UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ADDRESSING THE OMICS DATA EXPLOSION: A COMPREHENSIVE REFERENCE GENOME REPRESENTATION AND THE DEMOCRATIZATION OF COMPARATIVE GENOMICS AND IMMUNOGENOMICS

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Ngan K. Nguyen

June 2014

The Dissertation of Ngan K. Nguyen
is approved:

Professor David Haussler, Chair

Professor Joshua Stuart

Dietlind L. Gerloff, PhD

Professor Martha Zuniga

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Addressing the Omics Data Explosion: a Comprehensive Reference Genome Representation and the Democratization of Comparative Genomics and Immunogenomics

by

Ngan K. Nguyen

Advancements in technologies have resulted in an explosion of data, the volume of which continues to increase at an exponential rate. The accumulating wealth of data is enabling numerous new research possibilities and is transforming the world profoundly. In genomics, new genomes are being regularly sequenced, with a growing number of individual genomes becoming available for many species. As the ability to have complete genomic information becomes the norm, the need for a reference genome that better represents the particular species population intensifies: It becomes important to utilize the newly emerged sequences to improve current references and ensure better quality for future assemblies and experiments. Additionally, the proliferation of data has necessitated the decentralization of computational resources together with the empowerment of users to a do-it-yourself system, in which users create their own assemblies, alignments, visualizations and analyses. This is because with the accelerating amount of data, it is impossible and undesirable to maintain the infrastructure model in which only a number of specialized institutions handle most if not all of the data and analyses.

Joining many other on-going efforts, the works in this dissertation attempt to address some of these rising demands. First, I describe the problem of constructing a pan-genome reference for a population and demonstrate that the resulting pan-genome reference is more representative of the population than is any individual genome, using both simulated and real data. Second, I describe a comparative genomic framework that allows for easy generation of collections of web accessible UCSC genome browsers interrelated by an alignment. The pipeline, named the comparative assembly hub (CAH) pipeline, is intended to democratize UCSC comparative genomic resources and facilitate public sharing via the internet. As a demonstration, I create comparative assembly hubs for 66 *Escherichia coli/Shigella* genomes and highlight comparative analyses on their pan-genomic, core genomic and phylogenetic relationships. Last, I report on comprehensive assessments of the T cell receptor (TCR) repertoires of the autoimmune disease Ankylosing Spondylitis and show example comparative analyses for finding evidence of antigen selection and identifying potential disease-associated clones. In addition, I describe an open-source software package for profiling and comparing TCR sequencing data, called the "Adaptive IMmunoSequencing ToolKit", or the *aimseqtk* package. The *aimseqtk* package is comprised of four main components addressing common analyses of this type of data: clone tracking, repertoire profiling, public clone identification and publication mining.

To my parents,

Nguyen Duc Hoan and Vo Thi Le,

for their unconditional, unlimited love,

and endless support of my education.

To my partner and best friend,

Alexander Atkins,

for putting up with me and believing in me.

And to all my sisters,

for always being there.

Acknowledgments

This dissertation is the result of many years of work with the help and support of many wonderful individuals:

I thank my adviser David Haussler for his invaluable insights and guidance, and for providing me the opportunity to pursue my research. David's rigorous scientific curiosity and enthusiasm as well as the depth and breath of his knowledge are unparalleled and have always been an absolute inspiration.

I am in debt to Benedict Paten, who has patiently mentored me the last several years. Without his direction and encouragement, this dissertation would have been more difficult if not impossible to accomplish. Benedict's intellect is admirable and even more so is his detailed and careful scientific conduct. He has been a brilliant teacher.

I thank my committee members, Martha Zuniga, Dietlind Gerloff, and Joshua Stuart for their time, advice, and feedback.

I thank my coauthors and collaborators: Brian Raney, Glenn Hickey, Martha Zuniga, Maximilian Haeussler, Hyunsung John Kim, Daniel Zerbino, Mark Diekhans, Joel Armstrong, Dent Earl, Nader Pourmand, Brent Culver, Hiram Clawson, Ann Zweig, Donna Karolchik and Jim Kent for all their contributions to my work in this dissertation and beyond. In addition, I thank Max for his mentorship in the immunosequencing project and his help in many other aspects. Max's observant and inquisitive nature has been inspirational. I thank John for all the great brainstorming and discussions that we had. I thank Brian for taking my many browser-related questions, and

Mark for his help and advice the many times I got stuck. I thank Dent for his help with a number of figures included in this dissertation.

