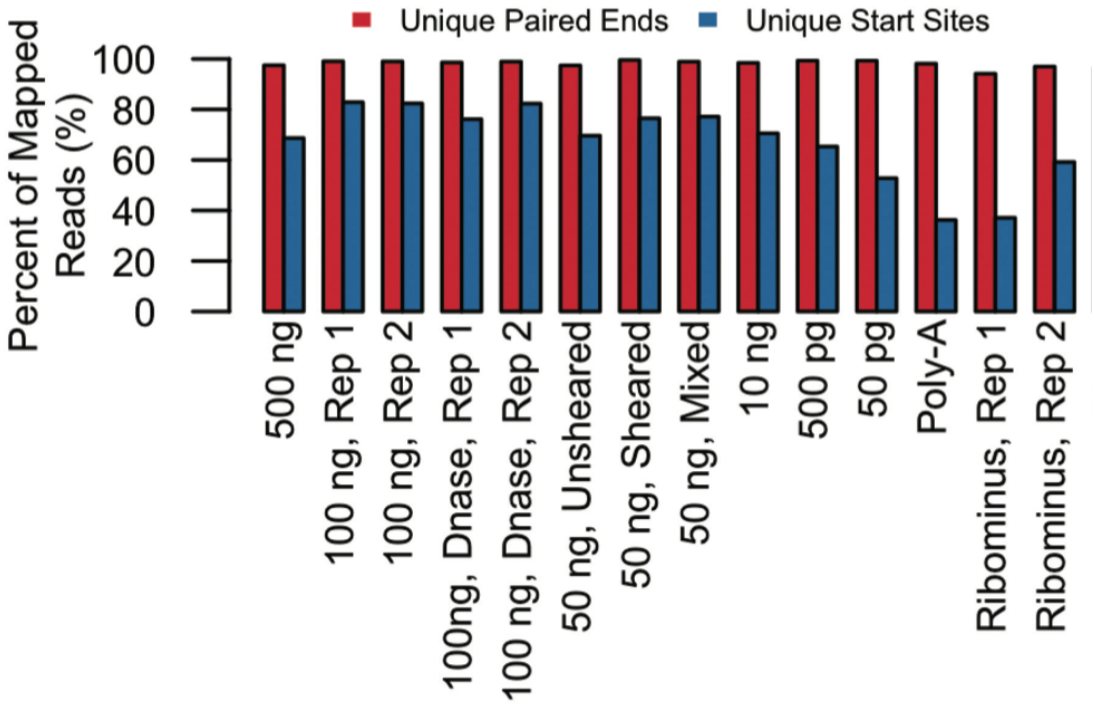

Figure 2.2: Complexity of libraries measured by percentage of reads with unique single or paired-end sites. All samples had similar levels of unique paired-end positions indicating low levels of PCR duplication overall. However, Poly-A selection and one Ribominus sample had a much lower percentage of unique single end start sites that may be consistent with 3' or 5' bias.

over transcripts was very spiky. This uneven coverage could be attributable to the use of random N-mers during first and second strand synthesis and could pose a problem for isoform quantification.

Of the three enrichment methods, Nugen displayed the least extreme 5' and 3' end bias overall. When assessed over percentiles of all transcripts, Poly-A enriched samples had a noticeable 3' bias, while RiboMinus had generally higher coverage on 5' end of transcripts (Figure 2.4.A). Nugen samples had a much smaller bias at the 3' end of transcripts. Similar observations were made when looking at nucleotide distance from the extreme 3' and 5' ends of long transcripts (Figure 2.4.B). When we assessed coverage bias of transcript length in addition to percentile, we observe good coverage over short transcripts, less than 1,000bp long for all methods (Figure 2.4.C). Biases

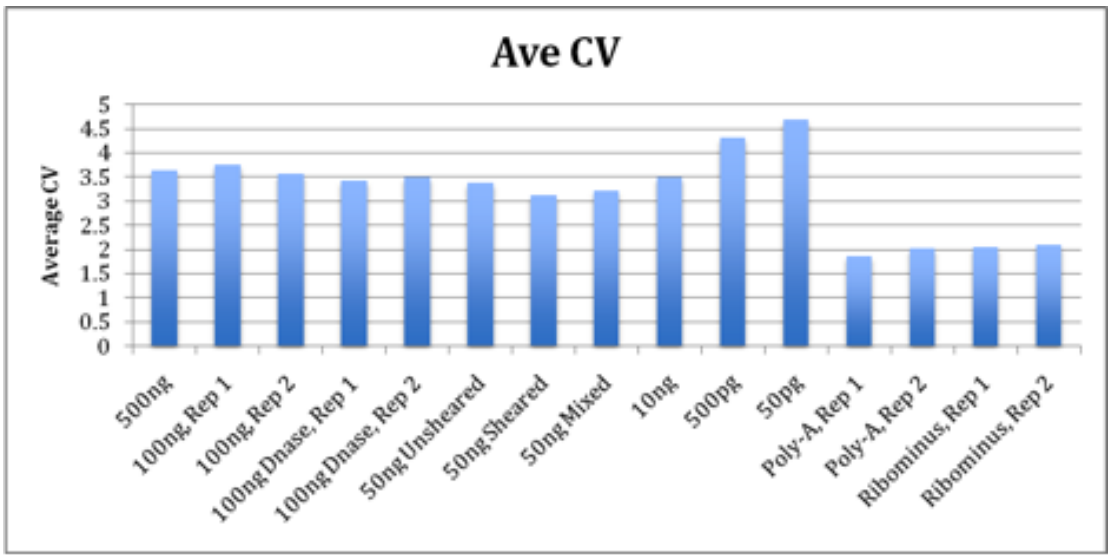

Figure 2.3: Average coefficient of variation (cv) of coverage across 50% of highest expressed transcripts. Poly-A and Ribominus samples are labeled. All other samples were generated using Nugen Ovation. Higher cv values are associated with less even coverage. Here Poly-A and Ribominus performed comparably. Samples prepared with Nugen Ovation had significantly less even coverage. Coverage of samples with lowest input material (500pg and 50pg) had least even coverage.

occur most prevalently for lengths greater than 1000bp long. Nugen samples exhibit a 3' bias for transcripts between 1,000 and 2,000 bp in length, but lacks 3' coverage for transcripts greater than 2,000bp in length.

Using Nugen samples with decreasing amounts of input material, we assessed the effect input material had on reproducibility of gene abundance estimation and sensitivity of expressed exon detection (Figure 2.5). Overall, RiboMinus and Poly-A enriched samples had highest correlations between each other. Technical replicates of RiboMinus had a Spearman's correlation of 0.982 versus 0.913 for technical replicates of Nugen using 100ng of starting material. Spearman's correlation of Poly-A replicates were low, but these were most likely due to errors during library preparation. Correlation between RiboMinus and Poly-A exceed correlations between Nugen technical replicates. Correlations between decreasing quantities of Nugen samples using 10ng or more input material were all above 0.9. Correlations between samples of 50 or 500 pg fell below, indicating a loss of reproducibility with minute quantities of RNA.

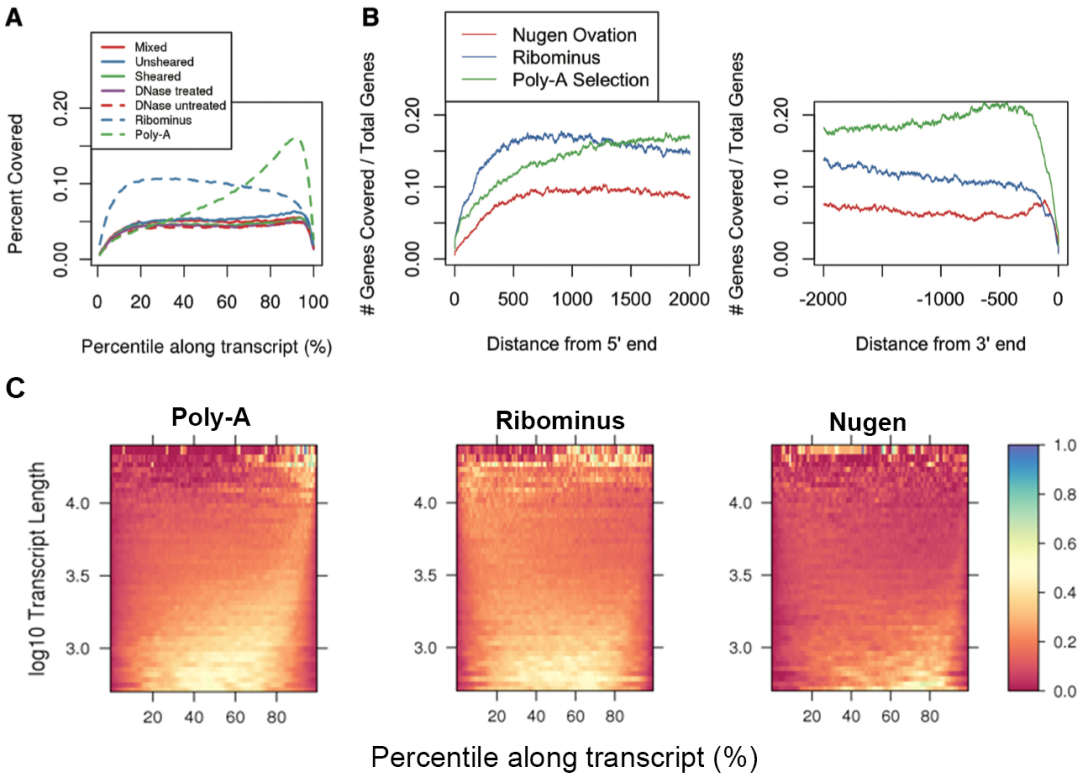

Figure 2.4: Extreme end bias in assessed over top 50% of expressed transcripts. A. Average coverage over percentiles shows extreme 3' bias in Poly-A samples, slight 5' bias in Ribominus samples and slight 3' bias in Nugen samples. B. Average coverage over long genes (>2,000 bp) from 5' end (left) or 3' end (right). C. Coverage bias displayed in a heatmap for Poly-A (left), Ribominus (center) and Nugen (right) binned by transcript length (y-axis) and percentile (x-axis). Colors indicate percentage of transcripts with coverage within a transcript length and percentile bin. In Poly-A samples, Extreme end biases are visible in longer transcripts (>1000 bp) in Poly-A and Ribominus samples. Nugen samples have 3' bias in shorter transcripts along with reduced 3' coverage in longer transcripts.

When equal numbers of reads were sampled, the ability to detect expressed exons was similar for all samples above 500pg, with approximately 50% of non-overlapping exons detected for all samples (Figure 2.6). However, 500pg and 50pg samples had reduced sensitivity with 44.7% and 28.9% of exons detected, respectively. Shearing of input material and DNAse treatment had little effect on sensitivity.

Finally, when we compared the Illumina GAIIx against the SOLiD sequencer,

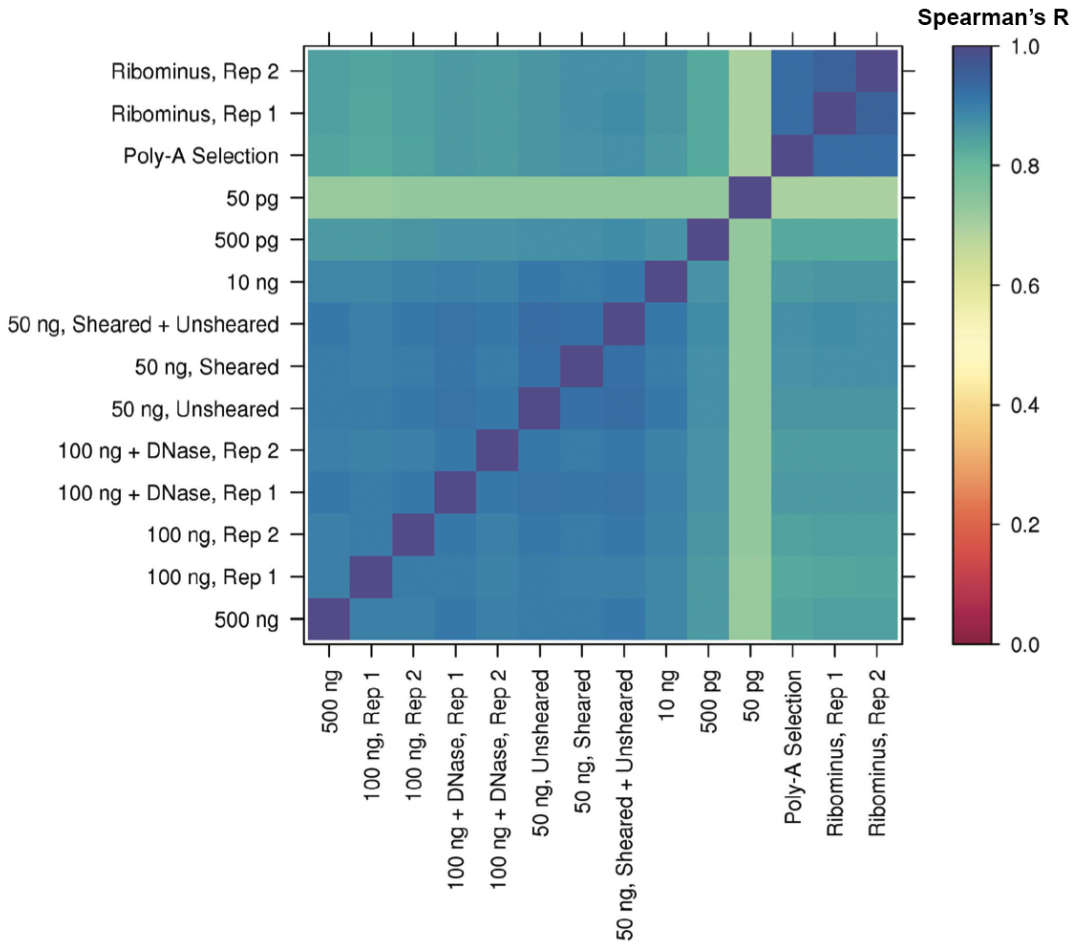

Figure 2.5: Pairwise Spearman's correlations between Nugen Ovation samples with decreasing amounts of starting material, Poly-A, and Ribominus samples. Poly-A and Ribominus samples show high correlations between each other; however, these samples do not share high correlation to all Nugen samples. Nugen samples had relatively high correlations to other Nugen samples, but correlations dropped for low input samples with 500pg or less.

we observed a difference in sensitivity and GC bias. 59% of exons were detected by both Illumina and SOLiD. However, 17% percent of exons were detected only by SOLiD, whereas only 3% of exons were detected only by Illumina. We performed a differential expression experiment between technical replicates sequenced on both SOLiD and Illumina. We found that statistically upregulated genes in SOLiD had a higher GC content than genes found to be upregulated in Illumina (Figure 2.7). Despite the observed GC

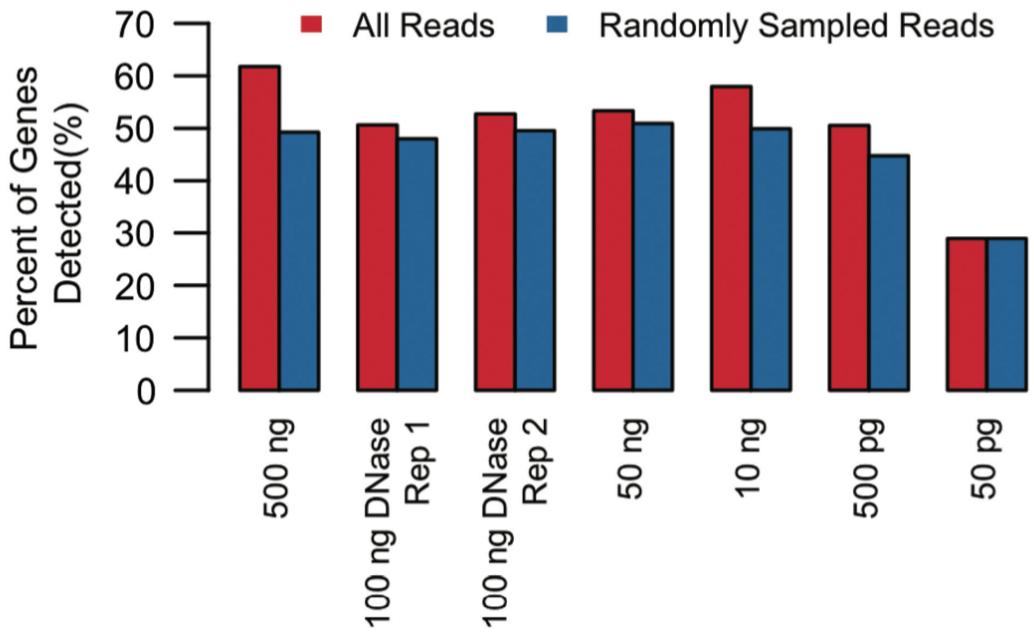

Figure 2.6: Sensitivity of detection of gene expression in Nugen samples with decreasing inputs. Bars in red represent the number of genes detected using all available reads. Bars in blue represent the number of genes detected when an even number of reads were randomly sampled from all samples. Sensitivity begins to decrease with the 500 pg sample and is greatly reduced in the 50 pg sample.

bias, we did not notice a substantial difference in the GC content of genes upregulated by either platform. However, the GC bias found in SOLiD may contribute to the increased exon detection sensitivity. Genic regions tend to be clustered within high GC regions of the genome [80].

2.1.4 Discussion

The strengths and weakness of the Poly-A enrichment, ribosomal depletion with RiboMinus and isothermal amplification with Nugen Ovation were assessed in this study. Overall, we found the enrichment of Poly-A tailed transcripts to have the greatest enrichment within exonic regions, but it had a significant 3' bias. RiboMinus resulted in a slight 5' bias and we observed slightly greater enrichment within exonic regions when compared to Nugen Ovation. Nugen Ovation generates reads with a slight 3' bias, but has significantly less even coverage. Spiky coverage may be an issue for isoform

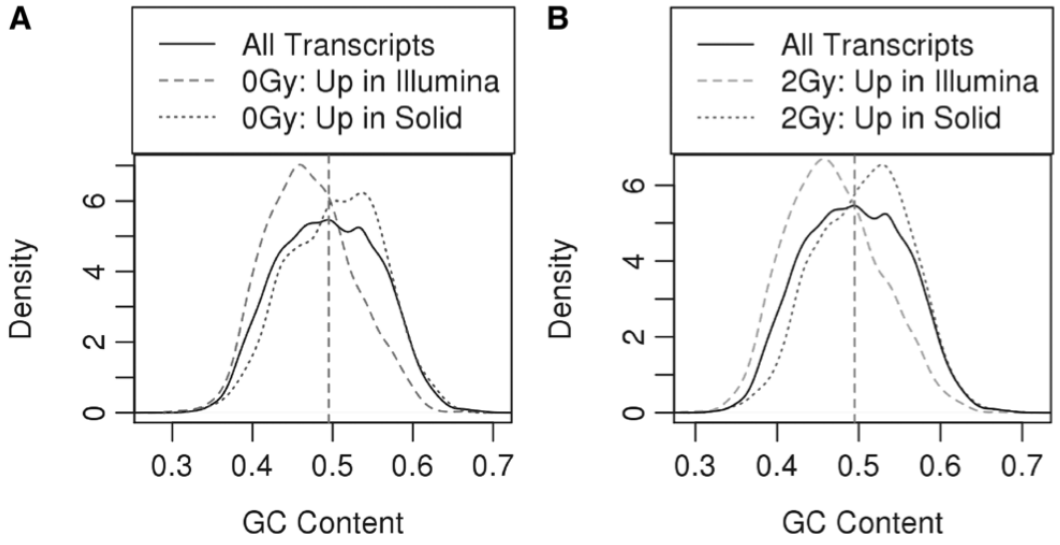

Figure 2.7: Density of GC content in genes found to be differentially expressed between SOLiD and Illumina. Genes with higher expression in SOLiD were found to have higher GC content, whereas genes with higher expression in Illumina were found to have lower GC content.0Gy and 2Gy mice are shown in A and B, respectively.

quantification, where many methods assume uniform coverage across a transcript.

Both Nugen and RiboMinus were effective for removing ribosomal reads while leaving the remainder of the transcriptome intact. Large numbers of intergenic reads for both libraries are consistent with reports of non-polyadenylated nuclear RNAs (ncRNAs) [23, 69, 63]. Poly-A enrichment methods would exclude ncRNAs from the final library. Thus, selection of enrichment and cDNA synthesis method should take into account the necessity of preserving ncRNA transcripts.

Poly-A enrichment and RiboMinus techniques were both highly reproducible, even with each other. Although Nugen Ovation samples were less reproducible, their expression estimates were consistent for vary amounts of input total RNA. Shearing of cDNA produced with Nugen generated no noticeable differences.

Based on our results, Poly-A enrichment appears most suitable when sufficient material exists and mRNA expression level estimates are the most important factor for RNA-seq experiments. RiboMinus is also preferable to Nugen Ovation when sufficient input material is available, due to higher reproducibility. Despite these shortcomings,

Nugen Ovation is a promising technique where whole transcriptome reads are desired and quantity of input RNA is a limiting factor. Nugen Ovation was shown to be reproducible and sensitive when as little as 10 ng of RNA was used as input. Additionally, we found the assay to be relatively sensitive and reproducible even when using 500 pg of RNA as input.

This experiment also demonstrates the feasibility of single cell RNA sequencing. A typical mammalian single cell contains 10-30pg of RNA. Although we saw reduced sensitivity and reproducibility at these levels, it is still possible to produce usable HTS data from extremely small quantities of RNA. The findings of this study have been further followed up by studies which have also benchmarked Nugen Ovation and other cDNA synthesis methods for degraded RNA samples (such as those from formalin-fixed paraffin-embedded tissue samples) and single cell studies [2, 174].

2.2 Minimizing Cellular Input and Studying Rare Follicular Helper T Cells

2.2.1 Introduction

In the previous section, we explored the extent to which we could reduce total input RNA for reproducible RNA-sequencing experiments. The ultimate goal of the experiments was the application of RNA-seq to study samples where input material is a limiting factor. One such application is the study of rare subsets of immune cells, such as follicular helper T-cells (T_{FH}).

T_{FH} cells are lymphocytes that play a large role in the regulation of B-cells. T_{FH} cells are a distinct T-cell subpopulation that specialize in providing help to B-cells in germinal centers [157]. T_{FH} cells are marked by their expression of CD4 and CXCR5. CD4 is the characteristic receptor expressed by helper T-cells (T_H). CXCR5 is an important receptor for the initial migration of T_{FH} cells into lymphoid follicles that contain germinal centers [17, 135, 74, 126]. A small subset of helper T-cells expressing CXCR5 circulate freely in blood, rather than being localized to germinal centers. It is likely that these circulating cells are the memory compartment of T_{FH} cells [102].

T_{FH} cells show remarkable plasticity and can mimic the effector functions

of T_{H1}, T_{H2} and T_{H17} cells depending on the type of infectious agent the immune encounters [130, 84, 91]. T_{FH} cells are often subdivided into three separate partitions: T_{FH1}, T_{FH2}, and T_{FH17} depending on the similarity of their expressed surface receptors and effector functions to a T_H subset. For example, following viral infection, T_{FH} cells excrete IFN-gamma, a cytokine typically produced by T_{H1} cells [96]. Infection of mice with Helminth parasites results in T_{FH} cells that produce IL-4, a cytokine typically released by T_{H2} cells [77, 46].

In order to study these special characteristics of T_{FH} cells, we devised a series of experiments to enable transcriptomic studies. In our first experiment, we optimize a method for cDNA synthesis directly from cell lysate using the Nugen Ovation kit. Directly synthesizing cDNA from lysate prevents losses from total RNA isolate. We identify an assay-specific minimum number of cells by sequencing technical replicates of FACS-sorted CD4+CXCR5+ blood T_{FH} cells with 100 cells up to 25,000 cells used as input. We also study the effects of in vitro activation of T_{FH} cells and report differentially expressed genes and gene ontologies.

Minimizing cellular input and total RNA loss is necessary because T_{FH} cells are relatively rare. A typical 5 ml blood sample contains $5x10^6$ PBMCs, 40% of which are CD4+ and only 4% are CXCR5+. Furthermore, T_{FH} cells are compartmentalized by their surface receptors and effector functions. We utilize the results of our first experiment to explore the subcompartments of T_{FH} cells within two individuals before and after vaccination with Fluzone. We observe large transcriptional differences between T_{FH} cells and other helper T-cell subsets, further supporting the idea that T_{FH} cells are a unique subclass of helper T-cells. Additionally, we observe expressional changes in the T_{FH1} compartment consistent the viral adaptive immune response.

Although the sample size of our study is low, it demonstrates our ability to study T_{FH} cells through RNA-sequencing. Our study serves as a useful pilot for the application of high throughput sequencing using small numbers of cells as input and shows its applicability for the study of T_{FH} cells.

2.2.2 Methods

2.2.2.1 Isolation of Cells

For titration studies, 45 ml of blood was drawn from two healthy subjects. Blood samples were centrifuged at 200xg for 10 minutes and the buffycoat was extracted. Helper T-cells were enriched using anti CD4+ magnetic beads. Enriched helper T-cells were then FACS-sorted and CD4+CXCR5+ T_{FH} cells were isolated. T_{FH} cells were split evenly into two pools. One pool was stimulated in vitro using anti-CD3/CD28 beads. Stimulated and unstimulated cells were aliquoted into 100, 1,000, 5,000 and 25,000 cell aliquots in duplicate. 25,000 cell samples were generated only from one of the two patients, due to the limited number of T_{FH} cells available.

For T_{FH} subcompartment studies, 45 ml of blood was drawn from two healthy subjects seven days following vaccination with Fluzone. T_H cells were enriched using methods identical to those used in the titration studies. Enriched T_H cells were FACS sorted into seven distinct groups: blood T_{FH} cells(CD4+CXCR5+), blood T_{FH1} cells (CD4+CXCR5+CXCR3+), memory blood T_{FH} cells (CD4+CXCR5+CCR6+), activated Th cells (CD4+CD154+), activated blood T_{FH} cells (CD4+CD154+CXCR5+), and activated blood T_{FH2} cells (CD4+CD154+CXCR5+ICOS+). Three replications of 5000 cells were used as input for blood T_{FH}, blood T_{FH1} and memory blood cells. 1-2 replicates whose quantity varied between 1,000 and 25,000 cells were used as input for activated Th cells, activated blood T_{FH} cells and activated blood T_{FH2} cells depending on rarity.

2.2.2.2 cDNA Synthesis and Sequencing

cDNA was synthesized robotically from cell lysate in a Nordiag M8000 liquid handler using methods described in Tariq et al, 2011 with the exception that cellular lysate rather than purified total RNA was used as input [148]. Sequencing libraries were prepared using robotic methods described in Hesson et al [38]. Samples were barcoded, pooled and sequenced across twelve lanes of an Illumina Hiseq2.

2.2.2.3 Alignment

Reads were aligned to the transcriptome using tophat2 with the hg19 human genome reference and UCSC knownGenes annotation [75, 57]. Parameters for the mean and standard deviation of insert size were set individually for each sample. MLE estimates of the mean and standard deviation of insert size were generated empirically from the insert sizes of one million randomly sampled reads aligned to the hg19 genome using bowtie2 with maximum insert size set to 1000 bp and only unique concordant alignments reported [81]. Only uniquely aligned concordant reads were utilized for downstream analysis.

2.2.2.4 Exon and Gene Abundance Estimation

Abundance estimation was determined using methods described in the previous section.

2.2.2.5 Sensitivity

Sensitivity of exon detection was determined through rarefaction analysis. For each sample, $1x10^6$, $2x10^6$, $4x10^6$, $8x10^6$, $1.6x10^7$ and $3.2x10^7$ reads were randomly sampled in triplicate. Exons were considered detected if at least one read was found to map within the exonic region.

2.2.2.6 Differential Expression

Differential Expression analysis was performed using DESeq2 on gene clusters [89]. For titration and in vitro activation studies, differentially expressed genes due to activation with anti-CD3/CD28 beads was determined by comparing 5,000 cell samples before and after activation. For T_{FH} subset experiments, differential expression was performed between technical and biological replicates of blood T_{FH} cells vs blood $T_{FH}1$ cells, blood T_{FH} cells vs blood T_{FH} memory cells, activated Th cells vs activated blood T_{FH} cells, and activated blood T_{FH} cells vs activated blood T_{FH2} cells. Genes with an FDR below 0.01 were considered to be differentially expressed.

2.2.2.7 Geneset Enrichment Analysis

Geneset enrichment analysis of genes differentially expressed by in vitro activation of T_{FH} cells was performed using GoSeq [178]. Geneset enrichment analysis of T_{FH} subsets were performed using enrichR [22].

2.2.3 Results

2.2.3.1 Optimal Number of Cells

Based on library functionality, reproducibility of gene expression estimates, exon sensitivity, we determined 5000 cells was optimal for use with Nugen Ovation cDNA synthesis directly from cell lysate.

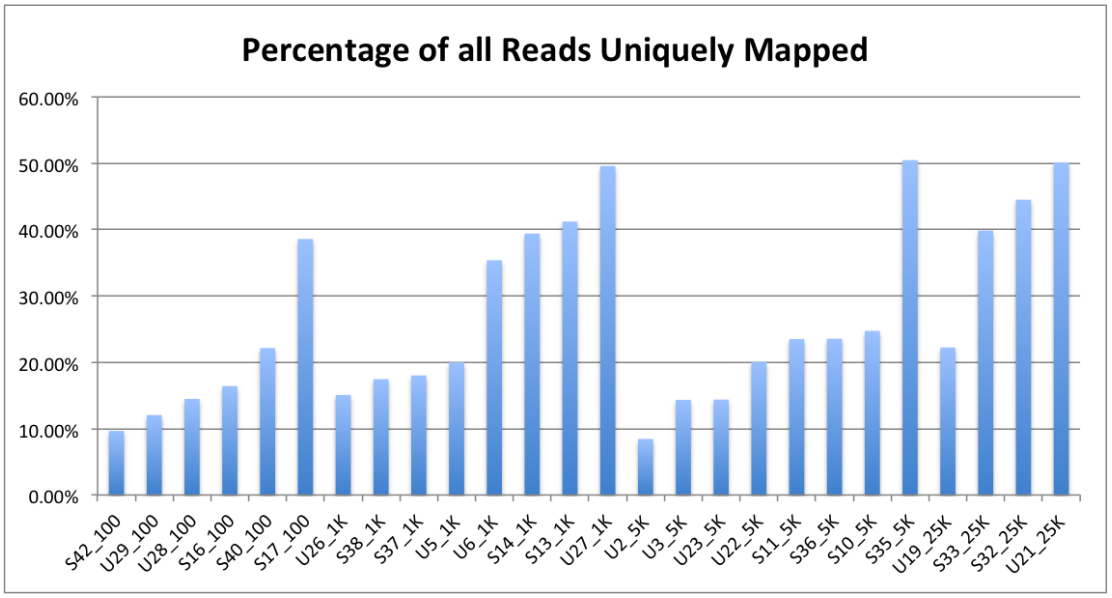

Figure 2.8: Percentage of uniquely mapped reads for each sample. The percentage of mapped reads is related to both library functionality and sequencing depth. IDs represent the patient number, the stimulation status (stimulated or unstimulated), the number of cells and the replicate.

The percentage of reads mapping uniquely to the genome was below 50% for all but two samples (Figure 2.8). 25,000 cell samples exhibited the highest mapping rates (mean 39%), followed by 1,000 cell samples (mean 30%). 5,000 and 100 cell samples exhibited low mapping rates with 22% and 19%, respectively. The highest individual

mapping rate belonged to a 5,000 cell sample, despite lower mapping rates across other 5,000 cell samples.