I thank Tracy Ballinger, Elinor Valesquez, Dent Earl, John Kim and all other fellow BME graduate students, past and present postdocs in the Haussler dry and wet labs, as well as in the department for their support and friendship. I thank Alex, Ben, John and especially Elinor for proofreading the many pages in this dissertation.

I thank everyone in the UCSC Genome Browser group and the CBSE team, including the browser staff, the system admins, and the office staff, especially Erich Weiler, Lynn Brazil and Daniel King, for a wonderful and supporting working environment.

Finally, I thank my family for their constant love and support. I am especially grateful to my sisters who have been taking great care of my parents the last four difficult years when my farther was ill and I was not able to be by their side.

Chapter 1

Introduction

We are undoubtedly witnessing the dawn of a new era, an era that overwhelms us with information. New data is being produced rapidly every day in a widespread number of fields: from consumption habits in marketing to fluctuations in stock markets, from trends and individual expressions in social media to astronomical data in the physical sciences, from climate observations in the earth sciences to genomic data in the life sciences. This immense amount of data has profoundly transformed our world, touching even the most basic aspects of our lives, as exemplified by targeted advertisements in media, customized results in internet searches and changes in social interactions and communications. As we move forward into this era, adaptations are required, demanding our attention to rethink and restructure the traditional ways of how everything is operated.

1.1 Reexamine the Reference Genome

In genomics, new genomes are now regularly sequenced, accelerating the availability of novel sequenced species and individual genomes, from both extensive public sequencing projects [10k-Community-of Scientists, 2009, 1000-Genomes-Project-Consortium, 2010] and individual efforts. As individual genomes become increasingly accessible, it is not only critical, but now most practical, to reexamine and to reestablish an essential component of the field: the representation of a species' reference genome.

The reference genome is typically a high quality individual genome that is used to represent the species of interest, providing a coordinate system, e.g an origin and a basis system, for genomes in the population or genomes of closely related subspecies. Previously sequencing costs were expensive, therefore the reference genome was assumed to be well represented by (mostly) a single genome. This representation, especially when the individual was selected partly by serendipity, is rather self-limiting because a single genome cannot best describe all the variations of an entire population.

It is proposed that a truly comprehensive universal coordinate system for the population variation must index a graph of aligned, common haplotypes [Li et al., 2010, Paten et al., 2014]. Given that all existing software (e.g. mapping, assembling and variation calling tools) heavily depends on a well established linear representation of the reference genome, I propose an intermediate solution for a linear reference genome that represents "the median point within the population diversity", inspired by the pan-

genome concept used in bacterial genomics. Due to horizontal gene transfer, within the same bacterial species, genetic content may differ dramatically from one strain to another. It was therefore apparent that describing a bacterial species by an individual strain was insufficient and consequently the pan-genome concept was introduced [Medini et al., 2005]. As traditionally defined in bacterial research, a pan-genome of a species is the union of genes of all strains within that species. This definition has since expanded to be the union of all homologous bases (alignment columns) of all individual genomes. In the first part of this dissertation, I explore the opportunities that arise from extending our standard definition of what is a reference genome to my novel pan-genome reference.

1.2 The Decentralization of Comparative Genomic Resources

Because of the increasing abundance of data, it has become impractical for public data centers to curate and/or manage all the data collections. Instead, a wealth of powerful tools is emerging to empower users to self-create visualizations and analyses. Contributing to this shift from a centralized system of computational resources to a user-oriented model are two software packages, namely, the CAH pipeline and the *aimseqtk* package. I created these packages to enable users to generate their own comparative genomic browser visualizations and perform their own comparative immunogenomic analyses.

1.2.1 The Comparative Assembly Hub (CAH) Pipeline

Visualization plays a critical role in research, not only by assisting us in analyzing, understanding and interpreting our data but also by providing us with visual cues for novel data interpretations as well as facilitating hypothesis formation. For example, the genome browser is one of the most powerful scientific visualization tools currently in existence. It incorporates many different annotations, spans different levels of resolution from whole chromosome to an individual base pair, in addition to being easily customizable. However, today's genome browsers are typically a single genome display, equipped with limited comparative capabilities. The CAH pipeline extends the single-genome scope of the UCSC browsers to display comparative genomics data, with multiple novel features incorporated to handle this type of data. The resulting representation takes advantage of the powerful features of the UCSC genome browsers, while at the same time, displays multiple alignments, different types of variations including structural rearrangements and duplications, and more importantly, provides consistent views when switching from one genome to another. Given a set of input genomes and available annotations, the pipeline generates a multiple sequence alignment, infers any pan-genome or ancestral genomes where appropriate, maps each input annotation from the original genome to other genomes of interest, and produces all necessary files to create one browser for each genome, all interconnected by the alignment. When the process is done, one of the output files is called the "hub.txt" file. Users can paste the location of this file to the UCSC Browser and the comparative assembly hub containing

the generated browsers with annotations is ready to use.