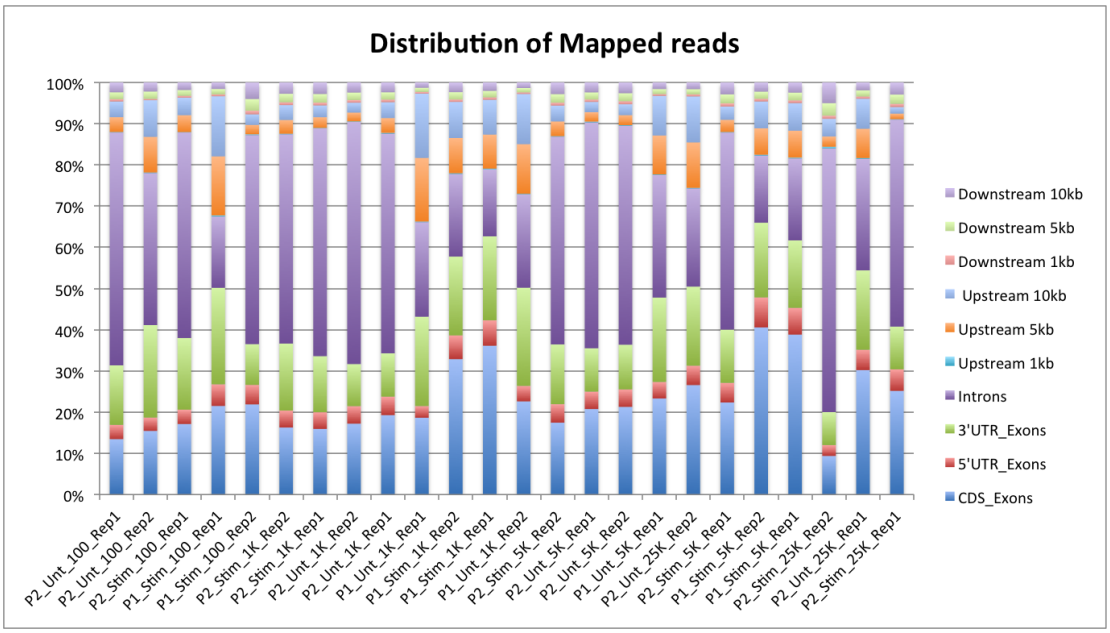

Figure 2.9: Distribution of uniquely mapped RNA-seq reads for each sample sorted by the percentage of reads usable for expression analysis. Reads are divided into coding, intronic, utrs and upstream and downstream flanking regions. IDs represent the patient number, the stimulation status (stimulated or unstimulated), the number of cells and the replicate. 1,000 and 5,000 cell samples show highest enrichment of reads mapped to exons.

5,000 cell samples also had the highest percentage (26%) of reads mapping to coding exons. Both 25,000 and 1,000 cell samples had 22% mapping to coding exons, while 100 cells samples had 18%. The percentage of reads mapping towards coding exons is reflective of reproducibility of gene expression estimation.

Reproducibility was assessed using pairwise correlation analysis of gene expression estimates between replicates with the same number of starting cells. 5,000 cell samples had greater linearity than any other starting input amount. 1,000 cell starting populations had a median correlation of 0.96 whereas 25,000 cell starting input had median correlation of 0.83. 100 cells samples behaved erratically and a median correlation was just 0.69. The distribution of pairwise correlation was also indicative of high reproducibility among 5,000 cell samples which ranged by from 0.94 to 0.98. 1,000

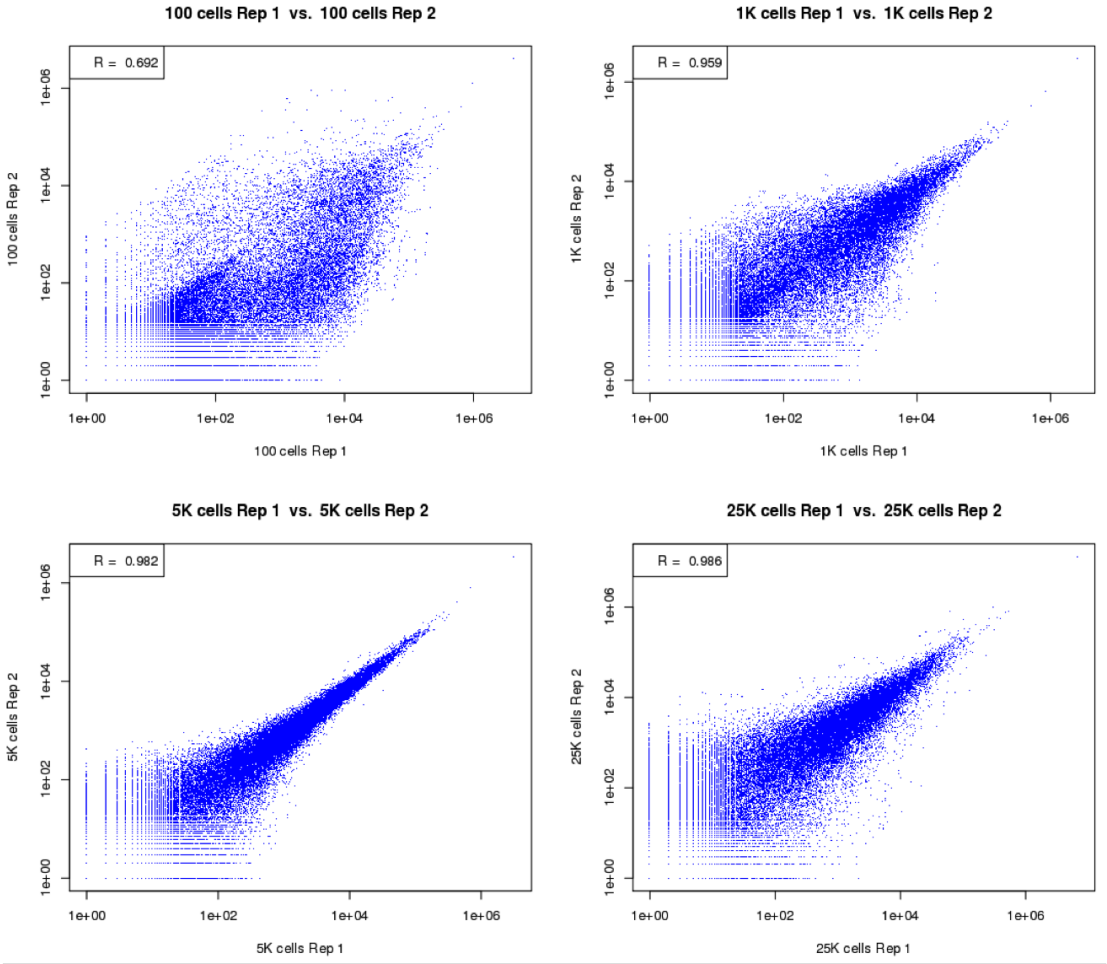

Figure 2.10: Representative scatter plots and Pearson's Correlations of replicates of increasing cell number. Here the samples with the median correlations are shown.

cell populations were also relatively reproducible and pairwise correlation values ranged from 0.9 and 0.98. Counterintuitively, 25,000 cell samples had lower than either 1,000 and 5,000 cell populations despite the larger number of cells. 100 cell populations had a very large spread and very low reproducibility.

Rarefaction studies also supported the choice of 5,000 cells as optimal (Figure 2.12). 5,000 cell samples were able to detect more exons at all sampling depths. 25,000 and 1,000 cell populations had similar levels of sensitivity with $4x10^6$ reads or less. With more reads, 25,000 cell samples were significantly more sensitive than 1,000 cell

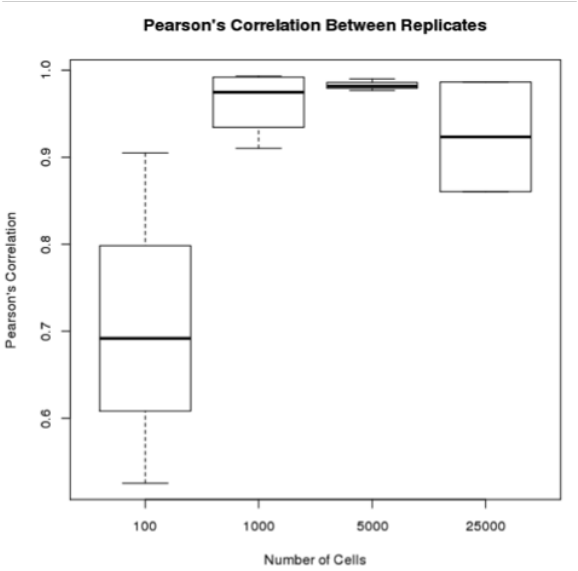

Figure 2.11: Pearson's Correlation between replicates with different number of starting cells. Samples with fewer starting cells have lower correlations between replicates and have a wider variance. Overall, 5,000 cells had the greatest reproducibility in the samples assayed here.

populations. 100 cell populations again performed poorly in this assay, and consistently detected less than half of exons detected in samples with more starting cells.

Differential expression analysis was performed on 5,000 cell populations before and after stimulation. Geneset enrichment analysis of differentially expressed genes resulted in enriched Biological Gene Ontologies associated with activation of the immune system (Table 2.3). This result further supports the legitimacy of using 5,000 cells as input for the study of rare cell populations.

2.2.3.2 Expressional Differences between T_{FH} Compartments

Pairwise Spearman's correlations of gene expression was generated between all samples. Hierarchical clustering using neighbor joining revealed distinct expression profile of CD154 expressing activated T-cells 2.13. Incredibly, the expression activated T_{FH} cells (CD4+CD154+CXCR5+) were highly correlated across multiple patients. Additionally, activated T_{FH} cells formed a separate subcluster from activated helper T cells.

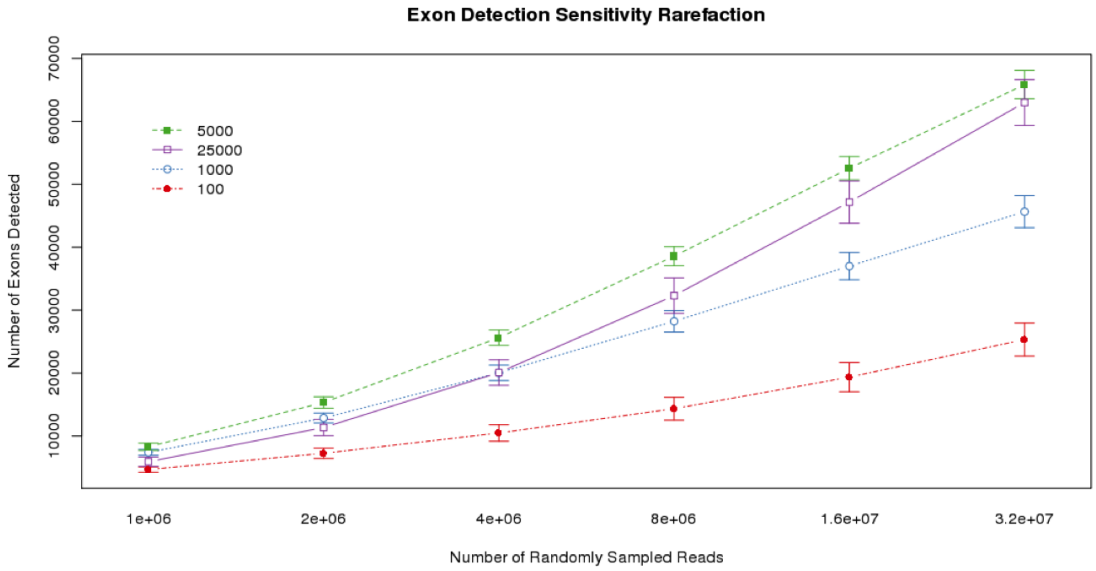

Figure 2.12: Number of exons detected as a function of sampling depth of uniquely mapped reads. Each library was sampled three times at each depth. Error bars show standard deviation. Here 5,000 cells showed the highest sensitivity.

Table 2.3: Selected enriched biological process gene ontologies from differentially expressed genes.

Selected Enriched Biological Process	False Discovery Rate
Cellular macromolecule metabolic process	8.3×10^{-15}
Regulation of cellular process	1.7×10^{-7}
Immune system process	1.9×10^{-7}
Immune response	5.8×10^{-7}
Regulation of immune system process	6.4×10^{-6}
Anti-apoptosis	0.001
Regulation of cell activation	.0003
Regulation of lymphocyte activation	.001
Regulation of leukocyte activation	.001

T_{FH} cells (CXCR5+) were significantly different from T_H cells (CXCR5-). 41 genes were up-regulated, whereas 530 were down-regulated. Up-regulated genes were highly enriched for both translational elongation and defense response to virus

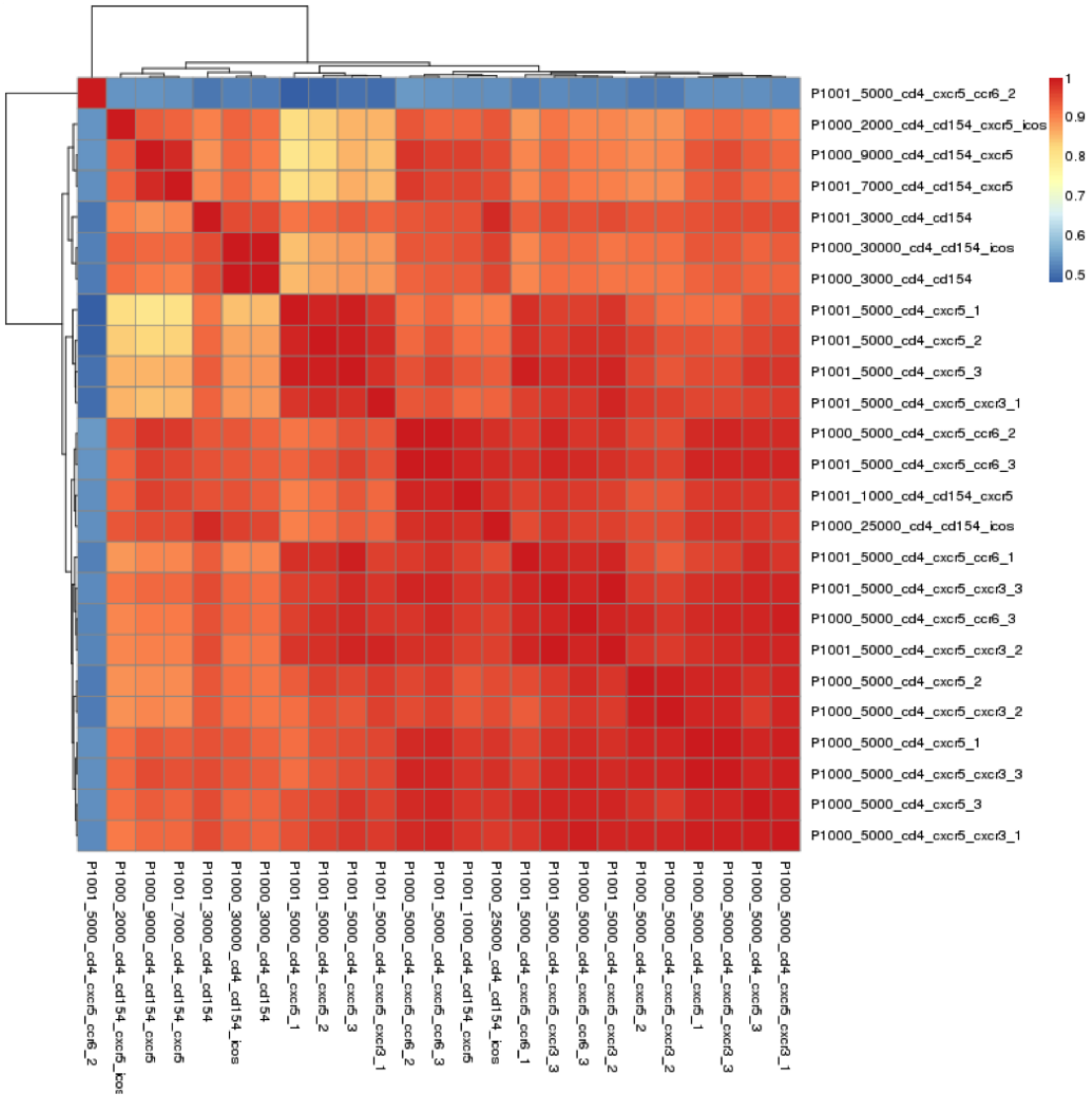

Figure 2.13: Pairwise Pearson's correlation between all samples that were hierarchically clustered using neighbor joining. Sample naming follows the convention of patient id, number of input cells and the list of surface markers selected.

gene ontologies. An increase in translational elongation suggests active production of proteins which may be associated with bursts in cellular activity. Defense response to virus is expected due to the vaccination with Fluzone. Down-regulated genes were highly enriched for endocytosis gene ontology, a biological process not typically associated with T_H cells. However, senescence and autophagy pathways were also found to be enriched

in down-regulated cells, indicating that non-T_{FH} cells were in an inactive state.

Pairwise differential expression between memory (CCR6+) and non-memory (CCR6-) samples revealed statistically significant genes. The CCR6 gene itself was identified as differentially expressed, serving as a litmus test to the effectiveness of the test. Also up-regulated and nearly as statistically significant was the gene AUTS2, a gene previously identified as being disrupted in a pair of autistic twins [146]. Studies have also shown a link between altered activation profiles for T-cells in children with autism spectrum disorders [6]. Studies have shown AUTS2 to be important for neural development and is not commonly associated with the memory subset of T_H cells [6]. AUTS2 was also implicated in GWAS studies associated with cytokine release in response to smallpox vaccination [72]. Down-regulated genes were enriched for positive regulation of NF-kappaB cascade, indicating that non-memory cells were participating in active adaptive immune response.

CXCR3 expression is a marker for the T_{H1} cells. 19 genes were up-regulated and 28 genes were down-regulated in CXCR3 expressing samples. Up-regulated genes were associated with response to virus, which is the canonical function of T_{H1} cells. All samples used in this comparison were CXCR5 expressing T_{FH} cells, recapitulating known plasticity of T_{FH} cells. Down-regulated genesets include the IL9 signaling pathway, exocytosis and DNA recombination.

Only three genes were found to be differentially expressed in ICOS expressing T_{FH2} cells. Geneset enrichment analysis of these genes (CDCA7, CST3, ZNF512) found no enrichment in any gene ontologies or pathways.

2.2.4 Discussion

In this study we optimized a cDNA synthesis method that can be performed directly on from cell lysate and applied the optimized protocol to rare subsets of circulating T_{FH} cells. We determined that 5,000 input cells were optimal for sensitive detection of expressed exons and reproducible gene-level expression estimates. Counterintuitively, using 25,000 cells as input material resulted in less reproducible results and lower sensitivity. Because all samples utilized the same amount of reagents regardless of the number of cells, it is possible that cellular debris from the larger amount

of lysate inhibited cDNA synthesis. To compensate for this effect, scaling the reagents used for cDNA synthesis to match the amount of cellular input could result in higher quality results for larger input amounts.

Smaller amounts of starting cells also resulted in poor sequencing results. 1,000 cell samples outperformed 25,000 cell samples for reproducibility of gene expression estimates. However, 1,000 cell samples also had lower sensitivity and missed many lowly expressed exons when compared to 25,000 cell samples. 100 cell samples performed poorly overall. Although suboptimal reagent to sample ratios may have affected the outcome, it is likely that limiting amount of input material and increased amplification resulted in poorer quality overall. Thus, studies which utilize very small quantities of input material, such as numerous single cell studies, should take extra care in accounting not only for single cell heterogeneity but also increased variance of cDNA synthesis with small quantities of RNA. This result mirrors results from the previous section indicating that very small quantities of input material are associated with loss of reproducibility.

Leveraging the result from our optimization experiment, we studied the difference in transcriptomic profiles over a variety of rare T_{FH} populations in two patients. These patients were vaccinated with Fluzone seven days prior to blood sampling. Although the sample size was small, we can make a few major conclusions. Although cell surface markers clearly separated each subset, the expression profiles were highly similar between all subsets. Activated T-cells clearly differentiated from non-activated T-cells according to hierarchical clustering of expression-based correlations.

We determined that CXCR5 expressing blood T_{FH} cells are a distinct subset of T_H cell. Differential expression analysis of CXCR5+ cells versus CXCR5- showed the largest number of significantly differentially expressed genes of any comparison. Up-regulation of the viral defense gene ontology confirms importance of T_{FH} cells in response to vaccination.

T_{FH} plasticity was also evident from this study. Differentially expressed genes between subsets verified expected behavior based on cell surface markers and vaccination with Fluzone. T_{FH1} subsets of cells were enriched for response to viral infection and high levels of translational activity implied active response to the vaccine. This demonstrates that T_{FH1} cells were acting in a manner similar to T_{H1} cells, whose activity is associated

with defending against intracellular bacteria and viral infection. On the other hand, T_{FH2} cells showed very little activity overall. T_{FH2} cells emulate the effector functions of T_{H2} cells, who typically respond to parasitic helminth infections. A T_{H2}-like response would not be expected following treatment with a vaccine designed to mimic a viral infection and elicit an IgG response.

One intriguing finding is the up-regulation of the gene AUTS2 in memory T_{FH} cells following vaccination. The gene has been associated with autism, alcohol consumption, attention deficit disorder and other neurological disorders [139, 68, 141]. Although studies have shown its importance for neurological development, AUTS2 is expressed across many tissues according to the Illumina Body Map project. AUTS2 is a nuclear protein thought to play a role in transcriptional repression by activation or deactivation of other transcription factors. [115, 114, 177]. The association of the adaptive immune response to a gene linked to autism is a particularly inflammatory subject.

Despite numerous studies disproving a causal link between vaccinations and autism etiology, public fears of vaccination remain [149, 32, 21]. Such fears are worsened by other studies demonstrating immune dysfunction in those with autism [116]. Some studies highlight differences in T-cell response in children with autism [6]. However, genetic studies support the hypothesis that a wide array of independent genetic factors contribute to autism spectrum disorder [59]. The identification of AUTS2 expression in memory T_{FH} cells serves as a link between neurological disorder and immune function. Indeed, AUTS2 has been associated as part of the SEMA5A regulatory network [24]. SEMA5A is an important gene both for guiding axons during development and as a modulator of the adaptive immune response [51, 24, 48].

Due to the complex interplay of cellular machinery, dysfunction in an upstream regulator may manifest in aberrant behavior in two seemingly unrelated systems. In this case autism and altered adaptive immune response may be linked by aberrant AUTS2 function or expression. This finding highlights the importance of differentiating between statistical correlations and causation.

Given the limited number of patients in this study, it is prudent to take caution when interpreting these results. However, we have demonstrated that studying rare

populations of circulating T_{FH} cells is possible using cDNA synthesis of direct cell lysate. This initial study was meant as pilot study and provides evidence that more robust large scale studies are possible using RNA-sequencing.

Chapter 3

Applications of HTS to Persistent Pathogenic Threats

3.1 HPV Integration Detection

3.1.1 Introduction

Human Papilloma Virus (HPV) is so ubiquitous that some medical professional consider to it be a part of the natural human flora [7, 64]. The DNA virus infects only keratinocytes and commonly manifests as benign epithelial tumors, such as warts [52, 29]. Although the vast majority of HPV infections do not progress to cancer, the virus has been identified as the primary cause of cervical cancer and is highly associated with other types of squamous carcinomas including cancers of the vulva, vagina, penis, anus and oropharnyx. HPV has been detected in 99.7% of cervical carcinomas, with high risk strains HPV16 and HPV18 detected in 50% and 20% of lesions, respectively [160].

Although the omnipresence of HPV within cervical cancers is worrisome, 90% of infections with high risk HPV types clear without medical intervention [55]. Of the 10% of infections that remain persistent, only a small fraction will develop into cancer [105, 106]. The disparity of HPV infection rate and the development of cancer can be partially explained by the integration of viral DNA into the host genome. Integration can drive the progression towards cancer by either over-expression of viral oncogenes or by insertional mutagenesis of host genes [65, 168].

The HPV genome encodes two viral oncogenes E6 and E7, which inactivate tumor suppressor genes p53 and pRB, respectively. The expression of E6 and E7 are normally regulated by two other viral proteins, E1 and E2. However, the integration event can often disrupt the coding sequence of E1 and E2, thus increasing transcription rates of E6 and E7 [67]. Transcription rates of E6 and E7 may also be affected by the proximity of insertion to transcriptionally active genes within the host genome[168].

Insertional location may also cause insertional mutagenesis, which can affect transcription rates or function of host genes. Notable examples have been presented in the literature with integrations occuring with the MYC proto-oncogene and the potential tumor suppressor genes ZBTB7C and CASZ1 [123, 151, 27, 34, 136].

HPV preferentially integrate within fragile sites within the human genome, however integrations have been detected over all chromosomes [151, 168]. Methods for detecting integration sites have adapted with changing technologies. Initial studies utilized bacterial cloning in conjunction with sequencing to identify junctions from DNA or cDNA [140, 8, 138]. A wide array of PCR-based assays were also developed for detection of integration from DNA, however these methods lack the sensitivity to detect all integration sites of HPV [151, 39, 90, 121]. More recently, targeted HTS and whole genome sequencing methods have identified many integration sites and elucidated the potential integration mechanism of HPV [175, 3].

In collaboration with Samuel Vohr, we developed a method for identifying HPV subtype(s), integration loci and oncogene conservation status from whole genome sequencing of cancerous or precancerous lesions.

3.1.2 Methods

3.1.2.1 Samples

The genomic DNA (40ng/μl) of twenty cervical cancer lesions were obtained from Stanford University. The samples were labeled cc1-cc20.

3.1.2.2 Sequencing Library Preparation

400 ng of genomic DNA was diluted with nuclease-free water to a total volume of 40 μl in thin wall PCR tubes. Genomic DNA was sheared using a Covaris

S220 focused-ultrasonicator. Parameters for shearing was 5% intensity, duty cycle 3, 200 bursts/second for a total of 110 seconds. Sequencing libraries were prepared from sheared DNA by first end-repairing sheared fragments, da-tailing, and sequencing adapter ligation via a semi-automated robotic protocol described in depth in Hesson et al 2010. Libraries were barcoded by 10μl of robotically-prepared sequencing library, PCR using 5μl of ABI 10x PCR buffer II, 5μl of MgCl2 (25mM), 5 μl of dNTPs (1.25mM), 5μl 10X PCR enhcancer with betaine (Epicentre), 0.5 μl inPE 1.0 forward primer, 0.5μl inPE 2.0 reverse primer, 0.5μl barcode primer and 0.25μl of Taq Titanium Polymerase in a total volume of 50μl for 8 cycles of PCR. Amplified and barcoded libraries were purified using a Qiagen gDNA spin-column and were quantified using a Bioanalyzer. Libraries were combined into two pools in equal concentrations.

3.1.2.3 Sequencing

Pooled libraries were sequenced with an Illumina GAIIx using a total of four lanes over two separate sequencing runs. Libraries were sequenced as paired-ends with 50bp reads for run 1 and 85bp reads for run 2.

3.1.2.4 HPV Typing

Paired reads were aligned to to 77 HPV reference sequences obtained from the PapillomiaVirus Episteme (PaVE) [153]. Reads that mapped to a unique reference sequence were kept. HPV type was determined by selecting the strain with the greatest coverage depth and breadth.

3.1.2.5 Identifying Chimeric Pairs

Forward and reverse reads were mapped to the UCSC human genome reference sequence (hg19) and the determined HPV strain sequence from the HPV typing step. Reads were first mapped to HPV sequences. Pairs with at least one read mapping to an HPV reference sequence were also mapped to hg19. Pairs with one end mapping to an HPV genome and the other mapping to the hg19 were considered to be "chimeric pairs". After it was discovered that many chimeric pairs were a result of reads mapping

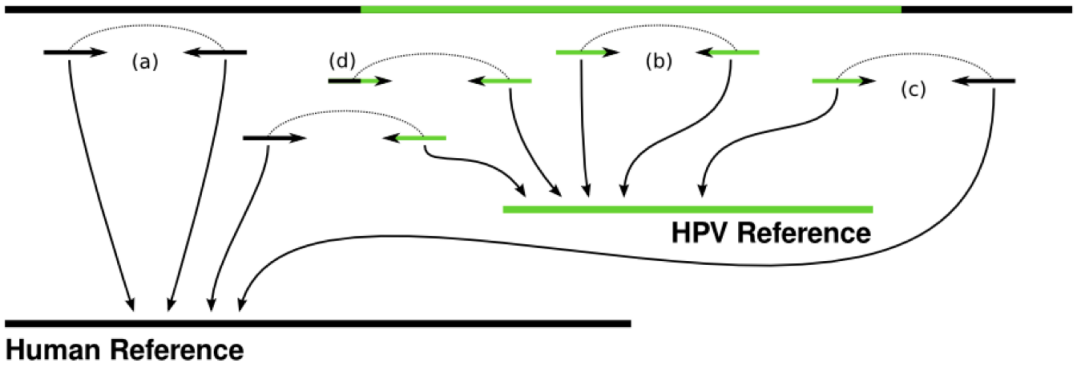

Figure 3.1: Summary of informative reads for the typing and integration detection of HPV. From the paired-end reads sequenced from a genome with a viral integration, we expect to see pairs that map only to the human reference (a), pairs that map only to the viral (b), chimeric pairs that map to both (c), and spanning reads that map to neither (d).

to low complexity regions of the HPV genomes, a low complexity filtered was added to remove artifactual chimeric pairs.

3.1.2.6 Identifying Integration Loci

Chimeric pairs were used to identify integration loci. Integration loci were expected to have a 5' and 3' integration boundary. Due to background level of discordant reads, an integration loci was not considered to be high confidence unless at least two non-duplicate reads were found to span the integration boundary. Reads were considered to be duplicate if the first base of the forward read and the last base of the reverse read were identical. Ideally, multiple reads would span both the 5' and 3' end of a viral integration. Orientation of integrated viral sequence was determined using the orientation of chimeric reads. Insert size range was also used to determine the approximate size of deletions in host genome as a result of integration.

3.1.2.7 Identifying Spanning Reads

After integration loci were determined, exact integration locations were determined by searching for "spanning reads". A spanning read is a single read which spans the actual integration location. A portion of a spanning reads maps to both the HPV

and human reference sequence. All reads that did not map to the hg19 human reference or the HPV sequence were checked.

Spanning reads were identified using a seeded-extension methodology. An index of 20mers was constructed for the detected HPV strain reference sequence and the region spanning the 5' and 3' integration boundaries. 20-mers from the 5' and 3' end of each read were matched against the index. Reads with a 20-mer matching against the indexes were extended one base at a time until bases no longer matched the reference sequence.

3.1.3 Results

Twenty cervical cancer samples were sequenced using an Illumina GA2x. Samples were sequenced with shallow depth with coverage ranging from 0.001x to 0.61x. Median coverage was 0.28x.

HPV types were determined for 17 out of twenty samples (Table 3.1.1). The three samples for which HPV strain could not be determined had coverage levels far below the median (0.04x, 0.002x and 0.001x).

High confidence integration loci were determined for 6 out of 20 samples. Of the detected integration loci, four of which belonged to HPV16, one to HPV18 and the last to HPV18. The integration event with the largest amount of evidence was observed in cc20.

The cc20 sample had multiple chimeric pairs and spanning reads on both the 5' and 3' integration boundaries (Figure 3.1.1). The integration resulted in a 20bp genomic deletion. 3,363 basepairs of the HPV58 viral genome were integration assuming that the detected ends cover the shortest possible integration of the linearized circular HPV genome. However, the integrated region did not contain viral oncogenes E6 and E7. Instead, 4 other coding proteins were preserved (E2, E4, E5 and L2). Despite the high confidence of integration boundaries detected, sequencing results indicated the presence of full length HPV genome. In addition to the high confidence integration detected in cc20, 9 other chimeric pairs were identified. However, none of these chimeric pairs were supported by more than one read.

We were also able to identify 5' and 3' integration boundaries in cc10. HPV16

Table 3.1: Sequencing statistics and HPV type determined for twenty cervical cancer samples.

Sample ID	Total Reads	Reads mapped to HPV	Chimeric Pairs	HPV Type
cc01	3,933,178	57	0	HPV45
cc02	4,779,660	210	0	HPV59
cc03	3,975,184	245	0	HPV18
cc04	5,135,128	339	0	HPV34
cc05	3,815,810	167	0	-
cc06	778,334	22	0	-
cc07	1,258,254	9	0	-
cc08	5,047,174	1,543	15	HPV16
cc09	19,920,888	1,085	8	HPV16
cc10	26,645,688	1,274	8	HPV16
cc11	19,355,944	967	6	HPV45
cc12	17,986,434	965	4	HPV45
cc13	19,625,450	3,088	17	HPV18
cc14	18,461,386	2,563	41	HPV16
cc15	23,974,886	1,877	3	HPV16
cc16	25,380,454	3,496	15	HPV16
cc17	26,634,494	1,885	8	HPV16
cc18	14,705,872	800	1	HPV16
cc19	16,026,746	1,318	4	HPV34
cc20	9,939,742	3,218	15	HPV58

was detected in opposite orientation to the reference. Chimeric pair analysis was able to identify a single integration locus on chromosome 2. However, spanning read analysis indicates the presence of multiple short integrations.

Integrations were often found near coding genes. For example, the HPV58 integration in cc20 was identified within the intronic region of ZFAT and the HPV16 integration of cc10 was found upstream of PREPL.

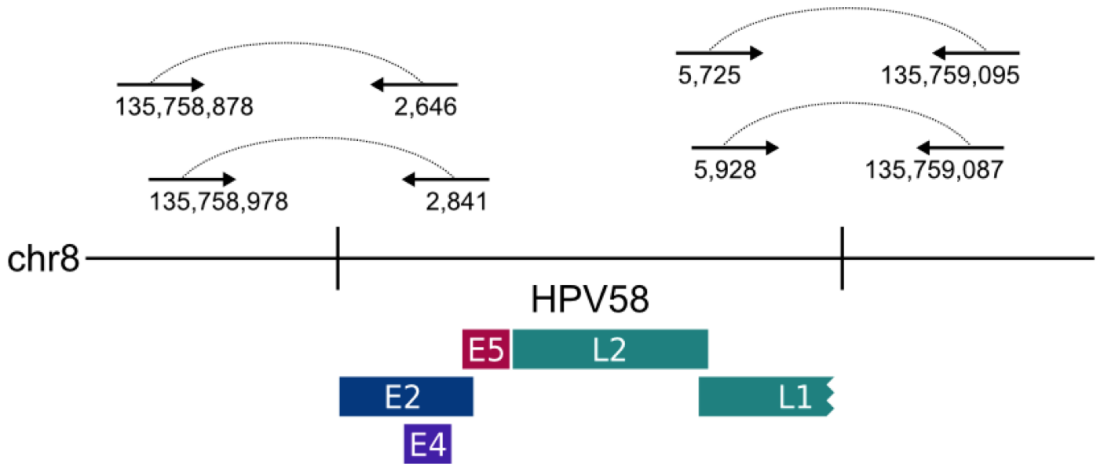

Figure 3.2: Integration details for sample cc20. Four chimeric pairs were found to span the integration of HPV58 into the host genome. The oncogenic genes E6 and E7 are not present in this integration.

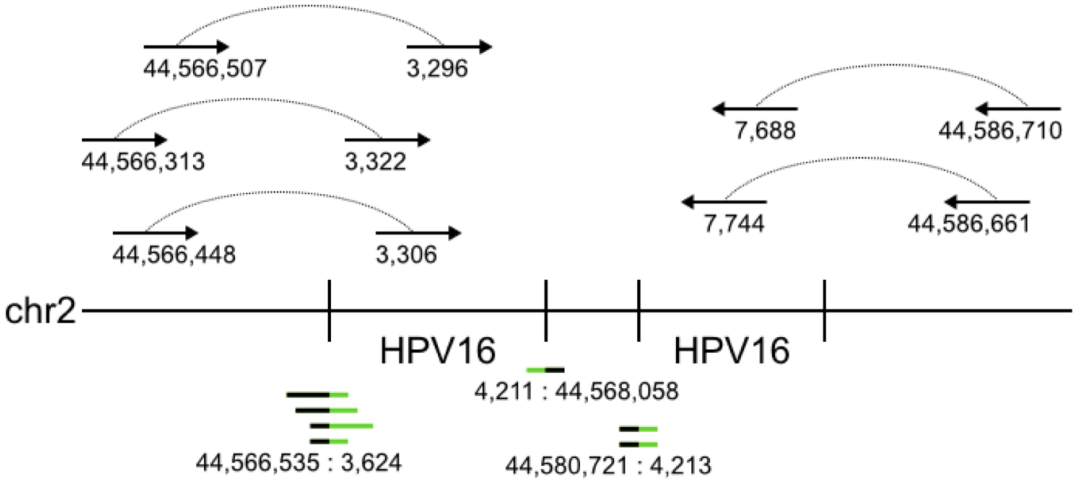

Figure 3.3: Integration of HPV16 into cc10. Chimeric pairs (top) suggest a single integration site. However, spanning pairs (below) suggest multiple small integration events.

3.1.4 Discussion

Identification of HPV subtype, preservation of oncogenes in DNA samples that contain integrated DNA of viral origin and genomic location of the integration event

may have a clinical impact for the progression of precancerous lesions. In this study, we were able to characterize all these features from whole-genomic sequencing data.

In 17 out of 20 cervical cancer samples, we were able to identify the HPV strain from which the integrated DNA originated. Most of the integrated DNA originated from high risk strains (HPV16, HPV18, HPV45, HPV58, and HPV59) [107]. We also identified HPV34 in two cervical cancer samples– a strain not typically considered to be high risk.

We were also able to identify integration loci in 6 samples. Identified integration loci were found to be more complex than a simple linear insertion within the host genome. For example, in cc20 the integration event was characterized with high confidence. In the integration event, we observe the insertion of a 3 kb portion of the viral genome. However, coverage over the viral genome indicates that the entire length of the viral genome was present in the sample. The disagreement between the detected integration and the HPV genomic coverage could be a result of an endogenous active infection with HPV or multiple integration sites. In cc20, we also observed multiple chimeric pairs, however none were supported by more than one read.

The integration event detected in cc10 was similarly complex. Although chimeric pair analysis identified a single 5' and 3' integration boundary, spanning reads showed the possibility that multiple short integrations occurred within the integration locus. Recently, Akagi et al 2013, proposed a new model for HPV integration within the host human genome [3]. In the proposed model, an integration event may insert multiple copies of an HPV-Host hybrid sequence within the human genome. The model may explain the observed complex piece-wise integration observed in cc10. In addition, multiple copies of an integration may contribute to our ability to detect integration boundaries despite relatively low levels of coverage.

Integration events detected in this study usually occurred near or in genes. The integration in cc20 occurs upstream of ZFAT, a zinc finger that likely binds DNA and functions as a transcriptional regulator involved in apoptosis and cell survival. It is possible that the integration altered the transcription of ZFAT and contributed towards the progression towards cancer. In cc10, the integration event occurred within the an intron of the PREPL gene. Although the gene, itself is not associated directly with

apoptosis, it is possible that endogenous expression of the gene drives up the expression of integrated viral genes.

Many features of integration may have a clinical impact for the analysis of precancerous cervical lesions. We can identify these features using HTS from whole genomic DNA. Recently, targeted methods for detection of integration events have been described [175]. However, as the cost of sequencing continues to plunge, it is only a matter of time before the cost of enrichment surpasses the cost of sequencing. When this occurs, general bioinformatic methods for the detection of HPV or other viral integrations can be applied to whole genome sequencing samples.

3.2 Identifying Drug-resistance in Clinical Isolates of *N. gonorrhoeae*

3.2.1 Introduction

Infection with *Neisseria gonorrhoeae* causes gonorrhea, a sexually transmitted disease estimated by the World Health Organization to have 106 million new cases in 2008 alone [172]. *Gonococcal* infections have a long history of developing resistance to antibiotics, first gaining resistance to penicillin in the 1970's. In response to penicillin-resistant strains, gonococcal infections were treated with fluoroquinolones, such as ciprofloxacin, as a second line of defense. Emergence of fluoroquinolone-resistant strains of *N. gonorrhoeae* caused the CDC to renege its use as a recommended treatment of gonorrhea in 2007. [112]. A familiar story emerged with the third generation of antibiotic used to treat gonorrhea. Cephalosporins such as ceftriaxone, were used as a primary treatment for just five years before the CDC recommended use of a combination of cephalosporins in addition to another antibiotic such as azithromycin [20].

It comes as no surprise that there exist strains of *N. gonorrhoeae* resistant to every antibiotic currently used to treat gonorrhea [47]. Although a single strain resistant to all antibiotics has yet to be discovered, the risk of such a strain emerging is high due to horizontal gene transfer in *N. gonorrhoeae*. Co-infections of multiple resistant strains may result in super resistant strains due to horizontal gene transfer. It is vitally important that gonococcal infections are properly treated with the correct

dosage of an antibiotic to which the infectious strain are susceptible.

Identification of drug-susceptibilities of isolated and cultured *N. gonorrhoeae* cells is important to the treatment of gonorrhea. The standard for identification of drug-susceptibilities within *N. gonorrhoeae* isolates are disk-diffusion and Etest susceptibility tests based on culture. Although these can be effective measures, they typically require 20-24 hours of time to culture [20]. More immediate results may improve the quality of treatment and prevent the unnecessary use of ineffective antibiotic treatments.

Sequencing assays have the potential to determine drug susceptibilities without culture. As sequencing technologies improve, the turnaround time for sequencing assays has drastically decreased. Targeted real time sequencing assays have already been developed that can predict fluoroquinolone resistance [44]. These assays are rapid and inexpensive, but only assay short regions previously identified to confer drug-resistance. Thus, new assays can be developed to test for resistances to all drugs.

Whole genome sequencing is becoming a viable alternative to culture or targeted sequencing assays. Exponential decrease in sequencing cost may soon make sequencing cheaper that either culture or targeted sequencing. With reduced sequencing costs come reduced sequencing time. Benchtop sequencers such as the Illumina MiSeq or Ion Torrent PGM offer turn-around times under 6 hours. Third generation sequencers based on nanopore sequencing can potentially reduce sequencing time to hours. High throughput sequencing methods have been applied to detect antibiotic resistance to cefixime resistant strains [50].

In collaborative work with Robert Shelansky, we demonstrate the utility of whole genome sequencing for the discovery of resistance-imparting variants. Forty clinical isolates, for which the minimum inhibitory concentration (MIC) of ciprofloxacin are known, were subject to high throughput sequencing. By using variant detection, phylogenetic clustering and enrichment within resistant populations, we are able to recapitulate known drug-resistance-imparting variants. Additionally, we report new loci potentially associated with drug-resistance and identify some variants that may cause hyper-resistance to ciprofloxacin.

3.2.2 Methods

3.2.2.1 Samples

DNA extracted from 40 clinical isolates and their corresponding MIC values were gratiously provided by Magnus Unemo, Department of Clinical Microbiology, rebro University Hospital. Sample collection and MIC determination are described elsewhere [44]. The 40 isolates were separated into three categories describing their resistance based on their MIC values. Samples with a MIC less than 0.0064 mg/L were considered susceptible to ciprofloxacin. Samples between 0.0064 mg/L and 0.25 mg/L were considered to have reduced susceptibility to ciprofloxacin. Finally, samples with a MIC greater than 0.25 mg/L were considered to be resistant.

3.2.2.2 Sequencing and Alignment

Sequencing libraries were prepared robotically. Samples were pooled and sequenced in one lane of an Illumina GAIIx. Demultiplexed reads were mapped to *N. gonorrhoeae* FA 1090 reference build obtained from UCSC. Alignments were performed using Bowtie2 with default settings [81]. Only pairs mapping concordantly to a unique genomic position were considered for downstream analysis.

3.2.2.3 Variant Calling

BAM alignments were used as input into FreeBayes with default parameters with the exception of ploidy which was set to 1. Variant calls that passed filters were annotated for function using VEP [99, 73]. A custom database was built using genomic sequence and gene annotations for *N. gonorrhoeae* FA 1090 downloaded from the UCSC microbial browser for use with VEP [137]. Variants were annotated by location relative to the nearest gene as either upstream, downstream or within a gene. Amino acid changes as a result of nonsynonymous variants were also annotated.

3.2.2.4 Phylogenetic Tree Building and Sample Clustering

A pairwise similarity score was measured between all samples. Similarity score was determined by summing the number of nonshared variants between two strains.

Pairwise similarity scores was used as input to a standard unweighted pair group method with arithmetic mean (UPGMA) to produce a phylogenetic tree. The R package, Agnes, was used to build the tree with default parameters [92, 60].

Samples were assigned clusters based on their proximity within the phylogenetic tree by setting a distance threshold. The distance threshold was set such that the maximum number of resistant clones were within a cluster without including any sensitive clones (Figure 3.4).

3.2.2.5 Collapsing Variants to Quinolone Resistance Determining Regions (QDRD)

Individuals variants were collapsed to Quinolone Resistance Determining Regions (QDRDs). For this study, a QDRD was considered to be the set of nonsynonymous variants within a protein coding gene.

3.2.2.6 Enrichment Analysis

Here we test for the enrichment of a feature in the resistant population or the susceptible population. Our enrichment method assigns a score between -1 and 1, with positive values indicating an enrichment in resistant clusters and negative values indicating enrichment in susceptible clusters. Here features were either nucleotide variants or sets of variants collapsed into QDRDs. The final enrichment score is described by the equation

$$E(x) = E_R(x) - E_S(x) \tag{3.1}$$

where $E(x)$ is the final enrichment score of feature x. $E_R(x)$ and $E_S(x)$ describe the enrichment of a variant within resistant or sensitive sets of clustered isolates. Each cluster of isolates contributes equal weight to the enrichment score of feature x.

$$E_R(x) = \frac{\sum_{C \epsilon R} E_C(x)}{n_C(R)} \tag{3.2}$$

Here the enrichment, $E_R(x)$,within resistant clusters, R, is equal to the sum of enrichment values within a cluster, C, normalized by the number of clusters within the resistant population, $n_C(R)$. A similar equation describes the enrichment within sensitive clusters.

Enrichment within a cluster, $E_C(x)$ is defined as the proportion of isolates within a cluster harboring the feature x.

$$E_C(x) = \frac{\sum_{i \epsilon C} I(x)}{n_i(C)} \tag{3.3}$$

where $n_i(C)$ is the number of isolates within cluster C and $I(x)$ is an indicator function equal to 1 if feature x is present in isolate i and 0 if it is not.

Enrichment values were determined for the set of all features. A gaussian distribution was fit to the resulting distribution of enrichment scores and Z-scores are converted to P-values to determine significance.

3.2.3 Results

Whole genome sequencing of DNA extracts of cultured clinical isolates resulted in sufficient coverage for variant calling. Mean coverage was 57x across all samples, with a minimum coverage of 5x and a maximum coverage of 167x.

Phylogenetic trees built using a pairwise similarity metric based upon the number of shared and unique variants between samples revealed that ciprofloxacin resistance do not share a common ancestor (Figure 3.4). Resistance isolates are spread out across the tree and can be highly related to susceptible clones. However, clusters of highly related resistant isolates correspond with ancestral subpopulations of resistant *N. gonorrhoeae*.

Enrichment analysis of variants resulted in 703 variants with a p-value less than 0.01. Within these enriched variants, we identified known resistance-imparting variants in gyrase A (gyrA), the known binding partner of ciprofloxacin. Variants at amino acid positions 91 and 95 of gyrA were discovered with p-values of $1.5\text{x}10^{-5}$ and $3.1\text{x}10^{-5}$, respectively. Variants at position 91 included a serine to phenylalanine mutation (S91F), whereas variants at positon 95 included an aspartic acid to glycine (D95G) and an aspartic acid to alanine (D95A) mutation. A third mutation at the 95 position, which caused an aspartic acid to asparagine (D95N) mutation was not correlated with ciprofloxacin resistance.

QRDR-level analysis identified mutations within parC as being enriched in resistant populations with a p-value of 1x10-4. Point mutations compressed within the

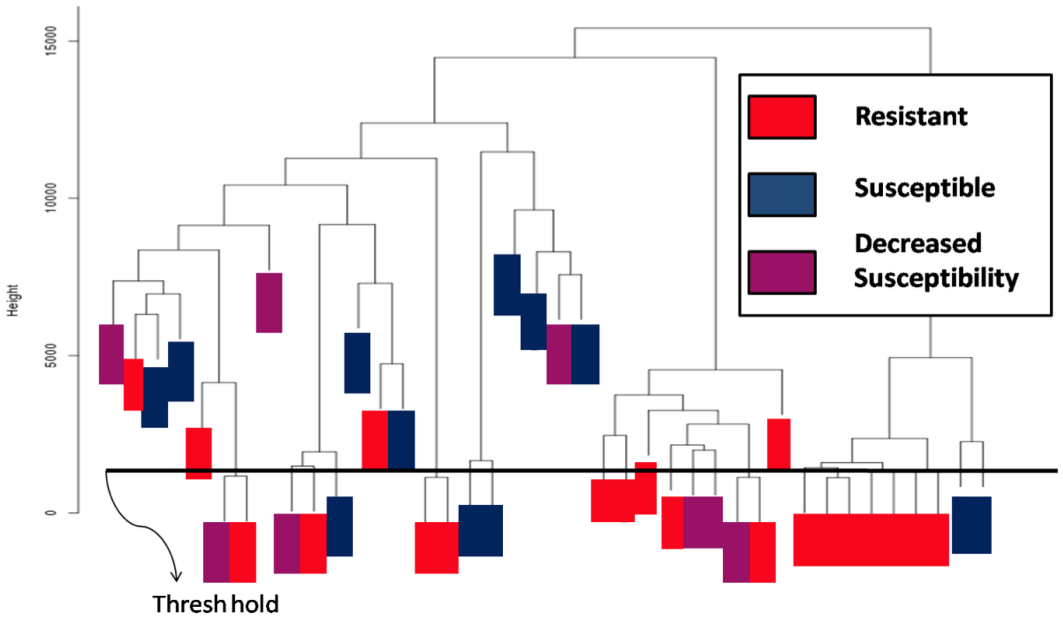

Figure 3.4: A phylogenetic tree built using UPGMA: Resistant isolates (Red), susceptible isolates (Blue), and decreased susceptible isolates (purple). A clustering threshold was set to group genetically similar resistant strains together while ensuring such that as many resistant strains occupied the same cluster without inclusion of a sensitive strain.

parC QRDR included serine to asparagine and arginine mutations at position 87 of the gene (S87N and S87R). In one isolate with moderate resistance (MIC = 0.75 mg/L), a S88P mutation was observed. In a super-resistant isolate (MIC >= 32 mg/L) glutamic acid was mutated to glycine at position 91 (E91G). The most enriched single nucleotide variant within resistant isolates was a mutation from aspartic acid to asparagine at position 86 (D86N)– a variant not reported in existing literature.

A study on quinolone resistance in *Pseudomonas aeruginosa*, a bacteria known to cause pneumonia, sepsis and other infections, have identified ABC transporter proteins to be associated with drug-resistance [181]. We also observed the enrichment of variants within many integral membrane transport proteins, including a variant annotated as a "putative drug-resistance protein" in the UCSC microbial genome browser.

A human gene, metnase, is also known the bind directly to ciprofloxacin. We identified variants within the insertion element IS1016 transposase which is a homolog

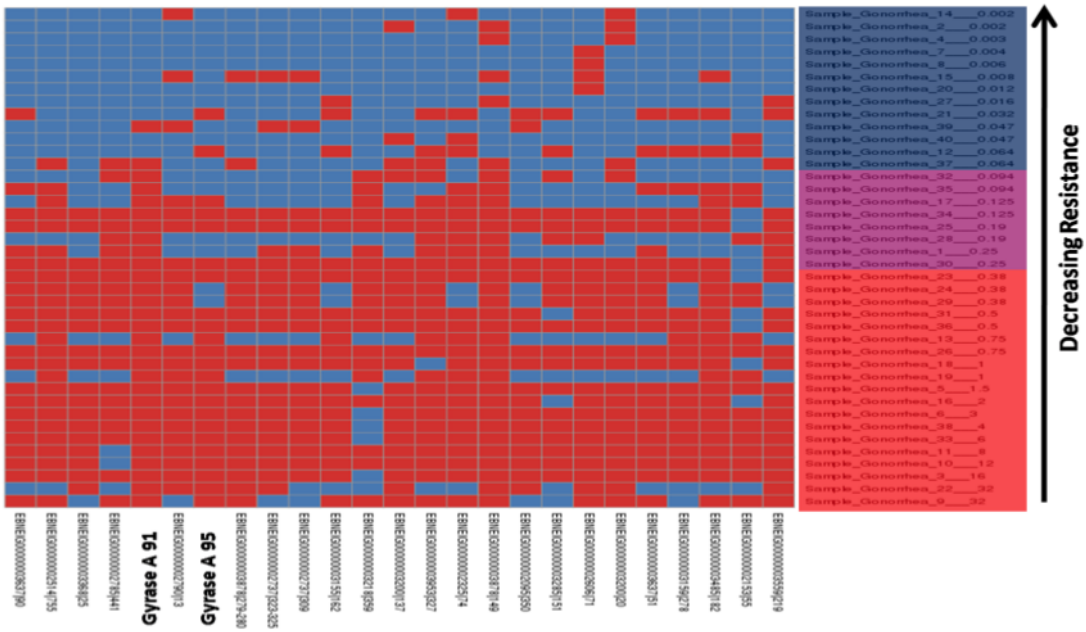

Figure 3.5: Top 25 most enriched SNVs using enrichment method. Each column represents a nonsense mutation in a gene. Each row represents an isolate. Rows are sorted by increasing resistance. Blue cells indicate a wild type base, while red cells indicate a nonsynonymous change.

of the human metnase [171]. Finally, we identified a point mutation in A-lyase 25 which occurs only within two super-resistant isolates.

3.2.4 Discussion

Phylogenetic analysis revealed that drug-resistant isolates did not originate from a single common ancestor. Drug-resistant strains can be found in almost all lineages indicating that any susceptible strain has the potential to evolve drug-resistance given favorable circumstances. Ciprofloxacin acts as a mechanical obstruction which prohibits the function of gyrase A [33]. Small genetic modifications can result in conservation of enzymatic function while negating the effectiveness of ciprofloxacin. This result coincides with results from the directed evolution studies presented in the previous section.

In this study, we observe that resistance-imparting mutations convert amino acids with relatively large and polar residues into amino acids with smaller functional

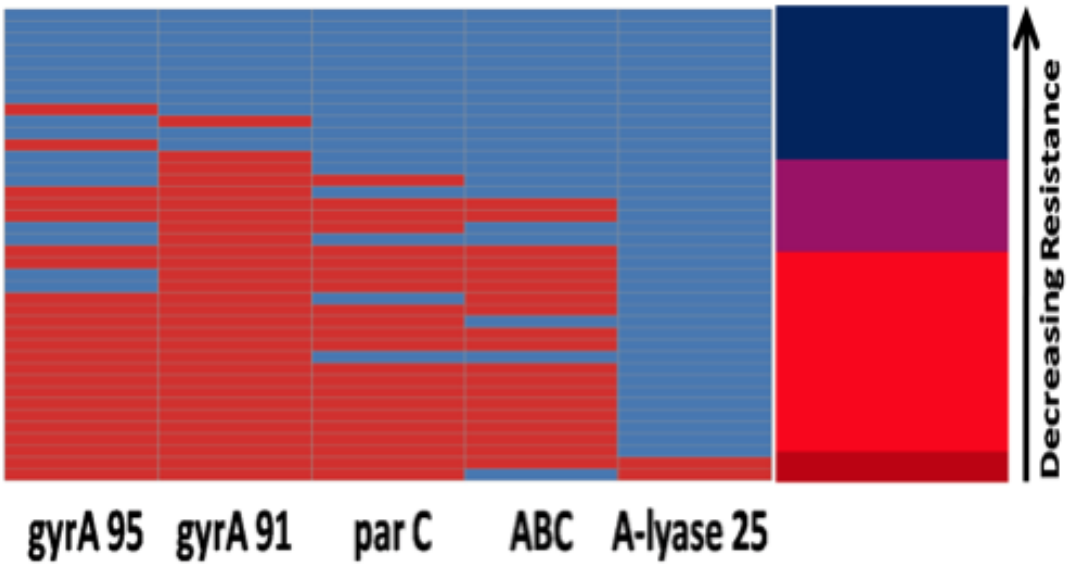

Figure 3.6: Biologically relevant QDRDs selected by our enrichment method.

groups. For example, D95A and D95G mutations associated with drug-resistance both replace a larger polar aspartic acid residue with smaller glycine and alanine residues. Additionally, the D95G mutations were also enriched for much higher levels of resistance to ciprofloxacin (MIC >1) and could be due to the smaller size of glycine compared to alanine. However, a D95N mutation which replaces aspartic acid with similarly sized arginine does not confer drug-resistance. Although the size of the residue is actually increased in the S91F mutation, the conversion of a highly polar residue to a highly non-polar residue could drastically alter the binding pocket of gyrase A, thus conferring drug-resistance. Additionally, the S91F mutation was requisite for the development of drug-resistance and was found in all drug-resistant samples.

When examining the results of our QRDR-level enrichment, we were also able to recapitulate known resistance-imparting variants in the parC gene. Mutations in parC has been previously identified as a marker for fluoroquinolone resistance [54]. parC is a subunit of topoisomeraise IV and shares significant homology with gyrA in the ciprofloxacin binding region. Of the variants that comprise the parC QRDR, D86N was the most enriched within the resistant population. However, isolates harboring this

mutation cover a wide range of MIC values. Two variants, S87R and E91G, in parC were found in only super-resistant strains. Like in gyrA, the E91G mutation replaces a large acidic amino acid with a smaller glycine and may have similar functionality.

Because parC and gyrA are direct binding partners of ciprofloxacin, it is no surprise that mutations in these genes confer drug-resistance. By identifying these two genes alone, we are able to classify clinical isolates as drug-resistant or drug susceptible. However, we are unable to regress the range of drug-resistance from these two genes alone. Although measurement error of the MIC may account for some of the variability, it is likely that regulatory alterations, transport modifications or competitive binding of ciprofloxacin to other protein products explain the remainder of variability in the observed MIC values.

We observed 703 values with a p-value of 0.01 or less that were highly enriched in resistant strains. Of these, there are a few notable examples that could explain the remainder of the variability of MIC values. ABC transporting proteins known for the exportation of molecules from bacteria are also known to be related to drug-resistance. We also observe an enrichment of variants within this class of genes in our experiment. Mutations in these genes could enhance drug removal from the bacteria allowing them to reduce the internal concentration of ciprofloxacin.

We also observe an enrichment of mutations in an insertional element, IS1016 transposase. The human homolog, metnase, is known to harbor ciprofloxacin sites. It is also possible that mutations in this gene sequester ciprofloxacin, thus reducing the efficacy of the drug. Finally, we observe a single variant in A-lyase 25 that is found only in super-resistant strains. Although lyase mutations are not typically associated with increased drug-resistance, it may be relevant to study the mutation in a follow up study.

By sequencing whole genome isolates and searching for enriched variants and QRDRs, we were able to recapitulate known drug-resistance-imparting mutations in addition to the discovery of many potentially significant variants. However, improvements to the methodology can be made. First, the methods used to generate phylogenies were crude and more precise methods utilizing de novo assembly of strains and more advanced methods could be utilized to generate more accurate phylogenetic trees. Ad-

ditionally, de novo assembly methods could potentially identify plasmid sequences that also contribute to drug-resistance [50]. Clustering of isolates required previous knowledge of MIC values for each isolate, but general methods could yield better results. Finally, we did not implement a regression based method to predict MIC value based on the variants we observed. A regression-based method would be necessary to translate sequencing based diagnostics to the clinic. However, the prediction of drug-resistance may not be completely amenable to logistic regression or naive bayes methods due to the complex interplay of variants.

Recently, other studies have utilized similar methodologies to identify drug-resistant variants across multiple drugs and species [173]. Their methods are similar to ours, but theirs may be more powerful due to the increased sample size. Ultimately, an increase in the number of samples will increase the specificity of drug-resistance determinant identification. The initial step of developing methods to both predict drug-resistance from whole genome sequencing samples and to predict drug-resistance-imparting mutations will only improve with time.

3.3 Evolution of Drug-resistant *E. coli* in Connected Microhabitats

Pathogenic resistance to antibiotics is a worldwide problem [85]. This is highlighted in the previous section, where clinical isolates of *N. gonorrhoeae* were found with varying levels of resistance to the antibiotic, ciprofloxacin. Strikingly, resistant isolates did not share a common ancestor and were often closely related to drug-susceptible isolates. This indicates that evolution of drug-resistance is not a rare event. In the study, we were able to identify many potential variants that confer drug-resistance. However, the study was limited in its ability to probe how such mutations occur and spread within an individual bacterium.

Other studies address the issue by performing directed evolution studies by growing bacteria in low concentrations of drug [117]. Although these studies are can be successful for identifying genetic variants associated with drug-resistance, they poorly model the conditions by which organisms encounter in the human body. Drug concen-

trations are not constant through the human body. Rather, concentrations of drugs are spatially heterogenous through different organs and tissues of the human body [35]. This spatial heterogeneity can accelerate evolution if subpopulations of pathogens are free to migrate between microenvironments of differing drug concentration [10].

Evolution within spatially heterogeneous interconnected microenvironments may be accelerated for two reasons. First, a mutation may be more likely to occur in an environment where the drug concentration is high enough to exert a stress on the organisms, but not high enough to kill the organism. Thus organisms will be able to replicate and develop mutations conferring increased fitness in the presence of stress. These newly adapted organisms can then migrate to microenvironments of higher drug concentrations and continue to adapt to even higher levels of stress. Secondly, competition for nutrients in a microenvironment with low stress may also drive the evolution of organisms to develop drug-resistance, thus allowing it to utilize nutrients in a high stress environment that were previously unavailable due to high concentrations of drug.

To test this hypothesis, collaborators at Princeton University developed a microfabricated chip of interconnected microenvironments. A ciprofloxacin gradient was spread across the chip allowing *Escherichia coli* to travel between compartments of varying drug concentrations. The study was able to greatly accelerate the time required to develop drug-resistance. With high throughput sequencing, we were able to identify causal mutations that conferred resistance.

3.3.1 Methods

3.3.1.1 Directed Evolution

A microchip was fabricated on a silicon wafer. The chip contained 1,200 hexagonal wells, 10 μm deep with 200 μm long sides. Each well was connected to neighboring wells via six microchannels 200 μm long, 10 μm deep and 10 μm wide. Wells at the edge of the chip contained nanoslits 100nM deep to allow the flow of nutrients and antibiotic into the chip into the interior of the array.

Two syringe pumps were used to create a drug gradient across the chip. The first syringe pump flowed LB media through the top section of the chip. The second

syringe flowed LB media with ciprofloxacin at a concentration of 10 mg/ml through the bottom portion of the chip. Both pumps were flowed under 30 μl/hour to form a gradient of ciprofloxacin across the chip.