1.2.2 The Adaptive Immunosequencing Toolkit (*aimseqtk*) Pipeline

Immunogenomics is an emerging field that specializes in sequencing and analyzing genomic data of the immune system, including T cell receptor (TCR) repertoires. With the immune system directly related to health and diseases, comparative analyses in immunogenomics data have a large number of applications, benefiting both basic research and clinical needs [Robins, 2013]. Consequently, there is already an abundant amount of this new type of data. However, similar to every other field, data analysis methods have yet to catch up with the speed and amount of the data that is being generated.

There is no publicly available software for comparative immunogenomics analyses. Research groups either have to write their own code or outsource the analysis to servicing companies. To facilitate research, I present the *aimseqtk* package, which is an open-source software solution for comprehensively profiling and comparing TCR repertoires.

Looking Forward

In the following chapters, I go into details on the pan-genome reference, the CAH pipeline and the *aimseqtk* pipeline, one chapter per topic. Each chapter is organized into these sections: overview, introduction, results, discussion and methods. The final chapter is an overall summary of the dissertation.

Chapter 2

Building a Pan-genome Reference for a Population[1]

[1]This chapter is derived from two manuscripts that Benedict Paten and I created together [Nguyen et al., 2014b, Nguyen et al., 2014c].

2.1 Overview

A reference genome is a high quality individual genome that is used as a coordinate system for the genomes of a population, or genomes of closely related subspecies. Given a set of genomes partitioned by homology into alignment blocks, formalized in this chapter is the problem of ordering and orienting the blocks such that the resulting ordering maximally agrees with the underlying genomes' ordering and orientation, to create a pan-genome reference ordering. We show that this problem is NP-hard, but also demonstrate, empirically and within simulations, the performance of heuristic algorithms based upon a cactus graph decomposition to find locally maximal solutions. We describe an extension of the Cactus software to create a pan-genome reference for whole genome alignments, and I apply it to construct a pan-genome reference for the human major histocompatability complex (MHC). I demonstrate that the constructed MHC pan-genome reference represents the population variants more comprehensively than individual reference genomes.

2.2 Introduction

2.2.1 Reference Genome, Utilities and Limitations

A reference genome is a genome assembly used to represent a species. Reference genomes are indispensable to contemporary research for several reasons. They provide a coordinate system for consistent descriptions of functional annotations, such as genes and regulatory elements [Coffey et al., 2011, ENCODE-Project-Consortium et al., 2011], and variation data, such as single nucleotide variants (SNVs) and structural variants [Sherry et al., 2001, 1000-Genomes-Project-Consortium, 2010]. Such reference coordinates form the basis of the genome browsers [Fujita et al., 2011, Flicek et al., 2011] upon which these annotations are shown. Reference genomes are used for mapping relatively short sequences, such as sequencing reads, to their appropriate place within the genome [Li and Durbin, 2009, Trapnell et al., 2009]. Such reference-based sequence mapping forms the basis of an overwhelming number of contemporary functional assays [Wang et al., 2009, Park, 2009, ENCODE-Project-Consortium et al., 2011]. Inversely to mapping, reference genomes are used in the design of primer sequences, which are used to target a specific subsequence of a genome, e.g. in a polymer chain reaction [Bartlett and Stirling, 2003]. Reference genomes are used widely in comparative genomics, particularly in the construction of genomic sequence alignments [Miller et al., 2007, Paten et al., 2008]. Finally, once a reference genome for a species is available it can be used to assist in the assembly of related genomes, e.g. by ordering

scaffolds [Wheeler et al., 2008, Chimpanzee-Sequencing-Analysis-Consortium, 2005].