An *E. coli* K12 strain was transformed to express GFP via a plasmid, but was otherwise wild type. Approximately 106 bacteria was inoculated by drilling a hole in the center of the chip, pipetting 2 μl of culture into the chip and sealing the chip with a silicon elastomer. Bacteria were cultured for 48 hours. Cell growth was monitored optically. A 470 nm LED was used for illumination and a Cannon 5d CCD camera was used to obtain fluorescent intensities.

Multiple experiments were performed to ensure the emergence of drug-resistant bacteria was not due to pre-existing mutants in the inoculated population. First, bacteria was cultured in a 96 well format with increasing concentrations of ciprofloxacin in each well. Second, decreasing quantities of cells were used for inoculation into the microfabricated chip.

3.3.1.2 Sequencing and Bioinformatic Analysis

Six samples were subjected to high throughput sequencing on the Illumina GAIIx platform using 110x90bp reads. DNA was extracted from a population of wild type cells (WT) and three populations of cells subjected to directed evolution within the microfabricated chip (DG1, DG2 and DG3). For DG1 and DG2, two technical replicates were sequenced from the same starting DNA (labeled DG1-a, DG1-b, DG2-a and DG2-b). WT, DG1 and DG2 were also sequenced using a Roche 454 sequencer.

Resulting short reads were aligned to an *E. coli* K12 reference genome (NCBI accession NC_000913.2). Illumina short reads were aligned using bowtie with default parameters with the exception that only uniquely mappable reads were reported [cite bowtie]. 454 short reads were aligned using Roche's proprietary gsMapper algorithm using default parameters.

Resulting BAM alignments were used as input to the samtools pileup command. Positions where greater than 98% of reads supported a non-reference base were reported. Variants associated with drug-resistance were identified by looking for variants present only in drug-resistant samples.

Variants within in the gyrA gene resulting in nonsynonymous amino acid changes were examined in solved crystal structures. The crystal structure of gyrA which both originated from *E. coli* and bound to ciprofloxacin has not been resolved. However, a crystal structure of gyrA originating from *Staphylococcus aureus* bound to ciprofloxacin has been elucidated (PDB ID: 2XCT). To identify the proximity of mutated amino acids to the binding site of ciprofloxacin, a structural alignment comparing a crystal structure of gyrA from *E. coli* (PDB ID: 3ILW) to the ciprofoxacin-bound structure from *S. aureus*. A jFATCAT rigid structural alignment with default parameters was used on the PDB website was performed [Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9) 739-747.] Amino acid position in *E. coli* structure was mapped to the nearest aligned amino acid in the *S. aureus* structure.

3.3.2 Results

Drug-resistant mutants emerged in under 5 hours from 10^6 inoculated wild type E. coli. Following inoculation, competition for nutrients drove *E. coli* towards the perimeter of the chip where nutrients entered via nanoslits. Bacteria eventually fixed at a "goldilocks point", a region of the chip where normal and drugged media converge and stress gradients were maximal. Upon development of drug-resistant bacteria, motile bacteria migrated to areas of higher drug-resistance and eventually invaded the entire chip.

Inoculation of wild type *E. coli* within the chip led to a complex growth dynamic most likely due to competition between wild type and drug-resistant cells for nutrients. Reinoculation of a chip with drug-resistant *E. coli* led to logarithmic growth, thus indicating reduced overall fitness of wild type cells.

Two methods ensured that emerging resistant bacteria was not already present in the inoculation population. First, decreasing numbers of cells in starting inoculations were used to seed the chip. If small subpopulations of *E. coli* already contained drug-resistance conferring mutations, experiments seeded with lower numbers of starting cells should not result with growth of resistant bacteria. However, even with just 100 starting cells, drug-resistance eventually emerged (figure 2). Secondly, bacteria were cultured in

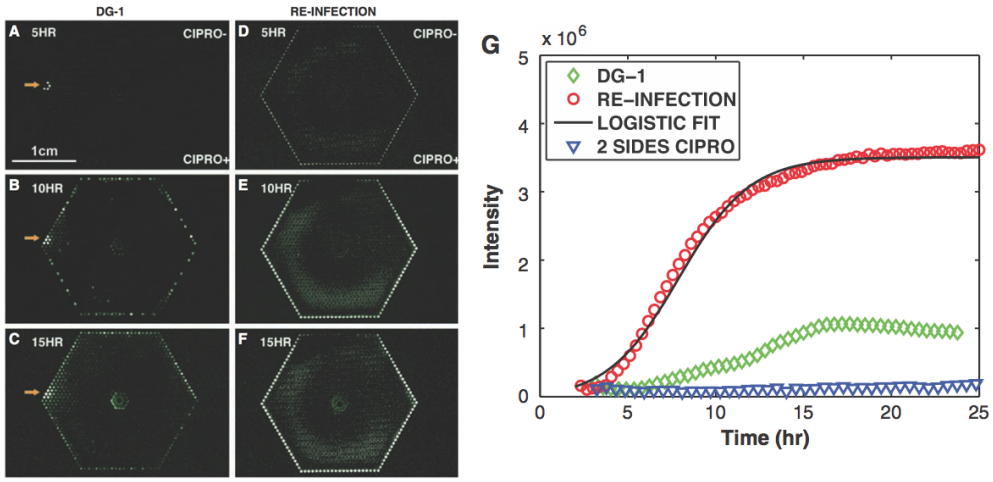

Figure 3.7: Growth and migration following inoculation with wild type *E. coli* (A-C) or evolved drug-resistant *E. coli* (D-F). A. Drug-resistant mutants emerge after 5 hours. Goldilocks point is denoted by orange arrow. B. Drug-resistant mutants spread to periphery of the chip. C. Mutants continue to spread across the entire chip. (D-F) Growth of drug-resistant *E. coli* results in faster growth and spread across the chip. G. Total GFP intensity is measured in wild type inoculate (DG-1) versus drug-resistant inoculate (RE-INFECTION). Re-infection with drug-resistant *E. coli* results in logarithmic growth whereas wild type inoculate grows slowly. Inoculating wild type into a chip with ciprofloxacin flowing from both sides (2 SIDES CIPRO) results in no growth.

96 well, each with increasing ciprofloxacin concentration. If drug-resistant *E. coli* were already present in the inoculation population, growth would have occured in wells where ciprofloxacin concentration was above the minimum inhibitory concentration (MIC). Such a phenomenon was not observed.

The genetic events that conferred drug-resistance were identified through high throughput sequencing with an Illumina GAIIx and Roche 454. Samples sequenced with Illumina resulted in 50x coverage per sample whereas samples sequenced on with 454 resulted in approximately 6x coverage per sample. Thirteen single nucleotide variants were discovered across all samples indicating this sample was slightly divergent from the strain used to generate the reference sequence. Four mutations were observed only in drug-resistant cells.

Gyrase subunit A (gyrA) is the protein targeted by ciprofloxacin. The SNP

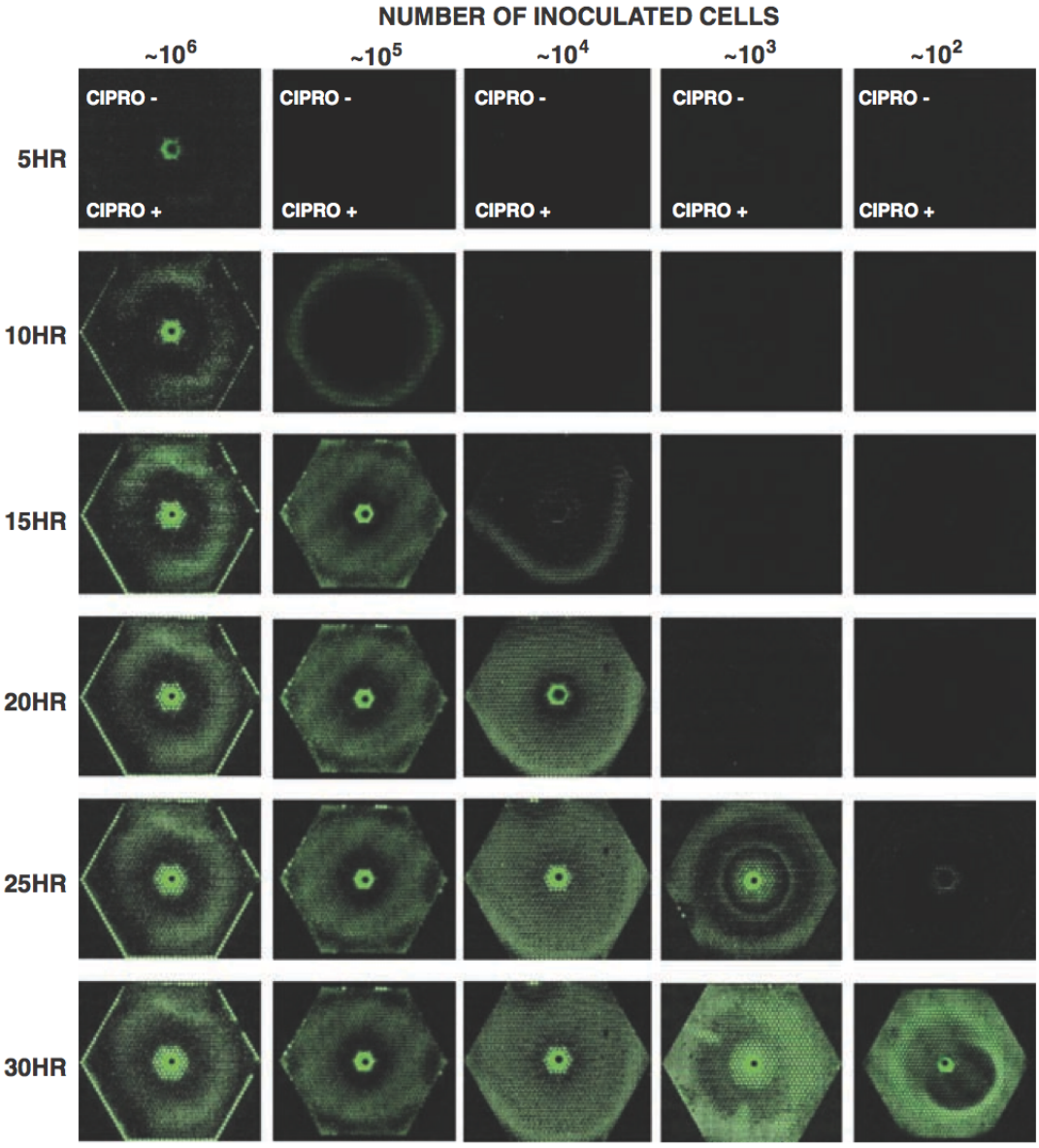

Figure 3.8: Growth and emergence of drug-resistant *E. coli* as a function of time and number of cells in initial inoculation.

occurred at base 2,337,183 in the *E. coli* k12 reference. This mutation occurs in codon 87 resulting in a missense mutation from an Aspartic Acid to a Glycine. The crystal structure was examined for clues to the function of the identified mutation. By structurally aligning a crystal structure of gyrA from *E. coli* with a structure from *S. aureus*

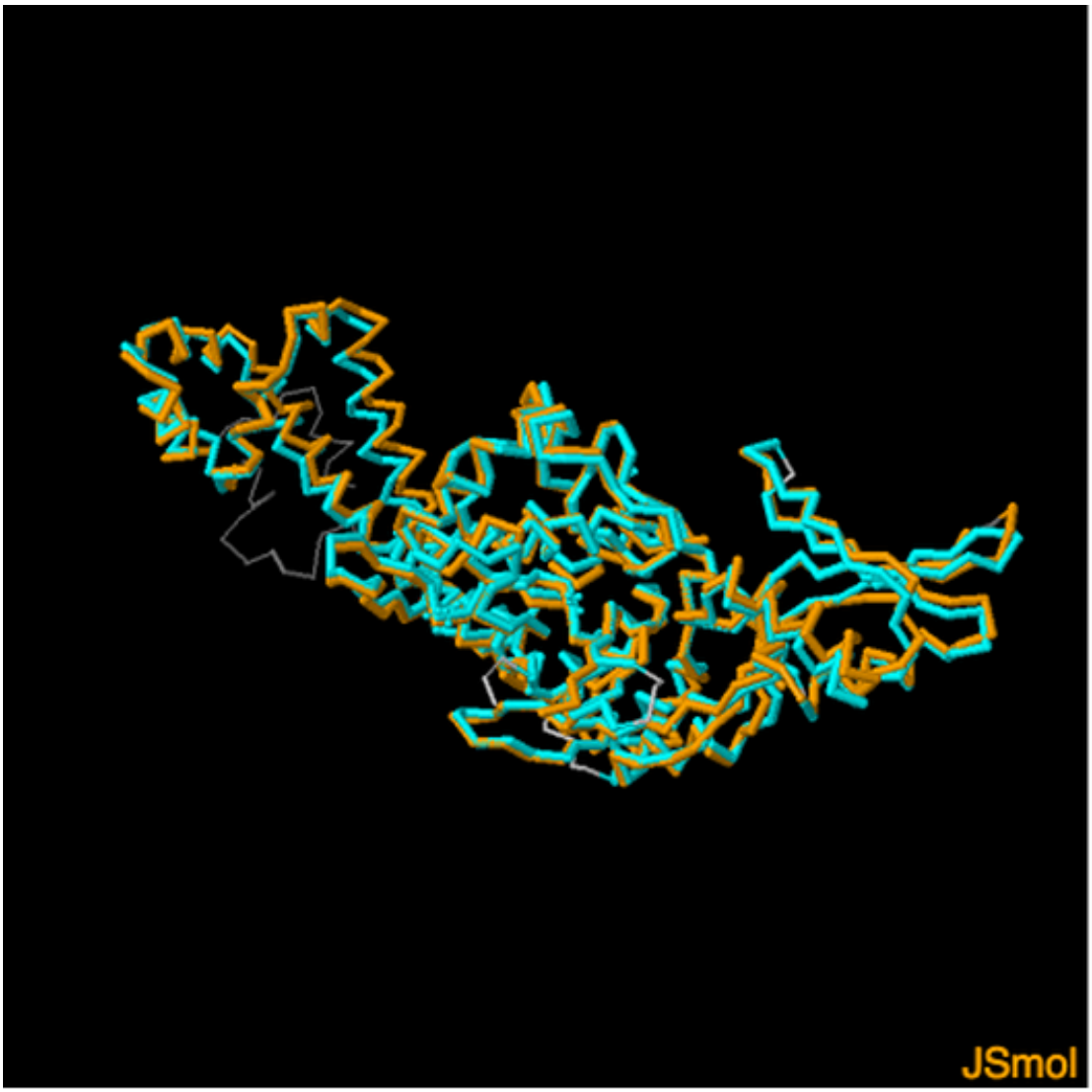

Figure 3.9: Structural alignment of *E. coli* K12 gyrase A (3ILW) and *S. aureus* gyrase A (2XCT) shows high conservation between two crystal structures.

containing the binding site of ciprofloxacin, we were able to map the position of the nonsynonymous amino acid change relative to the binding site of ciprofloxacin. We found that the position of the mutated amino acid was directly adjacent to the binding site of ciprofloxacin. Because the mutation converted the relatively large residue of aspartic acid to the small residue of glycine, it is likely that the mutation prevents the obstruction of enzymatic function by ciprofloxacin.

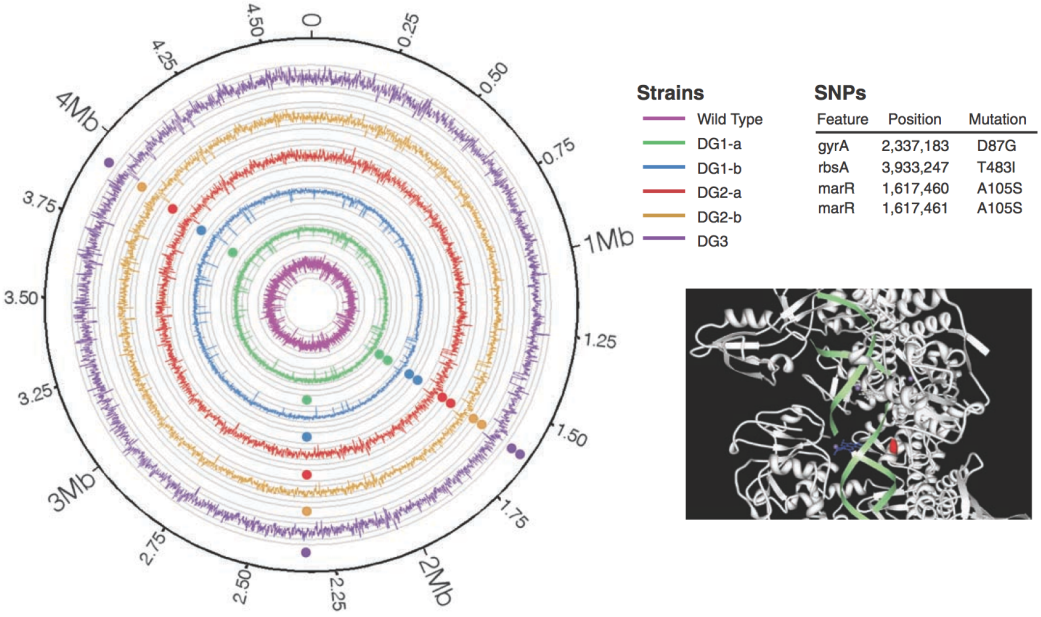

Figure 3.10: Summary of sequencing results. (Left) Circos plot showing the coverage (lines) and variants (circles) found in DG samples but not in wild type. (Top Right) A table of SNPs showing genomic position and amino acid change. (Bottom Right) Crystal structure of gyrase A from *S. aureus* (2XCT). Nicked double stranded DNA is highlighted in green, ciprofloxacin in blue and resistance-imparting variant in red.

Ribose Import ATP Binding Protein (rbsA) is a subunit of the ribose ABC transporter (RbsABC) responsible for binding ATP. This SNP occurs at base 3,933,247 (figure 2) and causes a missense mutation in codon 483 changing a threonine to an isoleucine. ABC proteins have been implicated in the export of other antibiotics such as erythromycin, tylosin, and macrolide out of prokaryotic cells (Fath 1993).

Two remaining mutations were observed in the repressor of the multiple antibiotic resistance operon (marR). Both mutations were missense mutations and were located at base 1,617,460 and 1,617,461 respectively. It is possible that these mutations affect the ability of *E. coli* to deal with antibiotics through increased expression of antibiotic resistance genes.

3.3.3 Discussion

The methods utilized in this study allow researchers to identify changes caused by drug resistance. The method simulates spatial heterogeneity found in nature– particularly in the human body. It is particularly powerful for studying the genetic changes that bacteria must undergo in isolation in order to develop drug-resistance.

The mutations in gyrA, rbsA and marR have all been previously identified as necessary changes to manifest bacteria resistant to clinically relevant concentrations of quinolones such as ciprofloxacin [62]. It is surprising that all four seemingly functional mutations are able to fix within a population in under 10 hours. It is very unlikely that all these mutations occurred at once and previous studies indicate they occurred sequentially. It is also probable that all four of these mutations were necessary for the drug-resistant bacteria to thrive at the high concentrations of ciprofloxacin used within this study.

In addition to the spatial heterogeneity mimicked by the microchip in this study, other levels of heterogeneity may also affect the development of resistant bacteria. Temporal heterogeneity caused by taking relatively large doses of antibiotic with low frequency may also drive the evolution of bacteria towards drug-resistance [9, 154]. Missed dosages exacerbate the problem by creating low levels of antibiotic. Periods of low antibiotic concentration are analogous to wells within our microchip that have low concentrations of drug. Research has shown that when bacteria grow in concentrations above the MIC but below mutation prevention concentration (MPC), drug-resistant mutations occur most frequently [180].

Drug-resistance-imparting mutations could be prevented by combining multiple antibiotics and this approach has been suggested not only for antibiotic treatments, but also for cancer therapy [look up some citations for this]. However, even with joint therapies, the emergence of drug-resistant bacteria is still possible due to not following physician recommendations when taking antibiotics.

Also, strategies used to prevent de novo resistance-imparting mutations in bacteria do not take into account horizontal gene transfer. In addition to genetic changes that can occur to generate quinolone resistance, numerous plasmids have been discovered that also contribute to drug-resistance [62]. Drug-resistance via plasmid is also a

problem for a wide array of antibiotics in a wide array of pathogens.

In the end, a small change in the type of antibiotics or dosage regimen may not be enough to overcome the increasing levels of drug-resistant pathogens affecting the world. A combination of personalized medicine, dosage and immune stimulation will be necessary to prevent complete domination of antibiotic resistant bacteria.

Chapter 4

HLA Typing from RNA-Seq data

4.1 Introduction

HLA typing is a key diagnostic for the diagnosis of disease and matching of donors and recipients on organ transplantation. There have been numerous attempts to identify HLA alleles using high throughput sequencing techniques [12, 37, 42, 56, 83, 86, 124]. The majority of these techniques exploit the high throughput of modern sequencing machines in order to generate high-quality low-cost genotypes. To accomplish this, the amplified HLA genes of many individuals are multiplexed and sequenced in a single run. These techniques can thus generate many genotypes at a relatively low cost due to economies of scale. However, these methods are also prone to failed amplification due to mismatches in the primer region.

As the ever-dropping cost of high throughput sequencing ushers in the age personalized medicine, it may one day become feasible to produce large amounts of sequence data for each patient individually without the need for amplification by PCR. Although genomic sequencing is useful for the detection of genetic predisposition to disease, RNA sequencing provides a snapshot of the current state of the cells in the body. If RNA sequencing experiments becomes routine in clinical diagnostics, it is feasible that many tests that once required specific assays could be combined into a single workflow requiring only a single test and followed by a multitude of bioinformatic analysis.

However, HLA genotyping is a difficult task due to number of unique alleles

that have been discovered in the human population. The MHC is the most polymorphic region in the genome [40]. Balancing selection has generated many thousands of possible alleles to ensure that the population has sufficient diversity to combat constantly evolving threats from other organisms [26]. Due to the extreme polymorphism in the region, short reads can map to many possible alleles making it difficult to assign genotypes to samples.

HLAforest is a utility that predicts HLA genotypes from high throughput sequencing data. It utilizes a unique knowledge-based approach that allows for accurate prediction of allele. It was built and tested on data generated from RNA-seq experiments, however, it has also been successfully applied to whole genome and exome sequencing experiments. Although other typing methodologies from RNA and exome sequencing have been developed [13, 164, 88, 36], HLAforest stands alone as the only open-sourced alignment-based prediction tool that can generate predictions at all resolutions from RNA-seq reads.

4.2 Methods

HLAforest is a knowledge-based technique that is designed to extract maximal evidence from each short read. Depending on read length, a singular read may contribute varying amounts of information used to identify an HLA allele. HLAforest exploits the natural hierarchy of HLA nomenclature in order to extract maximal information from each sequencing read. It does so using a series of steps. A filtering step extract reads that map to any allele in the IMGT database. Then all possible alignments to all alleles in the IMGT database are generated for each read. The evidence each read can contribute is then divided amongst all alleles using a novel tree building and pruning methodology. HLAforest also considers PHRED quality scores generated by the sequencer while distributing evidence, which contributes to its incredible robustness against sequencing error.

In addition to generate predictions, HLAforest is capable of generating more accurate expression estimates for HLA genes. It can also be extended to any gene that contains a high level of sequence homology amongst its alleles.

4.2.1 Alignment

Accurate alignment of short reads is critical to the quality of predictions generated by HLAforest. Alignments can be performed with any alignment algorithm, although for this work the suffix-tree based bowtie was utilized [82] . Two rounds of alignment are performed in order to generate the final set of alignments from reads to a database of HLA alleles. A reference index for bowtie2 was generated from version 3.10.0 of the IMGT HLA nucleotide database [133]. Null alleles were excluded from the set of known alleles, as these sequences sometimes contain intronic sequences and intronic reads from RNA sequencing experiments can artificially inflate scores of those alleles.

The first round of alignment filters out reads that do not align to the HLA database. Reads that aligned at least once to any allele sequence were kept for subsequent realignment to the reference database of HLA alleles, this time outputting all possible alignments to all alleles. Default parameters for bowtie were utilized except for trials where only exact alignments were allowed.

After the alignment step, each read contains a set of all possible alignments to all HLA alleles. In cases where a read aligns to the same allele in multiple locations, the most likely alignment is kept. The most likely alignment is considered to be the alignment with the lowest sum of sequence mismatch qualities (SMMQ). The SMMQ is generated by taking the sum of PHRED scores at mismatching positions within the reads. A default PHRED quality score of 20 is given to each inserted or deleted base in the read alignment.

4.2.2 The HLA Reference Tree

HLA Nomenclature divides types hierarchically, where each allele is defined at the gene, allele group (2-digit), protein (4-digit), nucleotide (6-digit), and intronic (8-digit) resolution. A tree reflecting all alleles in the IMGT database can easily be built by adhering to the nomenclature. In this implementation, a node with no biological meaning roots the tree and is known as the "root node". The descendants of the root represent genes, the highest level of classification as proposed by the nomenclature. The descendants of gene nodes are allele groups, whose descendants are proteins, etc. All

leaves in the tree represent alleles represented in the database, though some non-leaf nodes may also be represented in the database.

4.2.3 Alignment Trees

Alignment trees store all the alignment information generated during the alignment stage. An alignment tree can be thought of as a subtree of the reference tree where only nodes representing alleles aligned to the read are kept. Although they are a subtree of the reference tree, alignment trees are built using the HLA nomenclature of aligned alleles.

Alignment trees are important abstractions because they allow read mapping qualities to be distributed evenly despite unequal representation in the database.

4.2.4 Building Alignment Trees

A tree is built using a read and a set of aligned alleles (Figure 4.1.A-B). Aligned alleles are first split apart into the hierarchical units defined in the reference section. Once allele names are split apart into tiers, tree building begins with the creation of a root node. Aligned alleles are processed one at a time. The aligned allele is incorporated into the root node by first checking if the gene level node exists as a child of the root node. If no child exists, a node for the gene is created. The allele group, peptide, nucleotide and intronic levels are incorporated in a similar manner to the gene level. For each level, the tree is first checked if an equivalent node exists underneath the parent and creates it if it does not. For the allele group, the gene-level group is checked for equivalent a node equivalent to the allele group level of the allele. This process reiterates until all levels of all aligned alleles have been incorporated into the tree.

After this process completes, an alignment tree will describe all possible alignments from a read to alleles in the database. However, because alignment quality information has not been incorporated into the tree, all leaves are equally likely to be the allele from which the read originated.

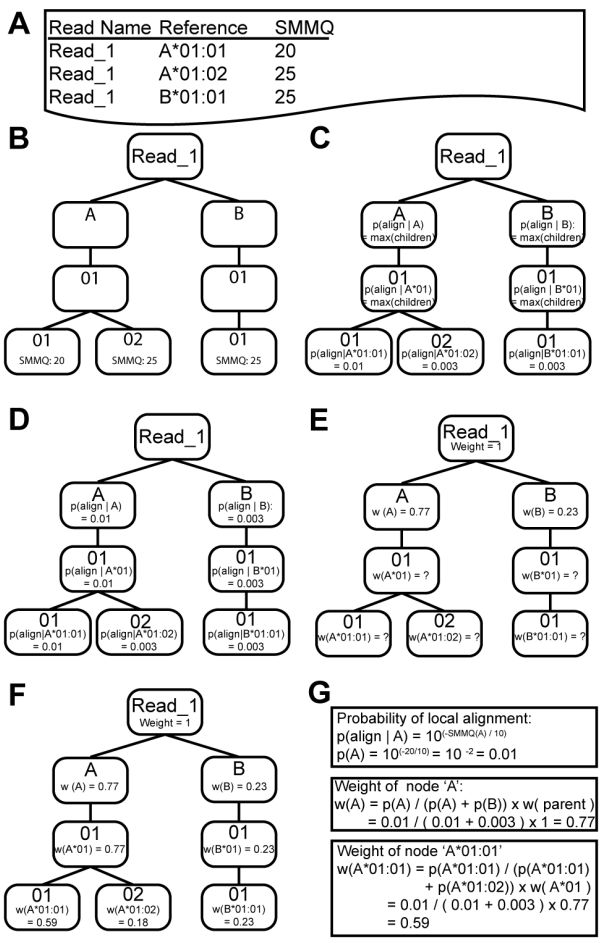

Figure 4.1: Method for building a weighted read tree. Given a set of alignments (A) for a single read, a tree is built such that all possible alignments are leaf nodes (B). Gene, allele group, peptide, nucleotide and intronic digits are represented as nodes on the tree. Sum of mismatch qualities (SMMQs) are converted to alignment probabilities for leaf nodes (C). Probabilities are then distributed upwards such that the probability of a parent node is equal to the maximum probability of its children (D). Weights are distributed downwards in such that the weight of a node is dependent on the local probability of the node and the weight of the parent child (E & F). Equations used for generating probability of an alignment and weights of example nodes are outlined (G).

4.2.5 Weighting Nodes in Alignment Trees

HLAforest considers all reads when generating predictions of HLA genotypes. However, not all reads contribute equally to the prediction. Some reads may map to

regions with high sequence homology resulting in many alignments for an individual read. Alternatively, some reads may map to a single allele uniquely. HLAforest uses a weighting scheme that distributes evidence based on the quality of alignment, the number of aligned reference sequences and their position in the read tree. Thus, reads that align to a single reference sequence has more influence on the final predicted genotype than a read which maps promiscuously.

Weighting nodes is accomplished first by calculating the local alignment probability at each leaf node or node with corresponding alignment(Figure 4.1.B). The local alignment probability can also be thought of as the probability that any mismatches, insertions or deletions in an alignment were due to sequencing error (Figure 4.1.C). The local alignment probability, p_{leaf} can be calculated with the equation:

$$p_{leaf} = -10^{-\sum^{i} \frac{q_i}{10}} \tag{4.1}$$

where q represents the PHRED quality value corresponding to mismatched base at position i in the read. For deletions in a read sequence, a default PHRED quality value of 20 is applied. These local alignments are then pass upwards to parents nodes such that

$$p_{node} = max(p_{children}). \tag{4.2}$$

Distributing probability this way ensures that the most likely alignment is considered for each level along the alignment tree. For example, if a read aligns better to A*01:01 than to A*01:02, this methodology ensures A*01 alignment probability is not unfairly penalized by the lower scoring alignment to A*01:02. Finally, alignment probabilities are converted to weights, which allows the evidence to be summed to produce a score for each allele. The weights are determined at each level following the equation:

$$w_{node} = w_{parent} \times \frac{p_{node}}{\sum p_{siblings}}, \tag{4.3}$$

where w_{node} is the final weight of a node, w_{parent}, is the weight of the parent node, and$p_{siblings}$ are the local alignment probabilities of nodes that share the same parent as the node (Figure 4.1.E-F). In this implementation, the weight of the root

node is set to 1 for all reads, however this weight can be modified to reflect variable read lengths or overall sequencing quality.

This weighting scheme ensures that the sums of weights at any particular tier (gene, allele group, protein, etc) sum to at most 1. This allows us to predict the genotype at any resolution by summing the weights of nodes in a tier. After all read trees have been built and weighted, the trees are fully prepared for use in genotype prediction.

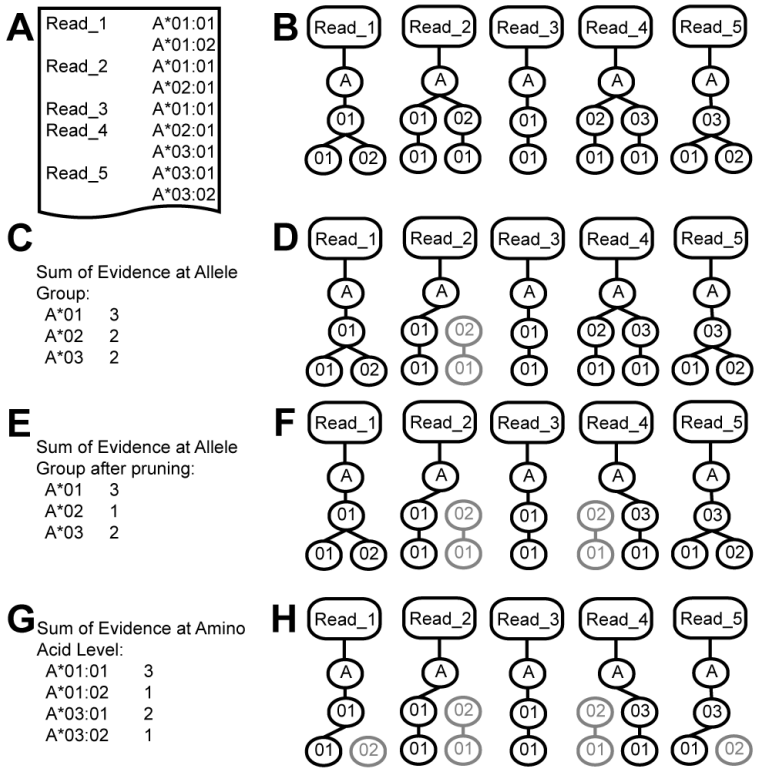

Figure 4.2: An example of the top-down pruning algorithm. Given a set of reads and their alignments (a), read trees are built for each read (b). The evidence for each allele group is determined by taking the sum of the evidence of all allele groups represented in the trees (c). Here it is assumed that the weight of each node is 1. The allele group with the maximum evidence is assumed to be the primary allele group for each gene and edges in trees containing the primary allele group are pruned (d). After pruning, the trees are reweighted and the evidence for each allele group (e). The second highest scoring allele group is then considered to be the minor allele. Read trees are then repruned such that only edges supporting the primary or secondary allele group remain (f). The process repeats itself iteratively until the most likely leaf nodes remain (g and h).

4.2.6 Type Prediction

HLAforest predicts HLA types using a top down, greedy, iterative pruning algorithm (Figure 4.2). The top down approach means that the highest level type are predicted first, e.g., the algorithm first picks the allele group of each gene before it picks the protein or nucleotide levels. The greedy selection of nodes allows HLAforest to pick alleles that explain the majority of the data. Finally, by pruning trees to contain only branches of most likely alleles, reads can have a larger impact on higher resolution predictions.

The algorithm begins by predicting a primary and secondary allele at the allele-group resolution for each gene. A "primary" type is defined here as the allele with the most evidence in an experiment and a "secondary" type is defined as the allele with less evidence. The primary type is chosen by summing the weighted evidence of all read trees at the allele-group level. For each gene in the reference database, the allele-group with the highest score is chosen as the primary type (Figure 4.2.c).

After the primary allele-group is chosen, a temporary set of pruned read trees is generated to predict secondary allele-groups. This step prevents reads originating from the primary allele-group from contaminating predictions of the secondary allele group. To accomplish this, read trees containing alignments to the primary allele-group are pruned such that all branches that do not correspond to the primary allele-group are pruned. Secondary allele-groups are then predicted by summing the weights of nodes at the allele-group level over all read trees (Figure 4.2.d). The top two scoring allele-groups after this step are considered to be the primary and secondary allele-groups.

Following the choice of the primary and secondary allele-groups, read trees are pruned such that only edges corresponding to the primary or secondary allele-groups or their children remain. Read trees containing alignments to both primary or secondary allele-groups, the allele-group with the lower weight is pruned. In the event that a read tree contains both primary and secondary allele groups of equal weight, the read tree is assigned to a random allele group. Read trees are then reweighted.

Homozygosity at the allele-group level is accounted for by checking the evidence assigned to the secondary allele-group. If the secondary allele-group contains less than 5% of the total evidence distributed to a gene, it is considered to be homozygous. This

threshold was determined empirically (data not shown). For homozygous allele groups, read trees are reweighted to removed the spurious secondary allele-group.

Peptide level predictions are generated using similar methodology used to predict allele-groups. The weights of peptide-level nodes are summed over all pruned read trees. The peptide node with the largest sum of weights for the primary and secondary genotypes are chosen. For samples where the allele-group was determined to be homozygous, the algorithm calls a primary and secondary genotype at the peptide-level if the evidence for the secondary peptide-level prediction exceeds 5%. For samples heterozygous at the allele-group level, only one peptide-level prediction is made.

The peptide-level prediction process reiterates for the nucleotide and intron level until all genotypes are predicted to the highest resolution.

4.2.7 Simulations

Simulations were used to test the predictive performance of HLAforest. For each simulation, two alleles were randomly chosen for all genes in the IMGT nucleotide reference. Reads were generated using ART, a utility that generates realistic high throughput sequencing reads using error profiles determined from empirical datasets [58]. Paired end 100 nt reads were generated with a mean fragment size of 250 nt and fragment size standard deviation of 50 nt. 25x coverage was generated for each allele with an inferred substitution rate of $8.7\text{x}10^{-5}$. 5,000 simulations using all genes were performed.

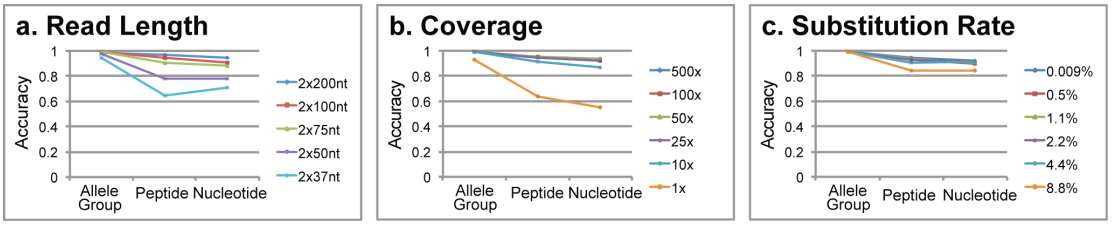

Figure 4.3: Simulation results showing the effect of read length (a), sequencing depth (b), and substitution rate (c) on average accuracy of HLA-A, HLA-B, HLA-C, and HLA-DRB1.

Simulations are particularly useful in comparison to real data because conflating factors can be tested in isolation. Here we tested the effect of varying sequencing

depth, sequencing error and read length on performance of HLAforest. To test the effect of these variables, two alleles were chosen for the genes HLA-A, HLA-B, HLA-C and HLA-DRB1 as these genes displayed the most diversity in the IMGT nucleotide database (Figure 4.3).

Simulations tested read lengths typically generated with Illumina sequencing runs. Paired reads with lengths 37 nt, 50 nt, 75 nt, 100 nt and 200 nt were tested. Sequencing error was likewise tested over realistic error ranges with inferred substitution rates of .009%, 0.5%, 1.1%, 2.2%, 4.4% and 8.8% tested. Coverage level in real data is dependent on sequencing depth, so a wide range of sequencing coverages were chosen to see at which depths HLAforest would be useful. Coverage levels of 1x, 10x, 25x, 50x, 100x and 500x were tested. Each variable was tested independently and the parameters other than the one being tested were kept constant. Default parameters were the same as used in the simulations using all genes. 5,000 simulations were run for each parameter.

To test the effect of alignment sensitivity, HLAforest was run twice for each simulation. One run allowed only exactly matching reads to be used as input, while the second run allowed all aligned reads to be used as input.

HLAforest predictions made from simulated data were considered to be accurate if the true allele (determined during the selection step of simulation) was consistent at the allele-group, peptide, nucleotide or intron level. Accuracy was assessed over all levels present in the true allele. If the true allele was only typed to the peptide level, accuracy was not assessed at the nucleotide or intron level. For example, the reference allele HLA-A*02: 90 is only typed to the peptide level and accuracy at the nucleotide level cannot be determined for this allele. Of the 8,631 sequences present in the IMGT database: 87 (1%) are typed to the allele group level; 6,014 (70%) are typed to the peptide level; 2,308 (27%) are typed to the nucleotide level; and 222 (2%) are typed to the intron level.

4.2.8 Real Data

Three datasets representative of typical use cases of HLA typing using RNA-Seq data were chosen based on availability of sequence-based typing (SBT) of HLA genes. The datasets include a parent-child trio, a large number samples sequenced at

high depth with short read lengths and a moderate number of cancer samples sequenced at moderate depth with high read length.

The first set of data was generated from three cell lines (gm12878, gm12891 and gm12892) representing a mother-father-daughter trio as part of the ENCODE project [25]. Raw FASTQ files containing paired 75 nt reads were downloaded from the UCSC Genome Browser, Encode release 4 (August 2012).

The second set of data was generated as part of the HapMap project. Fifty RNA-seq samples were sequenced in a study performed by de Bakker et al., 2006 HLA alleles were determined using SBT by Montgomery et al., 2010 [28, 101]. Short reads were downloaded from the NCBI SRA with the study accession id ERP000101. These samples were sequenced using paired 37 nt reads and represent the type of data typically generated in early RNA-seq runs or for runs where turnaround time is essential.

Finally, sixteen colorectal cancer samples were HLA typed in a study performed by Warren et al., 2012 [164]. These samples were initially sequenced in an attempt to identify fusion genes in colorectal cancer, but serve as a useful dataset for HLA typing. The samples were sequenced as paired 100 nt reads at moderate depth. The data were downloaded from the NCBI SRA under the study accession id SRP10181.

4.3 Results

4.3.1 Simulations

Accuracy was largely affected by read length, with longer reads providing substantially better results. Read lengths of 2x50, 2x75, 2x100, and 2x200 nt were tested for their effect on accuracy (Figure 3a). Greater read lengths provided more accurate results, with 2x200 nt achieving very high accuracy (96.7%) at the peptide level. At 2x100 nt an average accuracy of 94% was achieved. At lower read lengths, accuracy declined at the peptide level but increased at the nucleotide level. This effect is artifactual as accuracy is not assessed at the nucleotide level if the true allele is only typed to the peptide level. Allele group level suffered with 2x37nt and 2x50nt reads, achieving 94.5% and 97.6% accuracy, respectively. However, accuracy was above 98.7% for 2x75nt, 2x100nt and 2x200nt read lengths.

Simulations showed a minimal effect of increasing sequencing depth on accuracy. Total coverage amounts of 1x, 10x, 25x, 50x, 100x and 500x for each chosen allele were tested (Figure 3b). Although high resolution performance suffered at very low coverage (64% at 1x), accuracy jumped to 92% with just 10x coverage. Increasing coverage above 25x had minimal effect on high resolution accuracy. Peptide level accuracy was above 94.9% for all coverage levels above 25x. Allele group level accuracy was above 98.9% at coverage levels greater than 10x.

Substitution rates of 0.009%, 0.5%, 1.1%, 2.2%, 4.4%, and 8.8% were assayed in order to test the effect of sequencing error on the accuracy of the method. Substitution rate had minor effects on the accuracy of the method. With substitution rates below 2.2%, peptide level accuracy was above 93%. At 4.4%, accuracy declined to 90%. With a large substitution rate of 8.8%, peptide level accuracy was still respectable at 84%. Allele group level accuracy was above 98.9% for all substitution rates tested.

Simulations using all genes in the IMGT database rather than just HLA-A, HLA-B, HLA-C and HLA-DRB1 were conducted to see the accuracy of the method on each individual gene. Low resolution accuracy was above 98% for most genes except for a few selected genes (DPA1, DPB1, DRB6, MicA, MicB, Tap1, Tap2, V). Peptide level accuracy was above 91% for HLA-A, HLA-B, HLA-C, but fell below 90% for remaining genes.

Disallowing mismatches during the alignment step resulted in higher accuracy when substitution rates were sufficiently low. At 0.009% error, peptide level accuracy was 94.3% when mismatches were allowed, but increased to 95.6% when only exact matches were utilized. However, at higher substitution rates, performance declined achieving an accuracy of 55% at 1.1% error rate. At 0.009% substitution rate, accuracy was improved by 1-2% over all conditions tested when no mismatches were allowed during alignment. With a low substitution rate and sufficiently high coverage, it may be beneficial to restrict the number of mismatches during alignment. Boegel et al., 2012 report a similar effect and recommends allowing only a single mismatch during alignment [13].

An example demonstrating the effect of read weighting and tree pruning can be visualized in Figure 4.4. Here, one hundred 2x100 nt reads were sampled for the

HLA alleles HLA- A*02: 90 and HLA-A*26: 36. The difficulty of predicting HLA allele groups without read weighting or tree pruning can be seen in Figure 4a, which charts the maximum number of reads that map to a child member of an allele group. Although HLA- A*26 is ranked first in the number of reads supporting that allele, three incorrect allele groups rank above the true allele group of HLA-A*02.

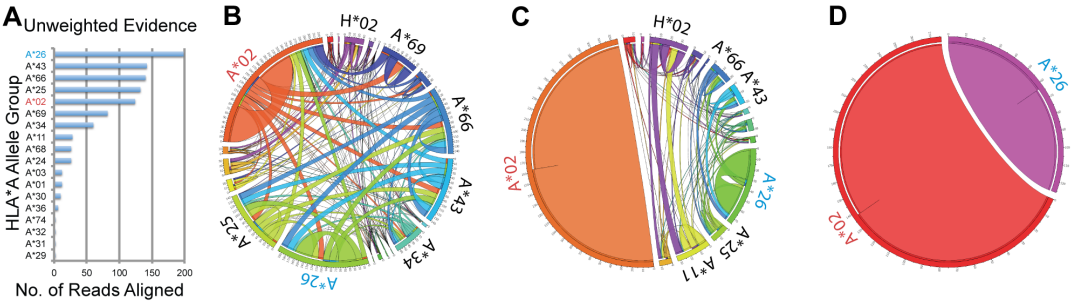

Figure 4.4: Effect of read weighting and tree pruning on predicting allele group. In this simulated example, HLA-A*02: 90 (labeled in red text) and HLA-A*26: 30 (labeled in blue text) are the true alleles. (a) The maximum number of reads mapping to any descendant of allele groups are shown. These results represent a naive attempt at predicting alleles from RNA-seq data where reads are unweighted. (b) Evidence for each allele group after building weighted read trees. Allele groups are labeled on the outermost circle. Arcs connecting allele groups have widths proportional to the amount of evidence that shared between connected allele groups. Here it is already evident that A*02 and A*26 have the most evidence, but other alleles have substantial evidence. (c) The effect of pruning read trees after selecting the primary allele (A*02) clearly distinguishes the secondary allele group (A*26) from other allele groups. (d) A final pruning step removes all ambiguous edges and assigns all evidence to the true allele groups.

Hierarchically weighting and pruning the simulated reads allows for the reduction of noise and for accurate prediction of allele groups in this example. The hierarchical weighting procedure provides multiple benefits. First it distributes all the evidence a single read could provide to all possible alignments, rather than giving equal weight to all alignments. Secondly, it allows for the visualization of the amount of shared evidence between all allele groups (Figure 4.4.B). After weighting has been applied, HLA-A*02 and HLA-A*26 contain the majority of evidence, however much of the evidence for the incorrect allele groups remain. The intermediate pruning step, where ambiguous evidence is assigned to the primary allele group, significantly reduces noise. Figure 4.4.C

shows that post pruning, the secondary allele group with the most evidence is clearly A*26. The final pruning step removes all ambiguous evidence (Figure 4.4.D).

4.3.2 Colorectal Cancer Data

In a study by Warren et al., 2013 on HLA typing using short read assembly, he presented a dataset of sixteen RNA-seq samples for which Class I molecules were typed by PCR [164]. Of the 96 total possible alleles, 87 were typed to the peptide level. Reads for this dataset were 2x100nt in length. On average, 3,500 reads aligned to the IMGT database. HLAforest predicted 85 (97.7%) of low resolution alleles correctly and 74 (85%) of high resolution peptides correctly.

Table 4.1: Accuracy of typing results from 50 HapMap samples with 2x37bp reads allowing or not allowing mismatched alignments to references.

	Mismatches	A	B	C	DQA	DQB	DRB	Total
Allele Group	Yes	.950	.930	.950	.980	.980	.976	.960
Peptide	Yes	.940	.838	.680	.804	.804	.872	.821
Allele Group	No	.950	.940	.950	1	.990	.963	.965
Peptide	No	.940	.848	.730	.804	.825	.872	.835

4.3.3 HapMap data

Fifty HapMap samples that have been both HLA typed with Sanger Sequencing and for which there are RNA-seq data available were analyzed using HLAforest. RNAseq data was 2x37nt in length. HLAforest was able to predict 96% of allele-groups correctly. 82% of peptide-level alleles were also predicted correctly. When alignments were restricted to disallow mismatches, HLAforest was able to predict 96.5% and 83.5% of peptide level alleles correctly (Table 4.1). On average, 127,000 reads aligned to the IMGT database with mismatches allowed during alignment and 91,000 reads aligned when only exact matches were reported.

4.3.4 Family Trio

Three cell lines, gm12878, gm12891, and gm12892, representing a daughter-father-mother trio, respectively, were used to test this method. It is expected that the daughter (gm12878) would carry a set of alleles from each parent. 2x75bp reads were used to predict alleles for all genes present in the IMGT database.

Table 4.2: Predicted alleles of major HLA genes on the daughter-father-mother trio of cell lines using exact alignments.

Gene	Father (gm12891)		Mother (gm12892)		Daughter (gm12878)	
	Primary	Secondary	Primary	Secondary	Paternal	Maternal
A	01:01:01**	24:02:01**	11:01:18**[3]	02:01:01**	01:01:01**	11:01:01**[3]
B	08:01:01**	07:02:01**	15:01:01**	15:01:20[1]	08:01:01**	56:01:01**[1]
C	07:02:01**	07:02:01*[2]	01:02:01**	04:01:01**	07:01:01**[2]	01:02:01**
DPA1	01:03:01	01:03:01	02:01:01	01:03:01	01:03:01	02:01:01
DPB1	04:01:01	03:01:01	14:01	06:01	04:01:01	14:01
DQA1	05:01:01	01:02:01	01:01:01	01:01:01	05:01:01	01:01:01
DQB1	02:01:01	06:02:01	05:01:01	05:01:01	02:01:01	05:01:01
DRA	01:02:03	01:02:02	01:01:01	01:01:01	01:02:02	01:01:01
DRB1	03:01:01	15:01:01	01:01:01	01:01:01	03:01:01	01:01:01

* Consistent at allele group level with types from Erlich et al., 2011 [37]

** Consistent at peptide level with types from Erlich et al, 2011.

[1] Inconsistent between parent and daughter at allele group level

[2] Inconsistent between parent and daughter at peptide level

[3] Inconsistent between parent and daughter at nucleotide level

When compared to types determined by targeted resequencing and Sanger sequencing as performed by Erlich et al., 2011 [37], all HLA class I alleles were recapitulated for gm12878 at the peptide level (Table 4.2). Most class I alleles were recapitulated for gm12891 and gm12892, except for those that were found to be discordant when compared to gm12878. In all, sixteen out of eighteen (89%) of class I molecules were called consistently with previous studies. Accuracy of other genes, assessed by looking for consistent predictions between the daughter (gm12878) and her parents, had accuracy

similar to major class I genes.

Accuracy of typing was assessed over the fifteen genes with sufficient coverage (greater than 1% of all mapped reads supporting the gene). Of the thirty alleles in these genes, twenty-six were predicted consistently at the peptide level, two were predicted consistently at the allele group level and two were completely miscalled.

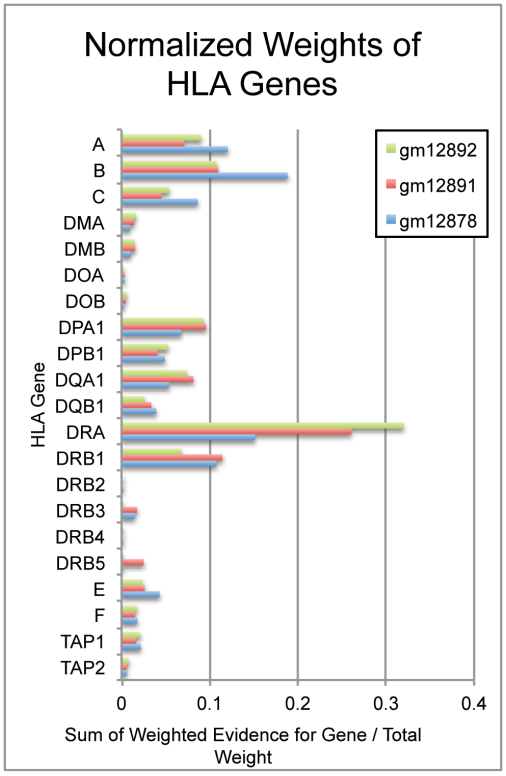

Figure 4.5: Final pruned weights supporting each gene in the IMGT database shows expression over major class I molecules (A, B, C) as well as over most major class I molecules (DMA, DMB, DPA1, DPB1, DQB1, DRA, DRB1). Some expression is seen in minor class I alleles (E, F) and non classical molecules (TAP1 and TAP2).

In addition to predicting alleles, the method allows for the estimation of expression of all genes. Figure 4.5 shows the number of pruned reads aligning to each HLA gene. We see that classical class I molecules (HLA-A, HLA-B, HLA-C) have moderate expression in all samples. Likewise there is moderate to high expression of some major class II genes (HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA and HLA-DRB1). There is lower expression of the minor class I molecules along

with some class II molecules (HLA-E, HLA-F, HLA-DMA, HLA-DMB, HLA-DOA and HLA-DOB).

4.4 Discussion

The method described here performed well in simulations with read length and substitution rates mirroring those of available sequencing technologies, namely Illumina sequencers. HLAforest has the advantage of scaling well with longer reads and it fully utilizes the phasing information present in paired-end reads. The method is generalizable to any set of genes that can be arranged hierarchically. It also has the major benefit of selecting alleles individually, thus reducing the complexity and combinatorial difficulty of selecting two alleles simultaneously. Some problems remain, including the inability to call novel alleles and report ambiguous alleles, but these can be addressed in the future.

Simulations show that this method can achieve a high resolution accuracy of 93% (2x100bp reads and 0.5% substitution rate) over the genes that represent the majority of diversity in the IMGT database (HLA-A, HLA-B, HLA-C, HLA- DRB1). Evaluation of this method on the daughter-father- mother cell line trio shows that 26/30 (87%) the daughter's alleles were predicted consistently at the peptide level. After comparison to results of Sanger sequencing and targeted resequencing with Roche 454, it was determined that errors in typing occurred within gm12891 and gm12892 (Table 4.2). Errors in typing may be related to the 2x75bp read lengths available for these cell lines. In order to make accurate predictions, HLAforest relies on the phasing information within individual reads, which are dependent on read length.

The majority of information extracted from short reads comes from the ability to phase discriminatory SNPs across the most diverse coding regions of the HLA genes. For class I genes, the majority of diversity are present in exons 2-4 [119]. Here, short reads of sufficient length and quality are able to phase the discriminatory SNPs that define each allele. Indeed, increasing the read length to 2x200 nt in simulations substantially improves the accuracy of the method with accuracy greater than 96.7% at the peptide level. This finding is significant as high throughput sequencing technologies such as Illumina and Ion Torrent have announced plans to release 2x200 nt sequencing

kits. The HLAforest method is best applied to RNA-seq data as reads are more likely to phase discriminatory SNPs across exons in fully spliced transcripts. If read lengths or insert size exceed the length of introns in these genes, this method can be extended to the whole genome or targeted sequencing datasets without loss of accuracy.

The robustness of this method is apparent in simulations testing effect of substitution rate on accuracy. Modern Illumina machines report an overall substitution rate <0.5% [125]. In simulations with comparable error rates, HLAforest is able to predict 93% of alleles at the peptide level. Even with substitution rates as high as 8.8%, HLAforest can predict 99% of low resolution alleles correctly.

This method has major benefits over earlier methods. First it can be generalized to any set of genes that are classified hierarchically. Secondly, as opposed to competing methods, there are no combinatorial issues with selecting two alleles to score simultaneously. In this method, alleles are chosen procedurally and this reduces the computation time necessary for scoring many hypothetical pairs of alleles.

Recently, HLAMiner has been published and shares many of the same benefits as HLAforest. The major distinction between the methods is HLAMiner's reliance on the de novo assembler, TASR [166]. HLAforest exploits highly efficient short read alignment algorithms, which have been the subject of major development in the bioinformatics field. This alignment step is easily parallelizable, as opposed to de novo assembly methods that require shared memory. Additionally, HLAforest uses all the phasing information within paired-end reads rather than attempting de novo assembly with shorter k-mers. Differences in the methodologies make direct comparison of the methods difficult. However, our simulations with 2x100 nt reads and 0.55% substitution rate show an average major class I accuracy of 92.7% at the peptide level as opposed to HLAMiner's sensitivity and specificity of 84.7% and 89.65% with the same parameters, respectively. When error rates were increased to 2%, average major class I accuracy with HLAforest dropped slightly to 92.7% whereas the sensitivity and specificity of HLAminer's simulations were 54.9% and 87.5% respectively. The data suggests that HLAforest's predictions are more accurate than HLAMiner even with larger error rates.

When compared to HLAMiner on sixteen colorectal cancer samples, HLAforest was able to predict 97.7% of low resolution alleles and 85% of high resolution alleles

for Class I molecules. HLAMiner reported 95.6% sensitivity and 99% specificity at low resolution, and 90.7% sensitivity and 93.5% specific at high resolution. It is worth noting that the sensitivity and specificity measures reported by Warren, et al. are not standard. Whereas HLAforest generated 96 predictions for the 96 possible peptide-level Class I alleles, HLAMiner generated 235 predictions.

HLAforest's predictions for colorectal cancer samples fell below the levels predicted by simulations with 2x100nt read lengths. This is perhaps due to the low number of reads aligning to the IMGT database, especially when compared to the HapMap samples (3,500 for colorectal samples vs 127,000 for HapMap samples). Insert size may also play a role in prediction accuracy.

HLAforest performed well on fifty HapMap samples using 2x37nt reads. It was able to predict 96% low resolution types accurately. Performance of HLAforest was comparable to seq2HLA, which predicted low resolution types with 100% sensitivity and 93% specificity. However, HLAforest was also able to predict 82% of high resolution types correctly. Remarkably, HLAforest performed better on these samples than it did in simulation with similar readlengths. This is perhaps due to the increased coverage in the HapMap samples as well as the reduced representation of alleles present in these samples. These cross study comparisons imply better performance with HLAforest; however, such comparisons should be interpreted cautiously until systematic benchmarking can be performed.

Although HLAforest presents many strengths, there are some shortcomings. First, typing class II MHC molecules may be impossible based on the cell type of the input RNA. Because only specialized antigen presenting cells (such as B cells, dendritic cells, etc) express class II MHC molecules, some of these cells must be present in the sample in order to generate the corresponding reads. Second, the method is restricted to alleles that already exist in the IMGT database. HLAforest chooses the closest matching allele, but does not reconstruct the actual allele. It is feasible to look for novel alleles by generating a consensus sequence from the reads supporting each predicted allele and then checking for novel SNPs. Third, HLAforest reports only two alleles and does not produce confidence scores at this time. Finally, this method does not yet incorporate population-based frequency data that has been shown to improve the accuracy of all

typing methods in well-studied populations [120].

Chapter 5

UCSC Immunobrowser

5.1 Background

The previous chapter discussed the MHC/HLA molecules that are important for displaying the peptides found intracellularly or extracellularly. HLA alleles encode the range of peptides that can be displayed. The genetic diversity of HLA alleles in the population is generated by balancing selection, the surface structure recognized by TCRs also include a processed peptide. The diverse surface structure of the MHC-peptide complex is therefore a result of a large number of HLA alleles in addition to the set of proteins found within the human body. Although the surface structures of MHC-peptide complexes are highly diverse, the peptide display functionality allows structural diversity to be encoded directly within the genome.

It is the T-cell's responsibility to analyze displayed peptides and to elicit an immune response when abnormal peptides are detected. T-cells are able to detect a wide array of possible MHC+Peptide complexes through the characteristic T-cell receptor (TCR) [109]. TCRs bind to peptides that have been processed and presented by MHC class I and II molecules. The wide array of displayed peptides are reflected by the number of unique TCRs that circulate freely through the lymphatic and circulatory systems [49].

Rather than encoding the detecting receptor of each peptide directly in the genome, T-cells utilize a system for generating a diverse array of binding pockets. VDJ Recombination is the molecular process that gives TCRs their diversity. This system

allows for the generation of T-Cells with 10^{15}-10^{20} different receptor sequences. In the process of recombination, each T-cell precursor undergoes a somatic rearrangement of multiple gene segments to create a novel DNA sequence. The rearranged DNA sequence defines the binding specificity of its receptor. Recombination contains multiple steps including combinatorial junction joining, junctional flexibility, palindromic addition (P-addition), and random nucleotide addition (N-addition).

Combinatorial junction joining is the process in which V, D and J segments are joined together randomly. Each V and J segment contain conserved sequences for the formation of the immunoglobulin fold, but differ significantly at the sequence level. Combinatorial joining of individual V, D and J segments account for 2.25 x 10^6 and 5.29 x 10^6 possible combinations in BCRs and TCRs. TCR-β and IgH chains recombine V, D and J units while TCR-alpha, Igλ and Igκ recombine only the V and J segments.

During the joining of V, D, or J segments, genetic sequence may be altered by deletion or by addition of palindromic sequence at the junctions of a segment. These additions are a result of the joining process that creates a hairpin at the end of each segment and creates a nick randomly. Following the creation of a nick, random nucleotides are added by terminal deoxyribonucleic acid transferase. These deletions and additions account for much of the diversity found in recombined TCRs.

T-cells express only a single receptor despite having two parental loci for each gene. Studies have shown that some T-cells have rearrangements in both parental alleles. Despite this, only a single parental allele for each chain is expressed. The method which silences a parental allele is known as allelic exclusion. If a productive VDJ rearrangement is made, a functional chain is translated and molecular mechanisms stop any further rearrangement. If a VDJ rearrangement is nonfunctional, the second parental allele is rearranged. If both rearrangements are nonproductive, the cell dies through apoptosis [97, 142].

Maturation of B-cells and T-cells share many similarities. Both B-cells and T-cells are created from hematopoietic stem cells in the bone marrow. B-cells remain in the bone marrow throughout maturation, whereas progenitor T-cells will migrate to the thymus to mature [87]. During maturation, sequential somatic recombination of receptor locii produce mature B-cells and T-cells. Self-reactive B-cells and T-cells are

deleted by apoptosis in a process known as negative selection[113]. B-cells that are self-reactive may undergo light chain rescue, in which alternative recombination of the light chain can reduce self-binding affinity [111]. T-cells also undergo positive selection in which T-cells that are unable to recognize self MHC are eliminated [144].

B-cells and T-cells have less in common during activation than they do during maturation. T-cell activation requires the binding of a TCR and its coreceptor CD4 or CD8 to a peptide-MHC complex. A costimulatory signal between B7 on the antigen presenting cell and CD28 on the T-cell are also required for activation of the T-cell. Stimulation of the CD28 receptor on the T-cell by CTLA-4, a close relative to B7, prevents activation and suspends the T-cell in a resting state known as anergy. Activated T-cells generate effector and memory T-cells. Memory T-cells are more easily activated than naive T-cells allowing for a faster response to second exposure of an antigen [76].

B-cells migrate from the bone marrow into the periphery where they are generally activated by helper T-cells. With lipopolysaccharides and other repetitive epitopes In some cases, B-cells can activate in the absence of T-cells. Activation of B-cells result in the production of both plasma B-cells, that produce antibodies in large quantities, and memory B-cells. A week after activation, germinal centers form in lymph nodes, which serve as a site of somatic hypermutation, antibody class switching, and clonal expansion of B-cells [76].

There is a discordance between the number of possible VDJ rearrangement and the number typically observed in sequencing experiments. This can be partially explained by the degeneracy of the genetic code multiple DNA sequences can code for the same amino acid sequence. This is most evident in infant samples, where due to a lack of TdT expression, the number of possible rearrangements are much lower than in adults [30]. There are many ways to generate the same amino acid sequence from varying junctional joinings. It has been theorized that these early rearrangements target common antigens allowing the young to build up a strong and capable immune defense against common environmental pathogens.

This observation has been seen in both the BCR and TCR loci. In a study of zebrafish, Jiang et al., 2011 found that Ig VDJ junctional usage between young zebrafish were highly correlated. A different correlation were observed in older fish.

The primary recombination rates in adult zebrafish were highly correlated, although clonal expansions lower the correlation at the total repertoire level [66]. Rudd et al., 2011 observed a similar phenomenon in young and old mice responding to a common antigen [134]. Another important observation in Rudd's study was that there were multiple different TCR rearrangements that targeted the same antigen.

It is important to consider HLA alleles in the shaping of the immune repertoire. A number of studies have found a relationship between VDJ usage of an immune response to a specific antigen and the allele of individuals [100]. The difference in structure of MHC class I and class II molecules also effect the VDJ usage in T-cells, where specific gene segments are preferred for specific MHC alleles [155]. In one example, adults positive for HLA-A*02 preferentially use TCR-α-10.2 and TCR-β-17 V gene segments during an immune response to the influenza virus [145]. Some of these public immune responses even share the same CDR3 amino acid sequence. Typically the shared amino acid sequence is largely derived from rearrangements with low numbers of insertions, further supporting the theory that early rearrangements provided a broad defense against common antigens. Recently, TCR sequencing experiments also provide some evidence for preferential VDJ usage among healthy individuals with the same allele that are not undergoing an immune response to an antigen [165].

TCR sequencing experiments have been useful beyond general study of the immune repertoire and the detection of public TCRs. Recent years have shown increasing study of the role of the immune system in cancer. The immediate applications of TCR sequencing experiments have obvious applications to T-cell driven Leukemia. Numerous studies have shown that TCR sequencing experiments are as effective at detecting lymphoid cancers such as ALL and CLL versus standard methods such as spectratyping or Sanger sequencing [95]. They are also effective for tracking the treatment efficacy of chemotherapies. Post treatment sequencing experiments are also useful for tracking of minimal residual disease (MRD) for the detection of cancer relapse.

Perhaps the greatest side effect of common standard of care cancer therapies such as chemotherapy and radiation is the destruction of all quickly replicating cells. Although these have the capability of destroying cancer, they also have the side effect of destroying the immune system, which relies on replication and clonal expansion for

effector functions of the adaptive immune system. In the case of lymphomas, immune depletion may be the intended effect where the immune system is then reconstituted by hematopoietic stem cell transplantation (HSCT). In either case, the reconstitution of the immune repertoire can be carefully monitored to ensure that healthy immune function is restored post therapy. As HSCT is extended for diseases outside of cancer, such as MS or HIV, proper immune reconstitution monitoring becomes another important use case for TCR sequencing.

5.1.1 Sequencing of TCR Repertoires

High throughput sequencing (HTS) or next generation sequencing (NGS) refers to a class of sequencing machines that produce a much greater throughput of sequence information when compared to traditional Sanger/capillary based sequencing [95]. Capillary based sequencing is typified by its long read length, but is only capable of sequencing a single amplicon at a time. HTS machines are capable of producing significantly greater amounts of sequencing information in a single run, but typically produce shorter read lengths than those afforded by Sanger sequencing. Of the many HTS technologies available, platforms by Illumina and Roche 454 are preferred for immunosequencing assays although Ion torrent sequencers have also seen increasing use [129]. Both technologies have their strengths and drawbacks and the selection of a sequencing platform depends on the main goal of the experiment.

The 454 titanium generates long read lengths (read lengths up to 1,000 bp have been reported) and produces around 106 reads per run. The long read lengths allow researchers to sequence an entire variable domain of BCR/TCR, which range from 350-400 bp in total length [179]. This is especially important for BCRs because of the somatic hypermutation that can generate SNPs across the entire domain. However, the relatively low number of reads limits the sampling depth of the repertoire and rare clonotypes are typically missed with this sequencer.

The HiSeq is Illumina's latest sequencing machine and is known for its generous throughput. A single run on a HiSeq is capable of producing upwards of $3x10^9$ sequencing reads. However, the read length is much shorter than that offered by the 454. The maximum read length of a HiSeq is 100 bp read from either end of a DNA

fragment. Known as paired-end sequencing, this method generates 100bp reads from both the 5' and 3' end of a fragment [179]. The shorter read length prevents researchers from sequencing an entire variable domain. Thus, researchers tend to concentrate on the CDR3 region of BCR/TCRs. Because TCRs do not undergo somatic hypermutation, the DNA sequence of the entire variable domain of a specific clonotype is less informative than for BCRs.

The choice of sequencing platform greatly dictates the sample preparation methods used to prepare the sequencing libraries. Because the 454 offers longer read lengths, sequencing primers can be designed in more conserved regions of the BCR/TCR. When generating a library for Illlumina, generally multiplex PCR techniques with primers designed for each V and J segment are necessary. As the read length of high throughput sequencing machines increase, it is likely that all immunosequencing experiments will sequence the entire variable domain in order to reduce primer bias and call alleles in each sample.

A number of approaches for amplifying CDR regions have been applied in human [16, 41, 78, 131, 132, 161, 165], other organisms [66, 128, 167] and phage-display libraries [127]. Once DNA or RNA is isolated from a sample, the VDJ region is amplified by a variety of techniques. The simplest approach is PCR where primers have been designed for conserved regions of an RNA transcript. In this scheme, a single reverse primer designed for the constant domain of each TCR/BCR can be utilized with a small number of forward primers.

DNA based approaches must utilize primers designed specifically for each V and J segment because the constant region is separated by a large intronic region. Because the recombination joins V segments to J segments randomly, primers designed specifically to each V and J are utilized. Multiplex PCR (mPCR) differs only from traditional PCR in that more than one pair of primers is used. Methods based on mPCR tend to have problems with quantification because small differences in template GC content and primer binding efficiency can result in a largely skewed final dataset [122, 70]. In order to combat primer bias in multiplex PCR schemes, Klaranbeek chose to first use bi-directional linear amplification followed by universal PCR (uPCR) primers [78]. In bi-directional linear amplification, forward and reverse primers are designed in the same

manner as PCR. The primers in this study also contained universal primer sequences at the 5' end which are utilized during the universal PCR step. During amplification, only a single directional primer is used preventing the exponential amplification that would have resulted in large biases. After the forward amplification is finished, the remaining primers are removed and a reverse directional amplification is done. After the linear amplification, traditional PCR with universal primers is done to increase the total amount of DNA material available for sequencing. This method reduces the effect of primer bias by amplifying exponentially only after the initial template is enriched by linear amplification.

Amplicon Rescue Multiplexed PCR (armPCR) is a proprietary technique that is conceptually similar to linear PCR followed by uPCR [53]. In the armPCR scheme, a low concentration of gene specific primers are used for a small number of PCR cycles. The left over primers are degraded with an exonuclease followed by a large scale amplification using universal primers. Because the primer bias is only in effect for a small number of cycles, it is believed that armPCR gives a more realistic viewpoint of the depth of clonal expansion within a sample.

5.1.2 Junctional Analysis Software

Immunosequencing data analysis is directed by the goal of the experiment. Thus the analysis techniques differ greatly from study to study. However, every experiment begins with junctional analysis software, which converts short sequencing reads into immunologically meaningful data. The types of results produced by these programs ultimately depends on the package. These tools, however, generally V and J gene segments. Some are able to call D segments, characterize insertions and deletions in junctional regions, translate the nucleotide sequence into a protein sequence, and determine whether a sequence produces a frame-conserving amino acid sequence.

Most publicly available software for junctional analysis was developed before the advent of immunosequencing. There are two types of tools: blast-based tools that utilize fast alignment algorithms to identify gene segments and HMM-based tools that are capable of producing statistically rigorous probabilities of each gene-segment call with higher accuracy.

Blast-based tools such as IMGT/V-QUEST+JVTA, Ab-origin, NCBI IgBlast, and JOINSOLVER are functionally similar. DNA sequences are blasted against a database of known immunoglobulin sequences and the highest scoring results are reported [143, 45, 18, 162]. Typically, the DNA sequences are blasted against V-segments and J-segments separately. Once both V and J segments are called, the frame of the V-segment is identified and the nucleotide sequence is translated to an amino acid sequence. A D-segment may also be called if it is available, but typically D-segments are difficult to call due to their inherently short length and the deletions/insertions created during the recombination process.

HMM-based methods work in a manner similar to blast-based tools, but more sensitive HMM based alignments are used in place of blast-based tools. HMM-based junctional analysis tools are designed to align specifically to the VDJ recombination region and incorporate insertions and deletions into their topography. HMMs such as iHMMune-align and SoDA report the most probable alignment, by using the Viterbi algorithm [43, 159]. SoDA2, the successor to SoDA, gains the capability of reporting the probabilities of multiple possible alignments by implementing the backward algorithm [108]. A benchmark by Jackson et al., 2010 shows that HMMs have better results than blast-based tools [61].

These tools were all designed with low throughput experiments in mind. Tools such as iHMMune-align and SoDA2 sacrifice runtime for increased sensitivity and power. Some attempts have been made to transform these algorithms for high throughput data (IMGT-HIGH, V-QUEST). However, IMGT is not suitable for automation due to a web-only interface and a 150,000 sequence limit per submission and only IMGT is able to perform junctional analysis on TCR sequences, the remaining tools have concentrated on IG sequences.

As the technology matures, numerous groups have developed software to better handle data at the scale of HTS. Decombinator and miTCR are two software packages that have been developed to handle the size of datasets being generated by modern sequencers. Decombinator uses a finite state machine in order to generate gene segment predictions in linear time [150]. Although the method is fast, the method fails to assign segments to 12% of reads with 1% sequencing error. miTCR identifies gene segments

using a seeded-extension algorithm based on the short motifs found in V and J gene segments [14]. miTCR performs well due by performing junctional analysis on high quality reads. Low quality reads are then mapped to high quality reads in linear time. The methodology is capable of eliminating >95% of sequencing errors and executes quickly enough to be applicable to real world datasets.

However, the majority of these tools are dedicated to the processing of raw sequencing data, calling VDJ gene segments in recombined receptors and mitigating sequencing error. Even software packages that offer visualization tools, such as miTCR, are not well suited for cross-sample comparison. While commercial services such as adaptiveTCR and iRepertoire offer web-based analysis and visualization tools, these are not open to the public and data generated using open protocols are not compatible with their systems. Thus, the need for an open academic resource for the comparison of T-cell receptor data is apparent.

5.1.3 Repertoire Analysis Methods

The types of repertoire-level analysis that are performed is application specific. Since its initial description in 2008, the assay has been utilized for characterizing general features of the TCR, tracking of cancer relapse and identification of public TCRs.

Because TCR sequencing offered an unprecedented sampling depth and sensitivity compared to previous methods, early studies often sought to character general characteristics of the TCR repertoire. Features such as gene-segment usage, CDR3 length, clonal expansion and upper bound estimates of the number of TCR sequences were assessed by numerous groups.

Gene-segment usage analysis can be explored to test for biased VDJ gene-segment usage or as a proxy for clonal expansion. The underpinning methods used for analysing gene-segment usage are simple. The fraction of reads or clonotypes which utilize a particular V or J gene segment ares counted and are plotted as a histogram where a categorical range is defined by the set of V or J genes. Finer granularity can be achieved by generating fractions for V-J pairs, or clonotypes that utilize a specific V and J. Comparison of V or J gene segment usage across multiple samples can be accomplished simply by plotting samples as lines or color-encoding samples in barcharts.

Visualizing V-J pairs is often performed with standard InfoVis techniques for two-dimensional histograms. Heatmaps, 3d histograms, scatterplots or chord diagrams are usually used to display V-J usage frequency. Both paired and unpaired gene-segment usage visualizations allow researchers to quickly identify overrepresentation of gene-segment usage.

By themselves, these plots are useful for identifying over-representation of a gene segment usage. Although these plots are generalizable to any TCR sequencing experiment, a variety of conclusions can be drawn depending on input material and data processing. For example, by looking at the non-coding subset of unique TCR sequences originating from DNA it was possible to determine that VDJ gene segment selection during recombination is indeed biased [110]. Alternatively, by looking at the overall frequency of coding sequences from DNA, it is possible to predict the total number of T-cells within the body sharing the same TCR sequence [165]. Or by sequencing both the RNA and DNA from a blood sample, it's possible to identify V-J gene segment subsets that are undergoing clonal expansion, as it is known that during activation and clonal expansion expression of the TCR is increased [76]. However, comparison of paired gene-segments usage across multiple TCR samples is difficult because each sample requires its own plot.

CDR3 length is another feature of TCR sequences that can be assayed and visualized at a repertoire level. CDR3 length can be determined at both the nucleotide and amino acid level. At the amino acid level, CDR3 length is distributed as a Gaussian. Due to selection for coding TCR sequences during development, nucleotide CDR3 length of multiples of 3 are also distributed as a Gaussian, whereas the remaining lengths are considered non-coding and underrepresented [156]. Nucleotide histograms of CDR3 length frequency recreate spectratypes, where the read out is fluorescent intensity measured during capillary electrophoresis. For RNA-sequencing experiments, the frequency of non-coding lengths should be near zero. As with V-J usage diagrams, clonal expansions are easily identified in these plots. When assayed in samples originating for DNA and by looking at only non-coding sequences, parameters for the distribution of insertions and deletions generated by the VDJ recombination process can be measured.

Ultimately, many studies focus on clonal expansion. Large clones can be read-

ily identified by sorting a list of all TCRs by their frequency. Large clones also present themselves in gene-segment usage plots as V-J pairs with large representation or in spectratypes as abnormally large peaks. Some studies utilize figures that plot the cumulative frequency of a sorted list of clones to show the extent of clonal domination within a sample. Likewise, statistical summaries such as Shannon's entropy also serve as a single number that represents the extent of clonal expansion where smaller numbers represent more clonal dominance within a dataset.

Identification of largely dominant clones is important for the detection of cancers such as chronic lymphoblastic leukemia or acute lymphoblastic leukemia. In cases of CLL or ALL, a single TCR sequence will dominate the majority of the represented repertoire (cite harlan and other groups that have measured this). Such a large effect is readily visible in all described plots as well as entropy measures. Once a cancer-associated TCR sequence is identified, subsequent sequencing experiments during therapy and remission can track the effectiveness of treatment and detect cancer recurrence.

Clone tracking and overlap analysis is a common for the tracking of cancer-associated sequences, TCR sequences across T-cell compartments, as well as for identifying TCRs shared across multiple samples and potential public TCR sequences. TCR sequences shared across multiple samples or individuals potentially target the same epitope. Identifying shared sequences from flat files can be computationally inefficient due to the need to process large files and store sequences in memory. Previously sequenced samples may also provide insight into function of TCR sequences, however, current tools do not enable the identification of sequences shared over many samples– only samples within the current study.

Once a TCR sequence of interest is identified, IMGT offers a blast-based service for identifying similar sequences annotated in Genbank. However, many studies do not submit their sequences to Genbank, instead opting to include the sequences in their main text or in their supplement.

5.1.4 Issues with Current Methods

Although current methods are powerful, there are areas that could be drastically improved. Ease of use, improved visualizations and simpler access to a large

repository of TCR sequences are not emphasized. Without the standardization of tools, analyzing TCR sequencing experiments requires a background in computer science and the ability to program.

5.1.5 Considerations for Designing the Immunobrowser

A public resource for the analysis and storage of TCR sequencing experiments should provide powerful analytic tools, serve as a repository for sequencing data, be easily accessible, expandable and shareable.

The tools should be usable by those without substantial programming experience. The browser should not only recreate typical analysis performed on TCR sequencing experiments, but also overcome existing issues with current analytic techniques and visualizations. In order to be useful as an analytic tool, it must also be powerful enough analyze subsets of the data that are of interest and to produce new results in real time. Tasks such as tracking shared clones must be fast and the identification of all samples which share the same clone must be done quickly. Analytical tasks must be easily altered and the results persistent. The tools should not only be publicly accessible, but provide persistent results that are easy to share.

Finally, because technology and analytical methods continue to improve, the design of the Immunobrowser should be such that it is easy to expand and maintain. Analysis methods and visualizations should be modular, reusable and should not affect the underlying data.

5.2 Implementation

The UCSC Immunobrowser was built following a web-based Model-View-Controller (MVC) software design pattern. MVC architecture is one of the oldest design patters for user interface (UI) development and applications designed using MVC principles are prevalent in modern web applications. The major benefit of the MVC design pattern is the separation between the underlying data (Model), "business logic" (Controller) and the user interface (View). The UCSC Immunobrowser utilizes the python package, Django [31], which allows for simplified implementation of web sites.

All data storage and manipulation are sequestered within the Model layer. A built in object-relational mapping (ORM) allows for pythonic objects to be stored directly within a relational database (RDB), in this case a MySQL database is used. In addition to data storage and retrieval, all methods for data manipulation prior to presentation are encapsulated within this layer.

The controller layer is responsible for identifying the types of data and manipulations that should be prepared for presentation to the user. In the context of the Immunobrowser, the views receive a command from the user, decides what data to return and chooses the appropriate view to render the data.

The view is the user interface of the Immunobrowser. All the visualizations, tables and statistics generated by the immunobrowser are described in the views. Although django is capable of generating static visualizations when paired with other python libraries such as matplotlib, interactive figures require the use of javascript. I utilize d3.js to generate interactive figures [15]. In addition, jQuery and bootstrap.js are both utilized to build user-interface modules throughout the browser.

5.2.1 Immunobrowser Model Layer

The Model Layer describes, stores and manipulates all the data utilized by the Immunobrowser. The objects stored in the browser are generally divided into two classes. The first class stores experimentally derived data such as patient information, blood sample from a patient, clonotypes, genetic recombinations and amino acid sequences. The second class stores downstream analysis data such as the set of filtering parameters used to generate visualizations and comparisons between multiple patients or samples.

Five models are associated with data generated directly from a TCR sequencing experiment (Figure 5.1). A patient model stores individual-specific information that does not change between blood draws including: name, gender, disease and birthday. A sample model reflects individual TCR sequencing experiment and stores associated information such as cell type, blood draw date and a patient. A single patient may be associated with many samples. Each sample object is associated with one or more clonotypes. Biologically, a clonotype refers to a clonal population of T-cells sharing the

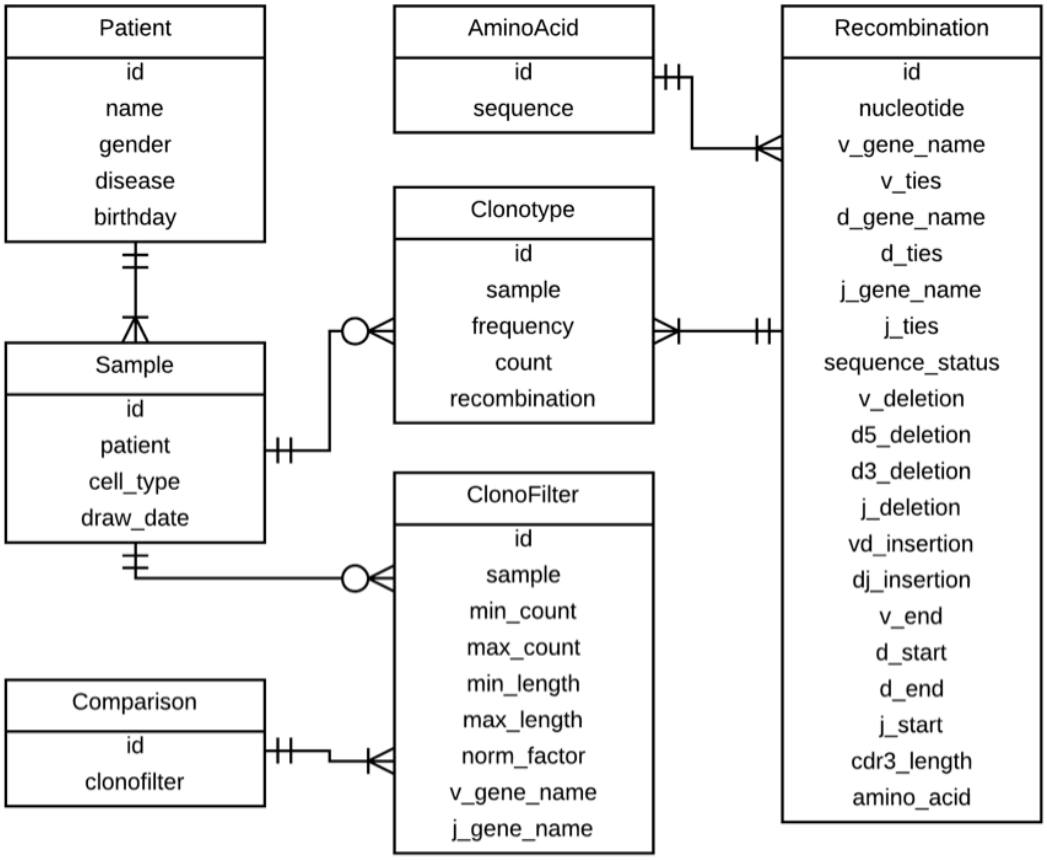

Figure 5.1: An entity relationship diagram of key Django models used to build the Immunobrowser. The Patient, Sample, Clonotype, Recombination and AminoAcid models store data generated from TCR sequencing experiments. The Clonofilter model and Comparison models store data associated with post-sequencing analysis.

same TCR. Here, the Clonotype model describes the abundance of T-cells expressing the same TCR nucleotide sequence in a sample. The Recombination model stores the genetic TCR sequence generated through VDJ recombination in vivo. A nucleotide sequence generated through targeted high throughput sequencing first undergoes junctional analysis using miTCR to identify the V, D and J gene segments used during recombination along with deleted and inserted bases. The annotated nucleotide sequence is then stored following the Recombination model. Finally, if a recombination is productive, it may be associated with an Amino Acid object that stores only a protein

sequence.

The ClonoFilter and Comparison models are used for comparative analysis of TCR sequencing experiments and are generated during repertoire analysis. ClonoFilters are associated with a single sample and store filtering parameters that allow researchers to visualize subsets of data. ClonoFilters are then used to retrieve a subset of clonotypes that adhere to filters present in a ClonoFilter. For example, a ClonoFilter with a minimum count of 10 would be used to query all clonotypes from a sample with at least 10 counts. ClonoFilters also have a normalization factor, which by default is set to the sum of observed counts. A comparison object is a set of clonofilters. ClonoFilter objects allow for many possible subsets of the data to be visualized on the fly, allow for persistent results with minimal data storage and simplifies sharing of results.

In addition to storing and retrieving data, the Model layer also contains methods for preprocessing data. For example, a method can return the total number of unique amino acid sequences found within a patient by performing an aggregate database query. More complex queries, such as returning the number of sequence reads associated with all V-J gene segment pairs are used to generate interactive visualizations presented by the Immunobrowser.

5.2.2 InfoVis Views for Exploratory Analysis of TCR Repertoires

Information visualization (InfoVis) is a subfield of data visualization that studies effective visual presentation of abstract data. The major focus of the field is to present large quantities of data in an intuitive and informative manner. Although the InfoVis field typically focuses on discovering and testing optimal methods for data presentation, the discovered techniques can be applied to the visualization of data generated from a array of fields including biology, economics, computer science, statistics and other fields where the structure of data does not have a rigid underlying visual structure. InfoVis techniques have also been utilized in TCR sequencing experiments to visually represent the usage of VDJ gene segments in the repertoire.

The UCSC Immunobrowser leverages common InfoVis techniques such as color coding, opacity and interactivity to enable visual analysis of TCR sequencing experiments. By incorporating these InfoVis elements, the shortcomings of traditional static

visualizations are avoided. In the following section, I review visualizations used in the Immunobrowser and compare their functionality to existing plots.

The majority of visualizations used in the Immunobrowser appear in the Comparison view. In the view, an arbitrary number of samples are compared through the use of a summary table, spectratype plot, scatterplot, functionality plot, domination plot and shared clones plot. InfoVis techniques are also utilized in other portions of the browser including the repertoire filter forms, TCR sequence view and the literature search view.

5.2.2.1 Summary Table

In some instances, a table can outperform a figure. This is the case with unrelated summary statistics generated from repertoire of a single sample. A summary table shows the number of unique amino acids, recombinations and sequencing reads observed for each sample. Shannon's entropy is also reported as a measure of clonal expansion in the repertoire. A low entropy value indicates the presence of highly dominant clones within a sample. It also contains a link to all the clonotypes within a sample that pass user-defined filters (Figure 5.2).

Summary

Sample	Reads	Recombinations	Amino Acids	Entropy	All Clonotypes
Wang 2009 None pan T rep1	15394	5603	5345	6.29	View all clonotypes
Wang 2009 None panT rep2	5958	3274	3106	7.03	View all clonotypes
Wang 2009 None panT rep3	66996	18377	17259	7.05	View all clonotypes

Figure 5.2: Summary table in the Compare view shows a numerical summary of filtered repertoires. Each row corresponds to a sample and shows the number of high throughput reads, genetic recombinations and unique amino acid sequences detected. In addition, the entropy of observed clonotype frequencies give a single numerical summary of clonal dominance in a sample. A hyperlink takes users to a detailed view of all clonotypes associated with the sample.

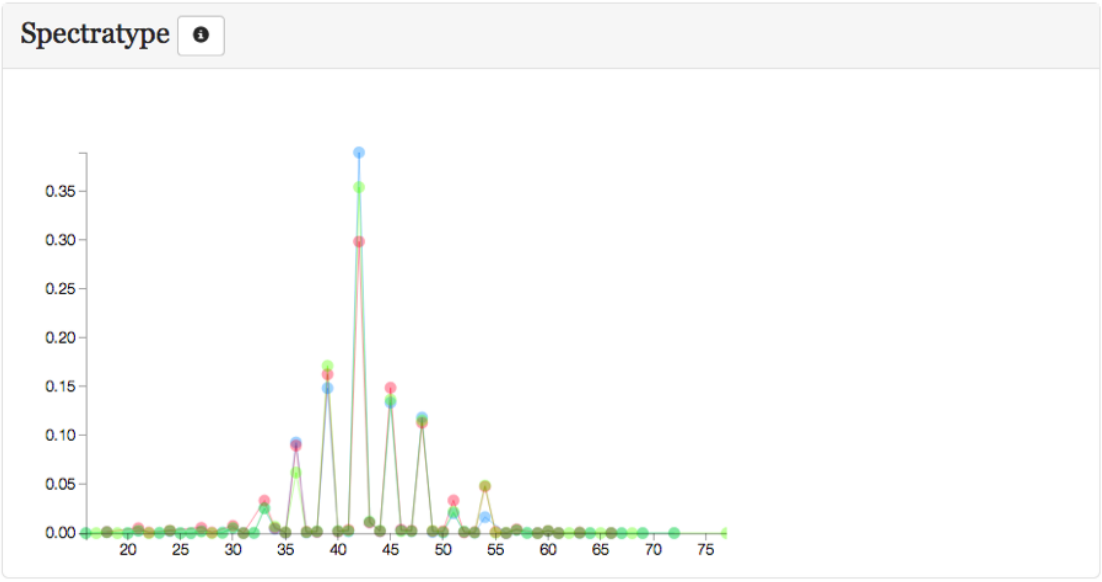

Figure 5.3: A spectratype shows the abundance of clonotypes (Y-axis) based on the nucleotide length (X-axis) of the recombined CDR3 sequence in the Compare view. In a sample derived from the DNA of a healthy non-challenged individual, it is typical to see peaks at lengths that are modal-3 (or every third length). Such a pattern is occurs due to selection of protein-coding receptor sequences during T-cell development. It is also typical to see peaks that are normally distributed in healthy individuals. Non-normal distributions often indicate an active expansion of clonotypes that often occurs during the adaptive immune response.

5.2.2.2 Spectratype Plot

The spectratype plot emulates an immunological assay where TCR sequence representation was assayed by length-based electrophoresis rather than sequencing [98]. Spectratypes can be interpreted as a histogram of observed CDR3 sequences binned by nucleotide length (Figure 5.3). In general, the spectratype shows spikes on CDR3 lengths that are multiples of three. This effect is attributable to the selection of coding TCR during T-cell development. CDR3s whose lengths are lengths are a multiple of 3 are distributed as a gaussian distribution in healthy adults who are not undergoing an active immune response. Large deviations from a normal distribution are often viewed as evidence of a clonal expansion.

Although spectratypes are commonly used for the study of TCR repertoire, they do not incorporate additional information from sequencing experiments such as V

and J gene segment usage.

5.2.2.3 VDJ Scatterplots

Visualizing VDJ gene segment usage is commonly performed using methods capable of displaying three dimensions of data. The dimensions represented in VDJ gene segment usage plots are the V, J and sometimes D gene segments used in recombination along with a number representing the overall abundance for a unique VDJ combination. Heatmaps, scatterplots, three dimensional histograms and chord diagrams are useful methods for displaying VDJ gene segment usage. However, comparing multiple samples in a single plot adds an addition dimension that the aforementioned techniques can not easily accommodate.

I address this issue by color-encoding samples, using opacity to display all samples simultaneously and utilizing interactivity to highlight individual samples or VJ combinations (Figure 5.4.A). Scatterplots display clonotype frequency based on joint V and J gene segment usage rather than CDR3 length. V and J gene segments are laid out on the X and Y axis respectively. Color-coded circles representing the sum of observed frequencies are then plotted. Circles with larger sizes represent higher observed frequencies of a V-J gene segment pair within a sample. However, studies have shown that the human ability to correlate area with a number is limited.

I again utilize interactivity to address this issue, preserving obvious visual differences in circle size while simultaneously providing numerical frequencies on demand. When a circle is highlighted with a mouse cursor, a tool tip displays decimal values of joint V-J usage frequencies for each sample (Figure 5.4.B).

In some instances, users may be more interested in the marginal V or J gene-segment usage rather than a joint V-J usage. To address this, V and J gene segment usage histograms flank the main scatterplot and share the same axis labels. By default, these histograms show the marginal usage of V and J gene segments in the dataset (Figure 5.4.C). However, the histograms can display a specific V or J gene segment usage histogram when an axis label is highlighted.

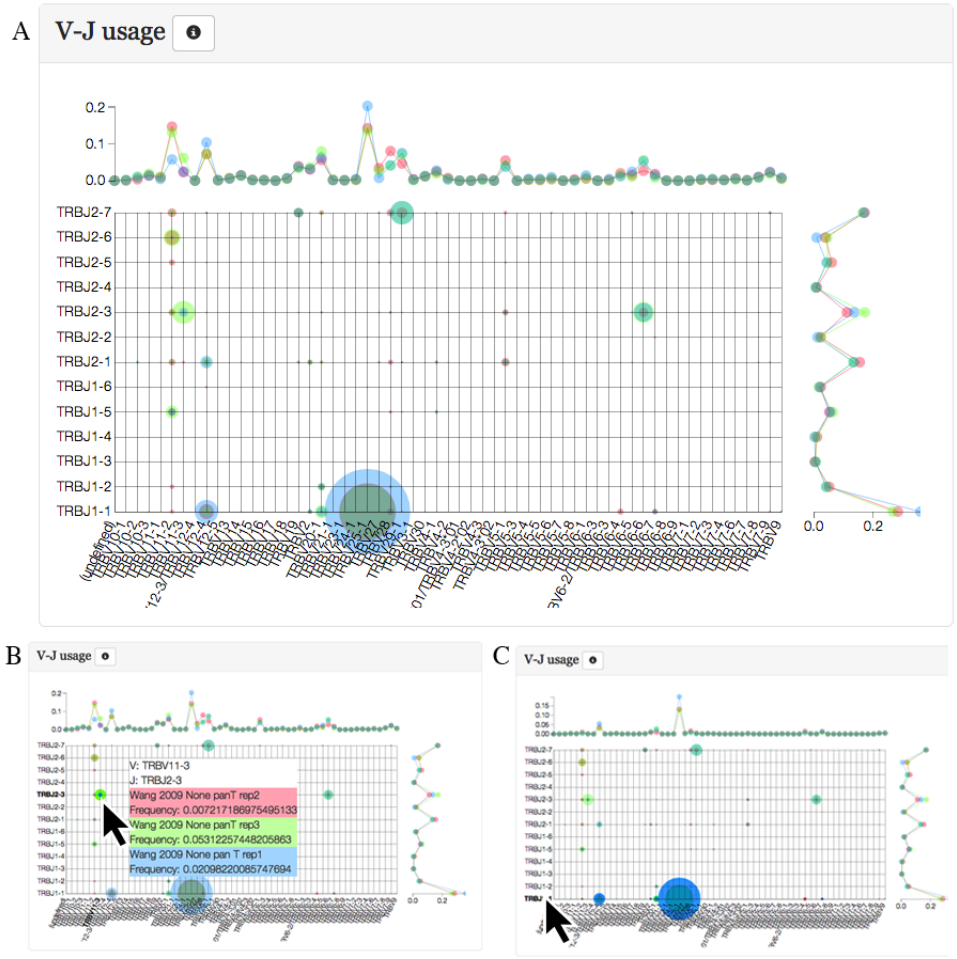

Figure 5.4: A scatterplot showing joint frequencies of V and J gene segment usage in the Compare view. Larger circles indicate larger frequencies. Histograms above and to the right of the scatterplot show the marginal frequencies of V or J genes. B) Hovering a mouse cursor over a circle shows the numerical frequencies of each sample in a popover tooltip. C) Hovering over an axis label updates the histogram to show frequencies of the selected gene segment. Here the above histogram shows the frequencies of V gene-segments given the highlighted J gene-segment, TRBJ1-1.

5.2.2.4 Functionality Plot

Although scatterplots and spectratypes are useful tools for repertoire visualization, not all TCR sequences in the repertoire encode functional receptors. The functionality plot is a stacked bar plot that shows the proportion of the sample representing coding TCRs vs sequences with stop codons and frameshift mutations (Figure

5.5). This plot allows researchers to assay enrichment for functional TCR sequences in a dataset, a metric that is especially useful for RNA-based sequencing assays where all detected clonotypes should be functional.

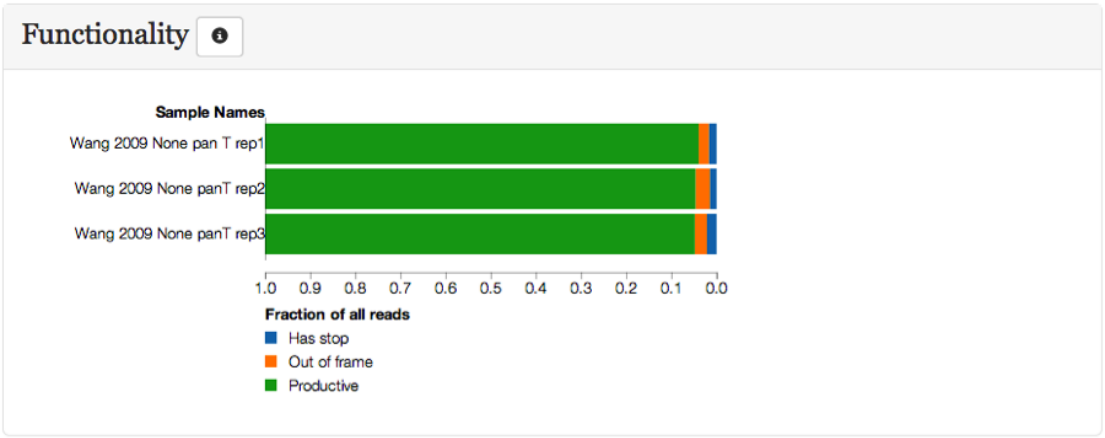

Figure 5.5: A stacked bar plot shows the fraction of protein coding status for each sample in the Compare view. During the formation of a T-cell receptor, the VDJ recombination process generates new nucleotide sequences that define the protein structure of the receptor itself. However, this process does not guarantee that each newly-generated receptor will produce a functional protein. A generated protein sequence can be: functional, contain a stop codon, or is out-of-frame. Generally samples from RNA will contain more functional receptors than a DNA sample.

5.2.2.5 Domination Plot

Clonotypes that represent large fractions of a total repertoire are especially interesting in tracking of blood cancers. A domination plot shows the total cumulative fraction of the repertoire that is represented by the top 100 clones with the largest frequency. A quickly rising line represents large clonal expansions that dominate the repertoire, while a slowly rising line indicates a repertoire with a more uniform distribution of clonotype frequencies (Figure 5.6).

5.2.2.6 Shared Sequence Plot

Identifying shared TCR sequences is useful in samples drawn from the same individual or for detecting "public" TCRs used to detect exposure to common pathogens.

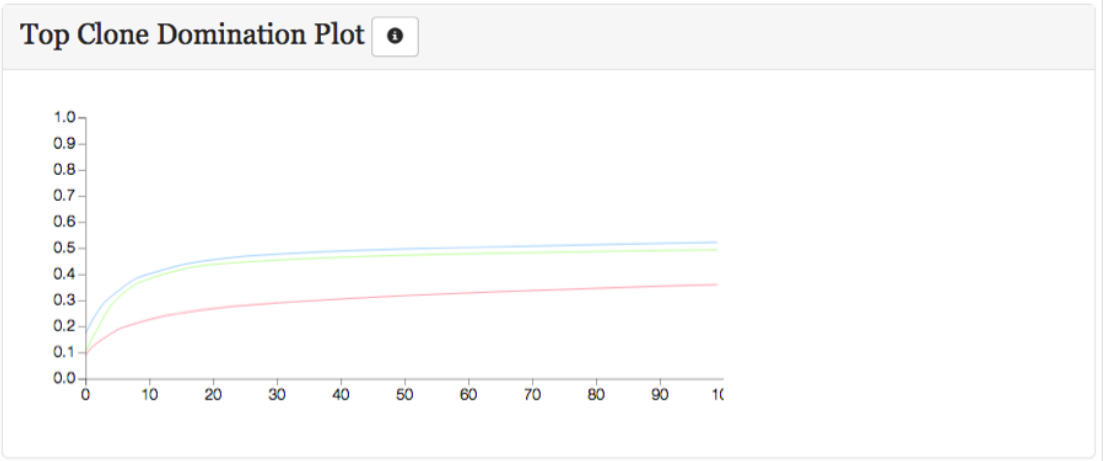

Figure 5.6: A line plot shows the shows the cumulative frequency (Y-axis) of the top 100 most abundant clonotypes for each sample (X-axis) in the Compare view. Larger area under the curve indicate lower levels of diversity in a sample.

The shared clones plot displays frequencies of amino acid sequences shared between all samples in a comparison. Frequencies for shared amino acid sequences are displayed as both a line plot and in a table. All shared amino acids contain a link to the amino acid detail view. Amino acids highlighted in the plot are also highlighted in the table, and vice versa (Figure 5.7).

5.2.2.7 TCR Sequence View

In addition to interactive sample comparison views, detailed views exist for individual clonotypes, recombination and amino acids. These views can be accessed via the "view all clonotypes" summary table, or from the shared amino acid plot in the comparison view. These views display the total number of reads represented in a clonotype, the V, D and J gene segments used in a recombination as well as the final coding amino acid sequence (if applicable) (Figure 5.8).

5.2.2.8 Literature Search

Users can search our database of CDR3 sequences mined from the literature by inputting their protein sequence in FASTA format. Alternatively, users can access this page from the amino acid detail view. Any resulting hits to the Literature database are

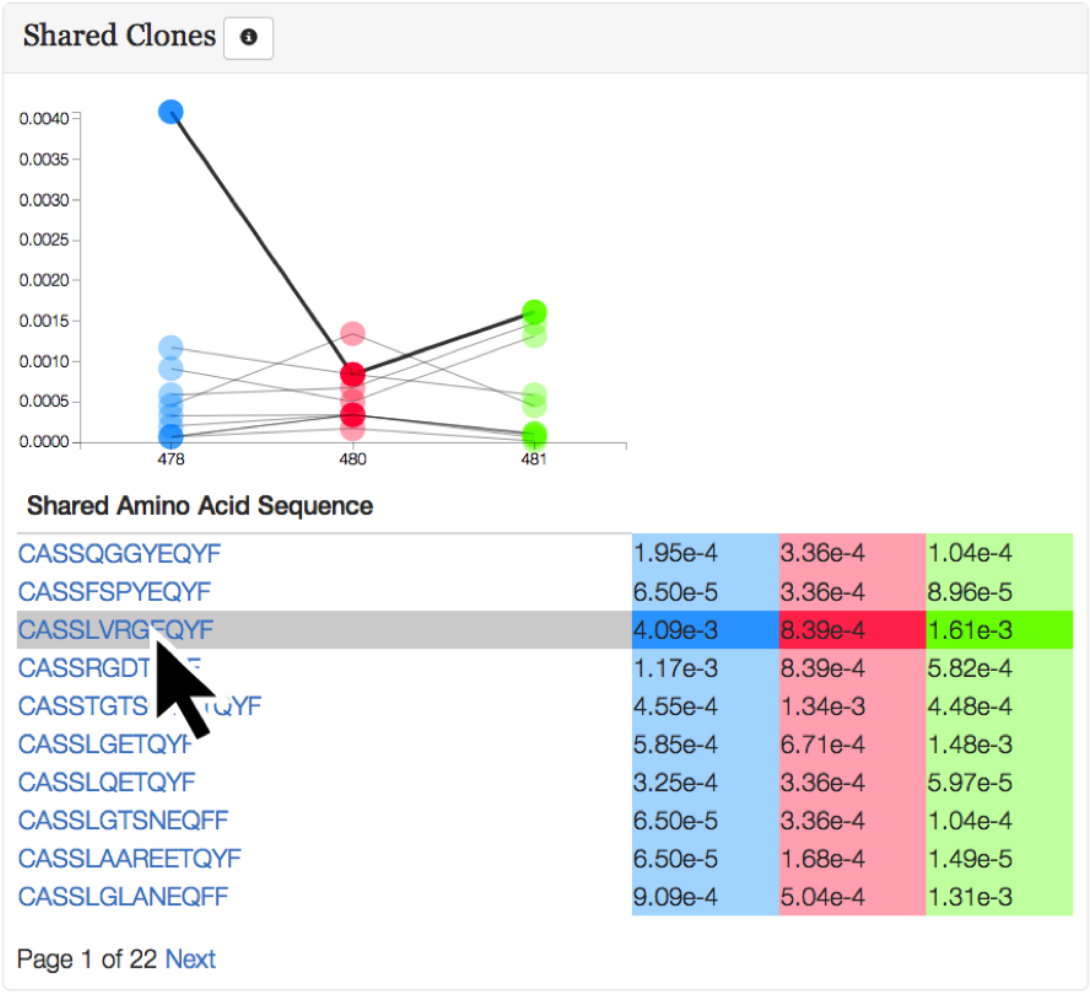

Figure 5.7: A graphical and tabular summary of shared amino acid sequences in the Compare view. *Upper Panel*: Frequency (Y-axis) of amino acid sequences in a sample is displayed as colored circles. Amino acid sequences across samples are represented as grey lines connecting circles. *Lower Panel*: A table displays the frequency of a shared amino acid for all samples. Hovering the mouse cursor over a row of the table highlights the corresponding line in the line plot and vice versa.

displayed with query-sequence alignments and links to the original publication (Figure 5.9).

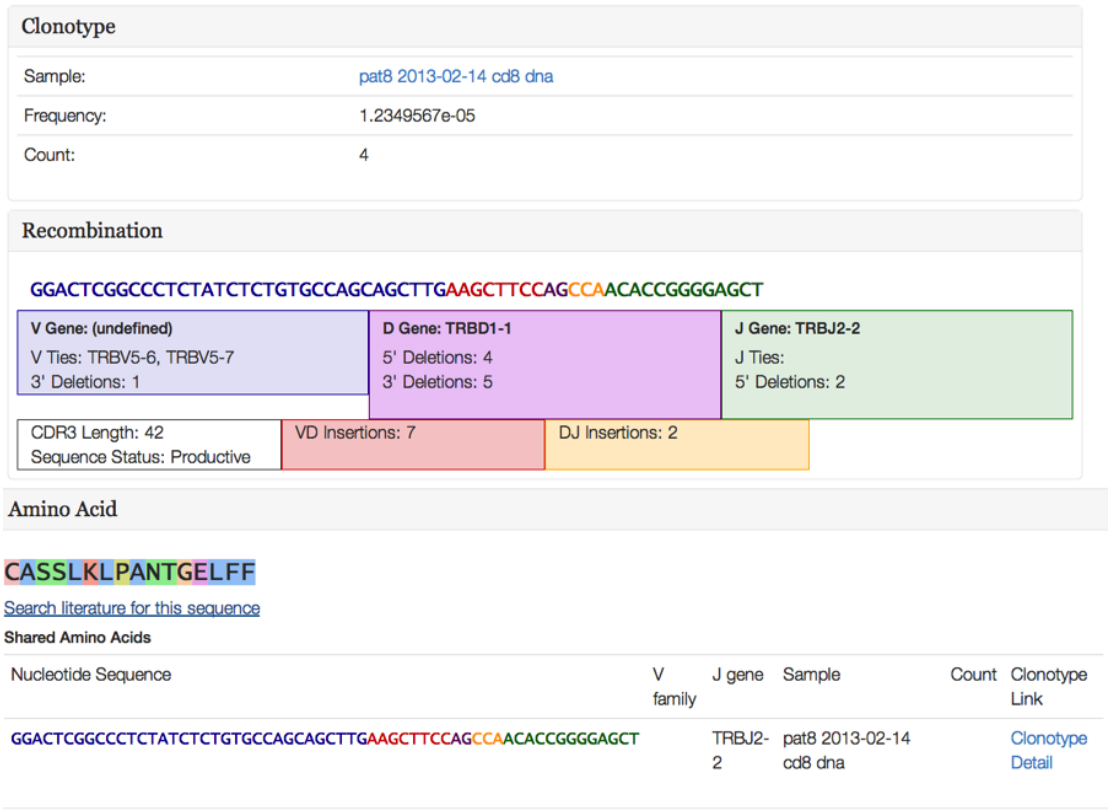

Figure 5.8: The detail view for a single clonotype. (Top Panel) The frequency and count of the clonotypes within a sample are shown. (Middle Panel) Recombination detail view shows the nucleotide sequence colorized according to membership of subsequence to a gene segment or insertion. (Bottom Panel) Amino acid detail view shows amino acid sequence colorized using a modified clustalw color scheme. A link allows simple access to the built in literature search tool. All clonotypes sharing the same amino acid sequence are listed in the table below.

5.3 Discussion

TCR sequencing experiments are utilized to study a variety of conditions involving T-cells in the adaptive immune system such as tracking residual disease in blood cancers, reconstitution of the adaptive immune system following bone marrow transplantation, identification of disease-specific clonotypes. Studies utilizing TCR sequencing perform custom analysis designed for each particular study.

The UCSC Immunobrowser provides a suite of tools that perform common

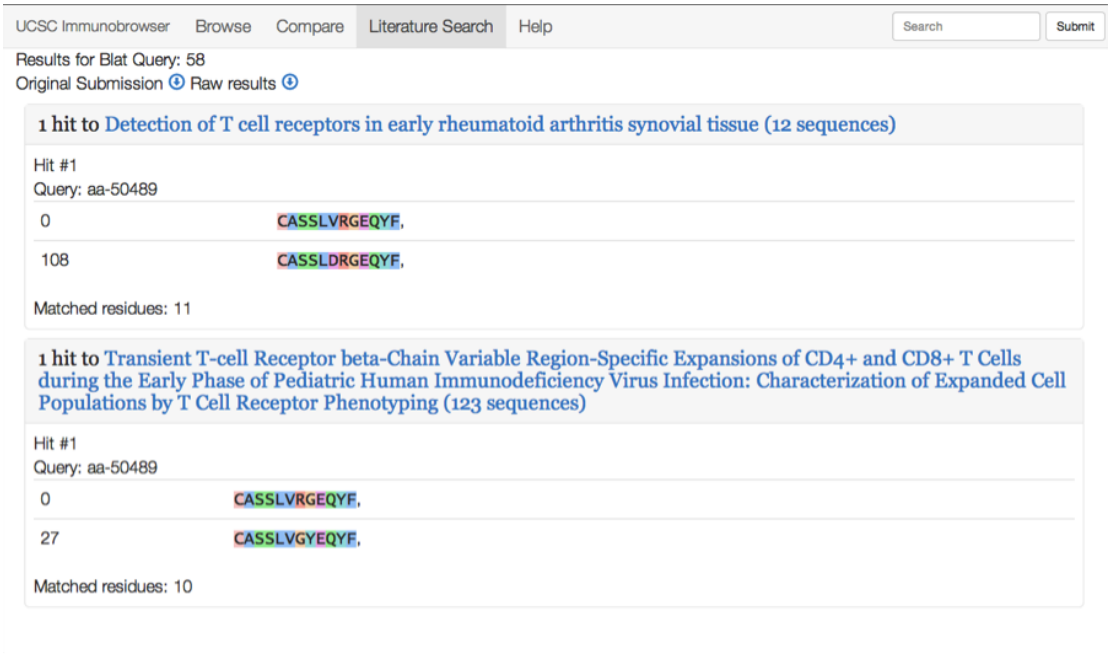

Figure 5.9: The results page of a literature search on the CDR3 amino acid sequence "CASSLVRGEQYF". Amino acid sequences are colorized using a modified clustalw color scheme.

analytic tasks of TCR sequencing experiments. For example, repertoire-level visualizations similar to the V-J scatterplot are present in numerous publications. However, the Immunobrowser leverages color, opacity and interactivity to display multiple samples in a single figure that simplify comparisons of V-J representation across multiple samples. For studies on immune system reconstitution, the ability to view all different time points in an individual's recovery simultaneously gives researchers a strong idea of when the immune system has returned to normal functioning levels.

Another common task is the identification, tracking and further investigation of aberrant clones. These clones can be identified in a number of ways using the Immunobrowser. Large clonal expansions are visually evident in the scatterplot and spectratype. These visualizations help researchers identify the V/J gene segments and CDR3 lengths of aberrant clones. The shared clone view directly tracks the frequency of the same amino acid sequence across multiple samples, which simplifies the spotting of large differences in the represented frequency of a clone in a repertoire. Once a TCR amino

acid sequence of interest has been identified, a literature search can identify previously published TCR's with similar sequences.

Both repertoire and clonotype level analysis can be easily performed on the Immunobrowser. As more data is imported into the browser, it becomes more useful as a public resource for TCR sequencing experiments. Future improvements may allow for user-uploaded sequences and data privacy.

Chapter 6

Closing Remarks

The utilities of HTS techniques are not limited to those presented within this dissertation. Because biology is exceedingly complex, assays and algorithms must consider each complexity individually. With improving technology, it is feasible that the sequencer follows the path of the computer. Whereas modern sequencers capable of sequencing a genome are relegated to large research institution, small personal sequencers affordable to the hobbyist are beginning to emerge. The personal sequencer may some day be as ubiquitous as the smart phone. Bioinformatic advancements are integral to the development of personalized medicine beyond treatment of disease and towards the real time monitoring of health.

Bibliography

[1] Paolo Actis, Michelle M Maalouf, Hyunsung John Kim, Akshar Lohith, Boaz Vilozny, R Adam Seger, and Nader Pourmand. Compartmental genomics in living cells revealed by single-cell nanobiopsy. *ACS nano*, 8(1):546–553, January 2014.

[2] Xian Adiconis, Diego Borges-Rivera, Rahul Satija, David S DeLuca, Michele A Busby, Aaron M Berlin, Andrey Sivachenko, Dawn Anne Thompson, Alec Wysoker, Timothy Fennell, Andreas Gnirke, Nathalie Pochet, Aviv Regev, and Joshua Z Levin. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods*, 10(7):623–629, July 2013.

[3] Keiko Akagi, Jingfeng Li, Tatevik R Broutian, Hesed Padilla-Nash, Weihong Xiao, Bo Jiang, James W Rocco, Theodoros N Teknos, Bhavna Kumar, Danny Wangsa, Dandan He, Thomas Ried, David E Symer, and Maura L Gillison. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome research*, 24(2):gr.164806.113–199, November 2013.

[4] S Anders. Differential expression analysis for sequence count data. *Genome Biology*, 2010.

[5] Christopher D Armour, John C Castle, Ronghua Chen, Tomas Babak, Patrick Loerch, Stuart Jackson, Jyoti K Shah, John Dey, Carol A Rohl, Jason M Johnson, and Christopher K Raymond. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature methods*, 6(9):647–649, September 2009.

[6] Paul Ashwood, Paula Krakowiak, Irva Hertz-Picciotto, Robin Hansen, Isaac N

Pessah, and Judy Van de Water. Altered T cell responses in children with autism. *Brain, Behavior, and Immunity*, 25(5):840–849, July 2011.

[7] G Astori, D Lavergne, C Benton, B Höckmayr, K Egawa, C Garbe, and E M de Villiers. Human papillomaviruses are commonly found in normal skin of immunocompetent hosts. *The Journal of investigative dermatology*, 110(5):752–755, May 1998.

[8] C C Baker, W C Phelps, V Lindgren, M J Braun, M A Gonda, and P M Howley. Structural and transcriptional analysis of human papillomavirus type 16 sequences in cervical carcinoma cell lines. *Journal of virology*, 61(4):962–971, April 1987.

[9] A M Bal, A Kumar, and I M Gould. Antibiotic heterogeneity: from concept to practice. *Annals of the New York Academy of Sciences*, 1213(1):81–91, 2010.

[10] Fernando Baquero, Maria-Cristina Negri, Maria-Isabel Morosini, and Jesús Blázquez. Antibiotic-Selective Environments. *Clinical Infectious Diseases*, 27(Supplement 1):S5–S11, August 1998.

[11] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and David L Wheeler. GenBank. *Nucleic Acids Research*, 33(Database issue):D34–8, January 2005.

[12] G Bentley, R Higuchi, B Hoglund, D Goodridge, D Sayer, E A Trachtenberg, and H A Erlich. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue antigens*, 74(5):393–403, November 2009.

[13] Sebastian Boegel, Martin Lower, Michael Schafer, Thomas Bukur, Jos de Graaf, Valesca Boisguerin, Ozlem Tureci, Mustafa Diken, John C Castle, and Ugur Sahin. HLA typing from RNA-Seq sequence reads. *Genome medicine*, 4(12):102, December 2012.

[14] Dmitriy A Bolotin, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, Maria A Turchaninova, Ivan V Zvyagin, Olga V Britanova, and Dmitriy M Chudakov. MiTCR: software for T-cell receptor sequencing data analysis. *Nature methods*, 10(9):813–814, September 2013.

[15] M Bostock, V Ogievetsky, and J Heer. D³ Data-Driven Documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, December 2011.

[16] Scott D Boyd, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew Z Fire. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine*, 1(12):12ra23, December 2009.

[17] D Breitfeld, L Ohl, E Kremmer, and J Ellwart. Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *The Journal of ...*, 2000.

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

[19] Catherine A Brownstein, Alan H Beggs, Nils Homer, Barry Merriman, Timothy W Yu, Katherine C Flannery, Elizabeth T DeChene, Meghan C Towne, Sarah K Savage, Emily N Price, Ingrid A Holm, Lovelace J Luquette, Elaine Lyon, Joseph Majzoub, Peter Neupert, David McCallie, Peter Szolovits, Huntington F Willard, Nancy J Mendelsohn, Renee Temme, Richard S Finkel, Sabrina W Yum, Livija Medne, Shamil R Sunyaev, Ivan Adzhubey, Christopher A Cassa, Paul IW de Bakker, Hatice Duzkale, Piotr Dworzy ski, and William Fairbrother. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biology*, 15(3):R53, March 2014.

[20] Centers for Disease Control and Prevention (CDC). Update to CDC's Sexually transmitted diseases treatment guidelines, 2010: oral cephalosporins no longer a recommended treatment for gonococcal infections., August 2012.

[21] A Chatterjee and K Moffatt. Vaccines and autism—an unlikely connection. *Autism—a neurodevelopmental journey ...*, 2011.

[22] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, April 2013.

[23] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, Victor Sementchenko, Antonio Piccolboni, Stefan Bekiranov, Dione K Bailey, Madhavan Ganesh, Srinka Ghosh, Ian Bell, Daniela S Gerhard, and Thomas R Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science (New York, N.Y.)*, 308(5725):1149–1154, May 2005.

[24] Ye Cheng, Jeffrey Francis Quinn, and Lauren Anne Weiss. An eQTL mapping approach reveals that rare variants in the SEMA5A regulatory network impact autism risk. *Human molecular genetics*, 22(14):2960–2972, July 2013.

[25] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–640, October 2004.

[26] S Consuegra, A Ellison, J Allainguillaume, J Pachebat, K M Peat, and P Wright. Balancing selection and the maintenance of MHC supertype variation in a selfing vertebrate. *Proceedings of the Royal Society B: Biological Sciences*, 280(1754):20122854, March 2013.

[27] J Couturier, X Sastre-Garau, S Schneider-Maunoury, A Labib, and G Orth. Integration of papillomavirus DNA near myc genes in genital carcinomas and its consequences for proto-oncogene expression. *Journal of virology*, 65(8):4534–4538, August 1991.

[28] Paul I W de Bakker, Gil McVean, Pardis C Sabeti, Marcos M Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J Monsuur, Pamela Whittaker, Marcos Delgado, Jonathan Morrison, Angela Richardson, Emily C Walsh, Xiaojiang

Gao, Luana Galver, John Hart, David A Hafler, Margaret Pericak-Vance, John A Todd, Mark J Daly, John Trowsdale, Cisca Wijmenga, Tim J Vyse, Stephan Beck, Sarah Shaw Murray, Mary Carrington, Simon Gregory, Panos Deloukas, and John D Rioux. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature genetics*, 38(10):1166–1172, October 2006.

[29] Ethel-Michele de Villiers, Claude Fauquet, Thomas R Broker, Hans-Ulrich Bernard, and Harald zur Hausen. Classification of papillomaviruses. *Virology*, 324(1):17–27, June 2004.

[30] M R Deibel, L K Riley, M S Coleman, M L Cibull, S A Fuller, and E Todd. Expression of terminal deoxynucleotidyl transferase in human thymus during ontogeny and development. *Journal of immunology (Baltimore, Md. : 1950)*, 131(1):195–200, July 1983.

[31] Django. The Web framework for perfectionists with deadlines — Django.

[32] A Doja and W Roberts. Immunizations and autism: a review of the literature. *The Canadian Journal of Neurological Sciences*, 2006.

[33] K Drlica and X Zhao. DNA gyrase, topoisomerase IV, and the 4-quinolones. *Microbiology and Molecular Biology Reviews*, 61(3):377–392, September 1997.

[34] M Dürst, C M Croce, L Gissmann, E Schwarz, and K Huebner. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. *Proceedings of the National Academy of Sciences of the United States of America*, 84(4):1070–1074, February 1987.

[35] A M Elliott, S E Berning, M D Iseman, and C A Peloquin. Failure of drug penetration and acquisition of drug resistance in chronic tuberculous empyema. *Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 76(5):463–467, October 1995.

[36] Krisztina Rigó Tim Hague Attila Bérces Szilveszter Juhos Endre Major. HLA

Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data. *PLoS ONE*, 8(11):e78410, 2013.

[37] Rachel L Erlich, Xiaoming Jia, Scott Anderson, Eric Banks, Xiaojiang Gao, Mary Carrington, Namrata Gupta, Mark A DePristo, Matthew R Henn, Niall J Lennon, and Paul I W de Bakker. Next-generation sequencing for HLA typing of class I loci. *BMC genomics*, 12(1):42, 2011.

[38] Eveline Farias-Hesson, Jonathan Erikson, Alexander Atkins, Peidong Shen, Ronald W Davis, Curt Scharfe, and Nader Pourmand. Semi-automated library preparation for high-throughput DNA sequencing platforms. *Journal of biomedicine & biotechnology*, 2010:617469, 2010.

[39] M J Ferber, D P Montoya, C Yu, I Aderca, A McGee, E C Thorland, D M Nagorney, B S Gostout, L J Burgart, L Boix, J Bruix, B J McMahon, T H Cheung, T K H Chung, Y F Wong, D I Smith, and L R Roberts. Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. *Oncogene*, 22(24):3813–3820, June 2003.

[40] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit,

Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-Fran ccedil ois Olivier, Michael S Phillips, St eacute phanie Roumy, Cl eacute mentine Sall eacute e, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe rsquo er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gon ccedil alo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Todd Johnson, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, Imtaz Yakub, Bruce W Birren, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg,

Paul L rsquo Archev ecirc que, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature ...*, 449(7164):851–861, October 2007.

[41] J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert A Holt. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research*, 19(10):1817–1824, October 2009.

[42] Christian Gabriel, Martin Danzer, Christa Hackl, Guido Kopal, Peter Hufnagl, Katja Hofer, Helene Polin, Stephanie Stabentheiner, and Johannes Pröll. Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Human Immunology*, 70(11):960–964, November 2009.

[43] Bruno A Gaëta, Harald R Malming, Katherine J L Jackson, Michael E Bain, Patrick Wilson, and Andrew M Collins. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics (Oxford, England)*, 23(13):1580–1587, July 2007.

[44] Baback Gharizadeh, Michael Akhras, Magnus Unemo, Bengt Wretlind, Pål Nyrén, and Nader Pourmand. Detection of gyrA mutations associated with ciprofloxacin resistance in Neisseria gonorrhoeae by rapid and reliable pre-programmed short DNA sequencing. *International Journal of Antimicrobial Agents*, 26(6):486–490, December 2005.

[45] Véronique Giudicelli, Xavier Brochet, and Marie-Paule Lefranc. IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences. *Cold Spring Harbor protocols*, 2011(6), 2011.

[46] Arielle Glatman Zaretsky, Justin J Taylor, Irah L King, Fraser A Marshall, Markus

Mohrs, and Edward J Pearce. T follicular helper cells differentiate from Th2 cells in response to helminth antigens. *The Journal of experimental medicine*, 206(5):991–999, May 2009.

[47] Namraj Goire, Monica M Lahra, Marcus Chen, Basil Donovan, Christopher K Fairley, Rebecca Guy, John Kaldor, David Regan, James Ward, Michael D Nissen, Theo P Sloots, and David M Whiley. Molecular approaches to enhance surveillance of gonococcal antimicrobial resistance. *Nature reviews. Microbiology*, 12(3):223–229, March 2014.

[48] Jeffrey L Goldberg, Mauricio E Vargas, Jack T Wang, Wim Mandemakers, Stephen F Oster, David W Sretavan, and Ben A Barres. An Oligodendrocyte Lineage-Specific Semaphorin, Sema5A, Inhibits Axon Growth by Retinal Ganglion Cells. *The Journal of Neuroscience*, 24(21):4989–4999, May 2004.

[49] Ananda W Goldrath and Michael J Bevan. Selecting and maintaining a diverse T-cell repertoire. *Nature ...*, 402:6–13, December 1999.

[50] Yonatan H Grad, Robert D Kirkcaldy, David Trees, Janina Dordel, Simon R Harris, Edward Goldstein, Hillard Weinstock, Julian Parkhill, William P Hanage, Stephen Bentley, and Marc Lipsitch. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. *The Lancet infectious diseases*, 14(3):220–226, March 2014.

[51] Christiane Gras, Britta Eiz Vesper, Yarua Jaimes, Stephan Immenschuh, Roland Jacobs, Torsten Witte, Rainer Blasczyk, and Constança Figueiredo. Secreted semaphorin 5A activates immune effector cells and is a biomarker for rheumatoid arthritis. *Arthritis & rheumatology (Hoboken, N.J.)*, February 2014.

[52] G Gross, H Pfister, and M Hagedorn. Correlation between human papillomavirus (HPV) type and histology of warts. *Journal of Investigative ...*, 1982.

[53] Jian Han, David C Swan, Sharon J Smith, Shanjuan H Lum, Susan E Sefers, Elizabeth R Unger, and Yi-Wei Tang. Simultaneous amplification and identification of 25 human papillomavirus types with Templex technology. *Journal of clinical microbiology*, 44(11):4157–4162, November 2006.

[54] P Heisig. Genetic evidence for a role of parC mutations in development of high-level fluoroquinolone resistance in Escherichia coli. *Antimicrobial agents and chemotherapy*, 40(4):879–885, April 1996.

[55] Gloria Y F Ho, Robert Bierman, Leah Beardsley, Chee J Chang, and Robert D Burk. Natural History of Cervicovaginal Papillomavirus Infection in Young Women. *New England Journal of Medicine*, 338(7):423–428, February 1998.

[56] C L Holcomb, B Hoglund, M W Anderson, L A Blake, I Böhme, M Egholm, D Ferriola, C Gabriel, S E Gelber, D Goodridge, S Hawbecker, R Klein, M Ladner, C Lind, D Monos, M J Pando, J Pröll, D C Sayer, G Schmitz-Agheguian, B B Simen, B Thiele, E A Trachtenberg, D B Tyan, R Wassmuth, S White, and H A Erlich. A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue antigens*, 77(3):206–217, March 2011.

[57] Fan Hsu, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. The UCSC Known Genes. *Bioinformatics (Oxford, England)*, 22(9):1036–1046, May 2006.

[58] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–594, February 2012.

[59] Guillaume Huguet, Elodie Ey, and Thomas Bourgeron. The genetic landscapes of autism spectrum disorders. *Annual review of genomics and human genetics*, 14:191–213, 2013.

[60] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, February 2012.

[61] Katherine J L Jackson, Scott Boyd, Bruno A Gaëta, and Andrew M Collins. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics (Oxford, England)*, 26(24):3129–3130, December 2010.

[62] George A Jacoby. Mechanisms of resistance to quinolones. *Clinical Infectious Diseases*, 41 Suppl 2:S120–6, July 2005.

[63] Alain Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature reviews. Genetics*, 10(12):833–844, December 2009.

[64] Alfred B Jenson, Stanley Geyer, John P Sundberg, and Shin-je Ghim. Human Papillomavirus and Skin Cancer. *Journal of Investigative Dermatology Symposium Proceedings*, 6(3):203–206, December 2001.

[65] S Jeon, B L Allen-Hoffmann, and P F Lambert. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *Journal of virology*, 69(5):2989–2997, May 1995.

[66] Ning Jiang, Joshua A Weinstein, Lolita Penland, Richard A White, Daniel S Fisher, and Stephen R Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 108(13):5348–5353, March 2011.

[67] Mina Kalantari, Frank Karlsen, Gunnar Kristensen, Ruth Holm, Bjorn Hagmar, and Bo Johansson. Disruption of the E1 and E2 Reading Frames of HPV 16 in Cervical Carcinoma Is Associated with Poor Prognosis. *International Journal of Gynecologic Pathology*, 17(2):146–153, April 1998.

[68] Vera M Kalscheuer, David FitzPatrick, Niels Tommerup, Merete Bugge, Erik Niebuhr, Luitgard M Neumann, Andreas Tzschach, Sarah A Shoichet, Corinna Menzel, Fikret Erdogan, Ger Arkesteijn, Hans-Hilger Ropers, and Reinhard Ullmann. Mutations in autism susceptibility candidate 2 (AUTS2) in patients with mental retardation. *Human genetics*, 121(3-4):501–509, 2007.

[69] Dione Kampa, Jill Cheng, Philipp Kapranov, Mark Yamanaka, Shane Brubaker, Simon Cawley, Jorg Drenkow, Antonio Piccolboni, Stefan Bekiranov, Gregg Helt, Hari Tammana, and Thomas R Gingeras. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome research*, 14(3):331–342, March 2004.

[70] Takahiro Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering*, 96(4):317–323, 2003.

[71] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A Harte, Steve Heitner, Angie S Hinrichs, Katrina Learned, Brian T Lee, Chin H Li, Brian J Raney, Brooke Rhead, Kate R Rosenbloom, Cricket A Sloan, Matthew L Speir, Ann S Zweig, David Haussler, Robert M Kuhn, and W James Kent. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42(Database issue):D764–70, January 2014.

[72] Richard B Kennedy, Inna G Ovsyannikova, V Shane Pankratz, Iana H Haralambieva, Robert A Vierkant, and Gregory A Poland. Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Human genetics*, 131(9):1403–1421, 2012.

[73] Paul J Kersey, Daniel M Staines, Daniel Lawson, Eugene Kulesha, Paul Derwent, Jay C Humphrey, Daniel S T Hughes, Stephan Keenan, Arnaud Kerhornou, Gautier Koscielny, Nicholas Langridge, Mark D McDowall, Karine Megy, Uma Maheswari, Michael Nuhn, Michael Paulini, Helder Pedro, Iliana Toneva, Derek Wilson, Andrew Yates, and Ewan Birney. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Research*, 40(Database issue):D91–7, January 2012.

[74] C H Kim, L S Rott, and I Clark-Lewis. Subspecialization of Cxcr5+ T Cells B Helper Activity Is Focused in a Germinal Center–Localized Subset of Cxcr5+ T Cells. *The Journal of . . .*, 2001.

[75] D Kim, G Pertea, C Trapnell, H Pimentel, and R Kelley. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome . . .*, 2013.

[76] T J Kindt, R A Goldsby, B A Osborne, and J Kuby. Kuby immunology. 2007.

[77] Irah L King and Markus Mohrs. IL-4-producing CD4+ T cells in reactive lymph

nodes during helminth infection are T follicular helper cells. *The Journal of experimental medicine*, 206(5):1001–1007, May 2009.

[78] PL Klarenbeek, PP Tak, and BDC van Schaik. Human T-cell memory consists mainly of unexpanded clones. *Immunology letters*, 2010.

[79] Nurith Kurn, Pengchin Chen, Joe Don Heath, Anne Kopf-Sill, Kathryn M Stephens, and Shenglong Wang. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clinical chemistry*, 51(10):1973–1981, October 2005.

[80] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H Waterston, Richard K Wilson, LaDeana W Hillier, John D McPherson, Marco A Marra, Elaine R Mardis, Lucinda A Fulton, Asif T Chinwalla, Kymberlie H Pepin, Warren R Gish, Stephanie L Chissoe, Michael C Wendl, Kim D Delehaunty, Tracie L Miner, Andrew Delehaunty, Jason B Kramer, Lisa L Cook, Robert S Fulton, Douglas L Johnson, Patrick J Minx, Sandra W Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A Gibbs, Donna M Muzny, Steven E Scherer, John B Bouck,

Erica J Sodergren, Kim C Worley, Catherine M Rives, James H Gorrell, Michael L Metzker, Susan L Naylor, Raju S Kucherlapati, David L Nelson, George M Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, Andr eacute Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W Davis, Nancy A Federspiel, A Pia Abola, Michael J Proctor, Bruce A Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Bl ouml cker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L Aravind, Jeffrey A Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G Brown, Christopher B Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R Copley, Tobias Doerks, Sean R Eddy, Evan E Eichler, Terrence S Furey, James Galagan, James G R Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L Steven Johnson, Thomas A Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W James Kent, Paul Kitts, Eugene V Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V Moran, Nicola Mulder, Victor J Pollara, Chris P Ponting, Greg Schuler, J ouml rg Schultz, Guy Slater, Arian F A Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I Wolf, Kenneth H Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S Guyer, Jane Peterson, Adam Felsenfeld, Kris A Wetterstrand, Richard M Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R Cox, Maynard V Olson, and R... Kaul. Initial sequencing and analysis of the human genome. *Nature ...*, 409(6822):860–921, February 2001.

[81] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie

2. *Nature methods*, 9(4):357–359, April 2012.

[82] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[83] Simon M Lank, Roger W Wiseman, Dawn M Dudley, and David H O'Connor. A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class I genotyping. *Human Immunology*, 71(10):1011–1017, October 2010.

[84] Beatriz León, André Ballesteros-Tato, Jeffrey L Browning, Robert Dunn, Troy D Randall, and Frances E Lund. Regulation of T(H)2 development by CXCR5+ dendritic cells and lymphotoxin-expressing B cells. *Nature immunology*, 13(7):681–690, July 2012.

[85] Stuart B Levy and Bonnie Marshall. Antibacterial resistance worldwide: causes, challenges and responses. *Nature Medicine*, 10(12 Suppl):S122–9, December 2004.

[86] C Lind, D Ferriola, K Mackiewicz, S Heron, M Rogers, L Slavich, R Walker, T Hsiao, L McLaughlin, M D'Arcy, X Gai, D Goodridge, D Sayer, and D Monos. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Human Immunology*, 71(10):1033–1042, October 2010.

[87] Phyllis Jean Linton and Kenneth Dorshkind. Age-related changes in lymphocyte development and function. *Nature immunology*, 5(2):133–139, February 2004.

[88] Chang Liu, Xiao Yang, Brian Duffy, Thalachallour Mohanakumar, Robi D Mitra, Michael C Zody, and John D Pfeifer. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Research*, 41(14):e142–e142, August 2013.

[89] M I Love, W Huber, and S Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, 2014.

[90] Frank Luft, Ruediger Klaes, Matthias Nees, Matthias Dürst, Volker Heilmann, Peter Melsheimer, and Magnus von Knebel Doeberitz. Detection of integrated

papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells. *International Journal of Cancer*, 92(1):9–17, 2001.

[91] Katja Lüthje, Axel Kallies, Yoko Shimohakamada, Gabrielle T Belz, Amanda Light, David M Tarlinton, and Stephen L Nutt. The development and fate of follicular helper T cells defined by an IL-21 reporter mouse. *Nature immunology*, 13(5):491–498, May 2012.

[92] M Maechler, P Rousseeuw, A Struyf, and M Hubert. [CITATION][C]. *R package version*, 2012.

[93] Lira Mamanova, Robert M Andrews, Keith D James, Elizabeth M Sheridan, Peter D Ellis, Cordelia F Langford, Tobias W B Ost, John E Collins, and Daniel J Turner. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature methods*, 7(2):130–132, February 2010.

[94] Shrinivasrao P Mane, Clive Evans, Kristal L Cooper, Oswald R Crasta, Otto Folkerts, Stephen K Hutchison, Timothy T Harkins, Danielle Thierry-Mieg, Jean Thierry-Mieg, and Roderick V Jensen. Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC genomics*, 10:264, 2009.

[95] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*, 24(3):133–141, March 2008.

[96] Heather D Marshall, Anmol Chandele, Yong Woo Jung, Hailong Meng, Amanda C Poholek, Ian A Parish, Rachel Rutishauser, Weiguo Cui, Steven H Kleinstein, Joe Craft, and Susan M Kaech. Differential Expression of Ly6C and T-bet Distinguish Effector and Memory Th1 $CD4^+$ Cell Properties during Viral Infection. *Immunity*, 35(4):633–646, October 2011.

[97] Inga-Lill Mårtensson, Antonius Rolink, Fritz Melchers, Cornelia Mundt, Steve Licence, and Takeyuki Shimizu. The pre-B cell receptor and its role in proliferation and Ig heavy chain allelic exclusion. *Seminars in immunology*, 14(5):335–342, October 2002.

[98] Krystyna Maślanka, Teresa Piatek, Jessica Gorski, Maryam Yassai, Jack Gorski, and J Gorski. Molecular analysis of T cell repertoires. *Human Immunology*, 44(1):28–34, September 1995.

[99] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, 26(16):2069–2070, August 2010.

[100] John J Miles, Daniel C Douek, and David A Price. Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunology and cell biology*, 89(3):375–387, March 2011.

[101] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature ...*, 464(7289):773–777, April 2010.

[102] Rimpei Morita, Nathalie Schmitt, Salah-Eddine Bentebibel, Rajaram Ranganathan, Laure Bourdery, Gerard Zurawski, Emile Foucat, Melissa Dullaers, SangKon Oh, Natalie Sabzghabaei, Elizabeth M Lavecchio, Marilynn Punaro, Virginia Pascual, Jacques Banchereau, and Hideki Ueno. Human Blood CXCR5+CD4+ T Cells Are Counterparts of T Follicular Cells and Contain Specific Subsets that Differentially Support Antibody Secretion. *Immunity*, 34(1):108–121, January 2011.

[103] Olena Morozova, Martin Hirst, and Marco A Marra. Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics*, 10(1):135–151, 2009.

[104] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, July 2008.

[105] Anna-Barbara Moscicki, Nancy Hills, Steve Shiboski, Kim Powell, Naomi Jay, Evelyn Hanson, Susanna Miller, Lisa Clayton, Sepideh Farhat, Jeanette Broering,

Teresa Darragh, and Joel Palefsky. Risks for Incident Human Papillomavirus Infection and Low-Grade Squamous Intraepithelial Lesion Development in Young Females. *Jama*, 285(23):2995–3002, June 2001.

[106] Anna-Barbara Moscicki, Mark Schiffman, Susanne Kjaer, and Luisa L Villa. Chapter 5: Updating the natural history of HPV and anogenital cancer. *Vaccine*, 24:S42–S51, August 2006.

[107] Nubia Muñoz, F Xavier Bosch, Silvia de Sanjosé, Rolando Herrero, Xavier Castellsagué, Keerti V Shah, Peter J F Snijders, and Chris J L M Meijer. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *New England Journal of Medicine*, 348(6):518–527, February 2003.

[108] S Munshaw and T B Kepler. SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics (Oxford, England)*, 26(7):867–872, March 2010.

[109] K Murphy, P Travers, and M Walport. [CITATION][C]. *Inc New York and London*, 2008.

[110] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16161–16166, October 2012.

[111] D Nemazee. Revising B Cell Receptors. *Journal of Experimental Medicine*, 191(11):1813–1818, May 2000.

[112] Lori M Newman, John S Moran, and Kimberly A Workowski. Update on the management of gonorrhea in adults in the United States. *Clinical Infectious Diseases*, 44 Suppl 3(Supplement 3):S84–101, April 2007.

[113] G J Nossal. Negative selection of lymphocytes. *Cell*, 76(2):229–239, January 1994.

[114] Nir Oksenberg and Nadav Ahituv. The role of AUTS2 in neurodevelopment and human evolution. *Trends in genetics : TIG*, 29(10):600–608, January 2013.

[115] Nir Oksenberg, Laurie Stevison, Jeffrey D Wall, and Nadav Ahituv. Function and Regulation of AUTS2, a Gene Implicated in Autism and Human Evolution. *PLoS genetics*, 9(1):e1003221, January 2013.

[116] Charity Onore, Milo Careaga, and Paul Ashwood. The role of immune dysfunction in the pathophysiology of autism. *Brain, Behavior, and Immunity*, 26(3):383–392, March 2012.

[117] M C Orencia, J S Yoon, J E Ness, W P Stemmer, and R C Stevens. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature structural biology*, 8(3):238–242, March 2001.

[118] Fatih Ozsolak, Alon Goren, Melissa Gymrek, Mitchell Guttman, Aviv Regev, Bradley E Bernstein, and Patrice M Milos. Digital transcriptome profiling from attomole-level RNA samples. *Genome research*, 20(4):519–525, April 2010.

[119] P Parham, D A Lawlor, C E Lomen, and P D Ennis. Diversity and diversification of HLA-A,B,C alleles. *Journal of immunology (Baltimore, Md. : 1950)*, 142(11):3937–3950, June 1989.

[120] Vanja Paunić, Loren Gragert, Abeer Madbouly, John Freeman, and Martin Maiers. Measuring Ambiguity in HLA Typing Methods. *PLoS ONE*, 7(8):e43585, 2012.

[121] Panu Peitsaro, Bo Johansson, and Stina Syrjänen. Integrated human papillomavirus type 16 is frequently found in cervical cancer precursors as demonstrated by a novel quantitative real-time PCR technique. *Journal of clinical microbiology*, 40(3):886–891, March 2002.

[122] M F Polz and C M Cavanaugh. Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology*, 64(10):3724–3730, October 1998.

[123] Nicholas C Popescu and Joseph A DiPaolo. Preferential sites for viral integration on mammalian genome. *Cancer genetics and cytogenetics*, 42(2):157–171, October 1989.

[124] J Pröll, M Danzer, S Stabentheiner, N Niklas, C Hackl, K Hofer, S Atzmuller, P Hufnagl, C Gully, H Hauser, O Krieger, and C Gabriel. Sequence Capture and Next Generation Resequencing of the MHC Region Highlights Potential Transplantation Determinants in HLA Identical Haematopoietic Stem Cell Transplantation. *DNA Research*, 18(4):201–210, August 2011.

[125] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):341, July 2012.

[126] Ata Ur Rasheed, Hans Peter Rahn, Federica Sallusto, Martin Lipp, and Gerd Müller. Follicular B helper T cell activity is confined to CXCR5hiICOShi CD4 T cells and is independent of CD57 expression. *European journal of immunology*, 36(7):1892–1903, 2006.

[127] U Ravn, F Gueneau, L Baerlocher, M Osteras, M Desmurs, P Malinge, G Magistrelli, L Farinelli, M H Kosco-Vilbois, and N Fischer. By-passing in vitro screening–next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Research*, 38(21):e193–e193, November 2010.

[128] Sai T Reddy, Xin Ge, Aleksandr E Miklos, Randall A Hughes, Seung Hyun Kang, Kam Hon Hoi, Constantine Chrysostomou, Scott P Hunicke-Smith, Brent L Iverson, Philip W Tucker, Andrew D Ellington, and George Georgiou. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature biotechnology*, 28(9):965–969, September 2010.

[129] Sai T Reddy and George Georgiou. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Current opinion in biotechnology*, May 2011.

[130] R Lee Reinhardt, Hong-Erh Liang, and Richard M Locksley. Cytokine-secreting follicular T cells shape the antibody repertoire. *Nature immunology*, 10(4):385–393, April 2009.

[131] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, 114(19):4099–4107, November 2009.

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

[133] James Robinson, Jason A Halliwell, Hamish McWilliam, Rodrigo Lopez, Peter Parham, and Steven G E Marsh. The IMGT/HLA database. *Nucleic Acids Research*, 41(D1):D1222–D1227, January 2013.

[134] Brian D Rudd, Vanessa Venturi, Miles P Davenport, and Janko Nikolich-Zugich. Evolution of the antigen-specific CD8+ TCR repertoire across the life span: evidence for clonal homogenization of the old TCR repertoire. *Journal of immunology (Baltimore, Md. : 1950)*, 186(4):2056–2064, February 2011.

[135] Patrick Schaerli, Katharina Willimann, Alois B Lang, Martin Lipp, Pius Loetscher, and Bernhard Moser. Cxc Chemokine Receptor 5 Expression Defines Follicular Homing T Cells with B Cell Helper Function. *The Journal of experimental medicine*, 192(11):1553–1562, December 2000.

[136] Martina Schmitz, Corina Driesch, Katrin Beer Grondke, Lars Jansen, Ingo B Runnebaum, and Matthias Dürst. Loss of gene function as a consequence of human papillomavirus DNA integration. *International Journal of Cancer*, 131(5):E593–E602, 2012.

[137] Kevin L Schneider, Katherine S Pollard, Robert Baertsch, Andy Pohl, and Todd M Lowe. The UCSC Archaeal Genome Browser. *Nucleic Acids Research*, 34(Database issue):D407–10, January 2006.

[138] A Schneider-Gädicke and E Schwarz. Different human cervical carcinoma cell lines show similar transcription patterns of human papillomavirus type 18 early genes. *The EMBO journal*, 5(9):2285–2292, September 1986.

[139] Gunter Schumann, Lachlan J Coin, Anbarasu Lourdusamy, Pimphen Charoen, Karen H Berger, David Stacey, Sylvane Desrivières, Fazil A Aliev, Anokhi A Khan, Najaf Amin, Yurii S Aulchenko, Georgy Bakalkin, Stephan J Bakker, Beverley Balkau, Joline W Beulens, Ainhoa Bilbao, Rudolf A de Boer, Delphine Beury, Michiel L Bots, Elemi J Breetvelt, Stéphane Cauchi, Christine Cavalcanti-Proença, John C Chambers, Toni-Kim Clarke, Norbert Dahmen, Eco J de Geus, Danielle Dick, Francesca Ducci, Alanna Easton, Howard J Edenberg, Tõnu Esko, Tõnu Esk, Alberto Fernández-Medarde, Tatiana Foroud, Nelson B Freimer, Jean-Antoine Girault, Diederick E Grobbee, Simonetta Guarrera, Daniel F Gudbjartsson, Anna-Liisa Hartikainen, Andrew C Heath, Victor Hesselbrock, Albert Hofman, Jouke-Jan Hottenga, Matti K Isohanni, Jaakko Kaprio, Kay-Tee Khaw, Brigitte Kuehnel, Jaana Laitinen, Stéphane Lobbens, Jian'an Luan, Massimo Mangino, Matthieu Maroteaux, Giuseppe Matullo, Mark I McCarthy, Christian Mueller, Gerjan Navis, Mattijs E Numans, Alejandro Núñez, Dale R Nyholt, Charlotte N Onland-Moret, Ben A Oostra, Paul F O'Reilly, Miklos Palkovits, Brenda W Penninx, Silvia Polidoro, Anneli Pouta, Inga Prokopenko, Fulvio Ricceri, Eugenio Santos, Johannes H Smit, Nicole Soranzo, Kijoung Song, Ulla Sovio, Michael Stumvoll, Ida Surakk, Thorgeir E Thorgeirsson, Unnur Thorsteinsdottir, Claire Troakes, Thorarinn Tyrfingsson, Anke Tönjes, Cuno S Uiterwaal, Andre G Uitterlinden, Pim van der Harst, Yvonne T van der Schouw, Oliver Staehlin, Nicole Vogelzangs, Peter Vollenweider, Gerard Waeber, Nicholas J Wareham, Dawn M Waterworth, John B Whitfield, Erich H Wichmann, Gonneke Willemsen, Jacqueline C Witteman, Xin Yuan, Guangju Zhai, Jing H Zhao, Weihua Zhang, Nicholas G Martin, Andres Metspalu, Angela Doering, James Scott, Tim D Spector, Ruth J Loos, Dorret I Boomsma, Vincent Mooser, Leena Peltonen, Kari Stefansson, Cornelia M van Duijn, Paolo Vineis, Wolfgang H Sommer, Jaspal S Kooner, Rainer Spanagel, Ulrike A Heberlein, Marjo-Riitta Jarvelin, and Paul Elliott. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7119–7124, April 2011.

[140] Elisabeth Schwarz, Ulrich Karl Freese, Lutz Gissmann, Wolfgang Mayer, Birgit Roggenbuck, Armin Stremlau, and Harald zur Hausen. Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. *, Published online: 07 March 1985; — doi:10.1038/314111a0*, 314(6006):111–114, March 1985.

[141] Sally I Sharp, Andrew McQuillin, and Hugh M D Gurling. Genetics of attention-deficit hyperactivity disorder (ADHD). *Neuropharmacology*, 57(7-8):590–600, December 2009.

[142] B P Sleckman. Mechanisms that direct ordered assembly of T cell receptor beta locus V, D, and J gene segments. *Proceedings of the National Academy of Sciences*, 97(14):7975–7980, June 2000.

[143] M Margarida Souto-Carneiro, Nancy S Longo, Daniel E Russ, Hong-wei Sun, and Peter E Lipsky. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *Journal of immunology (Baltimore, Md. : 1950)*, 172(11):6790–6802, June 2004.

[144] Timothy K Starr, Stephen C Jameson, and Kristin A Hogquist. Positive and negative selection of T cells. *Annual review of immunology*, 21:139–176, 2003.

[145] GBE Stewart-Jones and AJ McMichael. A structural basis for immunodominant human T cell receptor recognition. *Nature ...*, 2003.

[146] Razia Sultana, Chang-En Yu, Jun Yu, Jeffery Munson, Donghui Chen, Wenhui Hua, Annette Estes, Fanny Cortes, Flora de la Barra, Dongmei Yu, Syed T Haider, Barbara J Trask, Eric D Green, Wendy H Raskind, Christine M Disteche, Ellen Wijsman, Geraldine Dawson, Daniel R Storm, Gerard D Schellenberg, and Enrique C Villacres. Identification of a Novel Gene on Chromosome 7q11.2 Interrupted by a Translocation Breakpoint in a Pair of Autistic Twins. *Genomics*, 80(2):129–134, August 2002.

[147] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin

Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, May 2009.

[148] Muhammad A Tariq, Hyunsung J Kim, Olufisayo Jejelowo, and Nader Pourmand. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, 39(18):e120, October 2011.

[149] Brent Taylor, Elizabeth Miller, CPaddy Farrington, Maria-Christina Petropoulos, Isabelle Favot-Mayaud, Jun Li, and Pauline A Waight. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *The Lancet*, 353(9169):2026–2029, June 1999.

[150] N Thomas, J Heather, W Ndifon, J Shawe-Taylor, and B Chain. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics (Oxford, England)*, 29(5):542–550, February 2013.

[151] Erik C Thorland, Shannon L Myers, Bobbie S Gostout, and David I Smith. Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene*, 22(8):1225–1237, February 2003.

[152] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, May 2009.

[153] Koenraad Van Doorslaer, Qina Tan, Sandhya Xirasagar, Sandya Bandaru, Vivek Gopalan, Yasmin Mohamoud, Yentram Huyen, and Alison A McBride. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*, 41(Database issue):D571–8, January 2013.

[154] Matthew Vanneman and Glenn Dranoff. Combining immunotherapy and targeted therapies in cancer treatment. *Nature reviews. Cancer*, 12(4):237–251, April 2012.

[155] Vanessa Venturi, PC Ng, ZS Ende, and T McIntosh. A Mechanism for TCR Sharing between T Cell Subsets and Individuals Revealed by Pyrosequencing. *The Journal of . . .*, 2011.

[156] Stephanie Verfuerth, Karl Peggs, Paulomi Vyas, Lorna Barnett, Richard J O'Reilly, and Stephen Mackinnon. Longitudinal monitoring of immune reconstitution by CDR3 size spectratyping after T-cell–depleted allogeneic bone marrow transplant and the effect of donor lymphocyte infusions on T-cell repertoire. *Blood*, 95(12):3990–3995, June 2000.

[157] Carola G Vinuesa, Sidonia Fagarasan, and Chen Dong. New Territory for T Follicular Helper Cells. *Immunity*, 39(3):417–420, September 2013.

[158] Ana P Vivancos, Marc Güell, Juliane C Dohm, Luis Serrano, and Heinz Himmelbauer. Strand-specific deep sequencing of the transcriptome. *Genome research*, 20(7):989–999, July 2010.

[159] Joseph M Volpe, Lindsay G Cowell, and Thomas B Kepler. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics (Oxford, England)*, 22(4):438–444, February 2006.

[160] Jan M M Walboomers, Marcel V Jacobs, M Michele Manos, F Xavier Bosch, J Alain Kummer, Keerti V Shah, Peter J F Snijders, Julian Peto, Chris J L M Meijer, and Nubia Muñoz. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of Pathology*, 189(1):12–19, September 1999.

[161] Chunlin Wang, Catherine M Sanders, Qunying Yang, Harry W Schroeder, Elijah Wang, Farbod Babrzadeh, Baback Gharizadeh, Richard M Myers, James R Hudson, Ronald W Davis, and Jian Han. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences of the United States of America*, 107(4):1518–1523, January 2010.

[162] Xiaojing Wang, Di Wu, Siyuan Zheng, Jing Sun, Lin Tao, Yixue Li, and Zhiwei Cao. Ab-origin: an enhanced tool to identify the sourcing gene segments in germline for rearranged antibodies. *BMC bioinformatics*, 9 Suppl 12:S20, 2008.

[163] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[164] R L Warren, G Choe, D J Freeman, and M Castellarin. Derivation of HLA types from shotgun sequence datasets. ..., 2012.

[165] R L Warren, J 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 research*, 21(5):790–797, May 2011.

[166] René L Warren and Robert A Holt. Targeted Assembly of Short Sequence Reads. *PLoS ONE*, 6(5):e19816, May 2011.

[167] Joshua A Weinstein, Ning Jiang, Richard A White, Daniel S Fisher, and Stephen R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science (New York, N.Y.)*, 324(5928):807–810, May 2009.

[168] N Wentzensen. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Research*, 64(11):3878–3884, June 2004.

[169] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing-zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard A Gibbs, and Jonathan M Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature ...*, 452(7189):872–876, April 2008.

[170] Brian T Wilhelm and Josette-Renée Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, July 2009.

[171] Elizabeth A Williamson, Leah Damiani, Andrei Leitao, Chelin Hu, Helen Hathaway, Tudor Oprea, Larry Sklar, Montaser Shaheen, Julie Bauman, Wei Wang, Jac A Nickoloff, Suk-Hee Lee, and Robert Hromas. Targeting the transposase

domain of the DNA repair component Metnase to enhance chemotherapy. *Cancer Research*, 72(23):6200–6208, December 2012.

[172] World Health Organization. Prevalence and incidence of selected sexually transmitted infections, Chlamydia trachomatis, Neisseria gonorrhoeae, syphilis and Trichomonas vaginalis: .... , *Published online: 07 March 1985; — doi:10.1038/314111a0*, 2011.

[173] Michal Wozniak, Jerzy Tiuryn, and Limsoon Wong. An approach to identifying drug resistance associated mutations in bacterial strains. *BMC genomics*, 13(Suppl 7):S23, December 2012.

[174] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, and Stephen R Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods*, 11(1):41–46, January 2014.

[175] Bo Xu, Sasithorn Chotewutmontri, Stephan Wolf, Ursula Klos, Martina Schmitz, Matthias Dürst, and Elisabeth Schwarz. Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS ONE*, 8(6):e66693, June 2013.

[176] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, Jin Huang, Ming Li, Xinglong Wu, Lu Wen, Kaiqin Lao, Ruiqiang Li, Jie Qiao, and Fuchou Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, September 2013.

[177] Itai Yanai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, Shirley Horn-Saban, Marilyn Safran, Eytan Domany, Doron Lancet, and Orit Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics (Oxford, England)*, 21(5):650–659, March 2005.

[178] M D Young, M J Wakefield, and G K Smyth. Method Gene ontology analysis for RNA-seq: accounting for selection bias. *PMC free article][...*, 2010.

[179] Jun Zhang, Rod Chiodini, Ahmed Badr, and Genfa Zhang. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*, 38(3):95–109, March 2011.

[180] X Zhao and K Drlica. Restricting the selection of antibiotic-resistant mutants: a general strategy derived from fluoroquinolone studies. *Clinical Infectious Diseases*, 33 Suppl 3:S147–56, September 2001.

[181] Jinsong Zhou, Dongyun Hao, Xudong Wang, Teimei Liu, Chengyan He, Feng Xie, Yanhong Sun, and Jin Zhang. An important role of a "probable ATP-binding component of ABC transporter" during the process of Pseudomonas aeruginosa resistance to fluoroquinolone. *Proteomics*, 6(8):2495–2503, April 2006.