Reference genomes are now available for most important model organisms, including prokaryotes [Blattner et al., 1997], single cell eukaryotes [Goffeau et al., 1996], invertebrates [Adams et al., 2000, Consortium et al., 1998], plants [Aradidopsis-Genome-Initiative, 2000] and vertebrates, including fish [Aparicio et al., 2002], birds [International-Chicken-Genome-Sequencing-Consortium, 2004], non-avian reptiles [Alföldi et al., 2011], amphibians [Hellsten et al., 2010] and mammals [Mouse-Genome-Sequencing-Consortium et al., 2002, Lindblad-Toh et al., 2005]. At the time of writing, the UCSC and the Ensembl genome portals each have more than 50 different eukaryotic reference genome browsers, each based upon a reference assembly. As the cost of sequencing further declines these numbers will grow exponentially, with projects such as Genome 10k [10k-Community-of Scientists, 2009] projecting the sequencing of ten thousand distinct vertebrate species in the coming few years.

The utility of a reference genome for a species is potentially limited by the degree to which it represents the population's genomes. A failure to represent other intra-species genomes can be caused by variation in the following ways. First, sufficiently dense SNVs may make it difficult to discern homology between a reference genome and other genomes, particularly when mapping relatively short subsequences. Second, insertions and deletions may cause a reference genome to exclude subsequences that are common in other genomes. Third, nonlinear rearrangements, such as inversions, may

create breakpoints with respect to a reference genome. Finally, duplications may result in some genomes having a different copy number of certain subsequences in comparison with the reference.

Given these potential failures of representation, for a given species it is natural to ask how well the existing reference genome represents the species. In addition, whether it is possible to impute a better consensus reference genome from the multiple individual genomes. In this chapter, I explore creating a consensus, pan-genome reference from a collection of genomes and attempt to answer these two questions empirically for a limited portion of the human genome: the major histocompatability complex (MHC). The notion of a human pan-genome has been briefly visited by Li *et al.* [Li et al., 2010] in 2010. In the study, the authors investigated the "novel" sequences, i.e. sequences that were not present in the current human reference, in two de novo human assemblies, one Asian genome and one African genome. They identified approximately 5 Mb of such novel sequences and highlighted the needs for more de novo assemblies of human genomes to obtain a comprehensive understanding of the human pan-genome. Here, I formalize the problem, which is constructing a pan-genome reference for a collection of genome assemblies. Additionally, in constructing a novel reference I impute a complementary, integrated map of the variation within the MHC, in which every variant is described with respect to the reference and is given context by alignment within one large multiple sequence alignment (MSA).

2.2.2 The Human Major Histocompatability Complex Reference

The human genome [International-Human-Genome-Sequencing-Consortium, 2004] is perhaps the best sequenced vertebrate reference genome at this point, with essentially one large monoploid scaffold representation of each chromosome and only 20 unlocalized scaffolds as of GRCh37.p4 [Church et al., 2011]. Due to recent population bottlenecks [Hey and Harris, 1999, Li and Durbin, 2011], humans have a relatively low degree of polymorphism with respect to one another [Marth et al., 2004, Traherne, 2008]. The human reference genome can therefore be considered a reasonable near best case scenario for the reliance on a single reference genome in mammals.

I choose the MHC primarily because it contains interesting patterns of mammalian evolution [Belov et al., 2006] and substantial variation within humans [Traherne, 2008], including regions with extreme haplotype divergence [Raymond et al., 2005]. Additionally, the MHC is important in human immunology and disease [Fernando et al., 2008, Traherne, 2008], as highlighted by recent genome wide association studies (GWAS) [Wellcome-Trust-Case-Control-Consortium, 2007, Fellay et al., 2007].

The reference sequence as of GRCh37.p4 for the MHC is a single haplotype, named PGF. It is the largest and most complete haplotype segment available in the human reference genome [Horton et al., 2008]. It also represents one of the most frequent haplogroups within European populations, a haplogroup being a collection of similar haplotypes. This is in contrast to most of the remaining human reference, which

is a chimera of a number of randomly selected genomic samples [Lander et al., 2001], but deriving mostly from one in particular [Osoegawa et al., 2001]. The MHC has been designated in the GRCh37 reference genome a "polymorphic region". As such, seven additional alternative, homologous haplotypes, also representing common European haplogroups, have been added as supplementary sequences to GRCh37. These haplotypes are not included within the actual reference sequence, but are included as valid alternatives. In spite of being high quality assemblies, the alternative haplotypes are often excluded by most mapping softwares and processing pipelines (an example is the 1000 Genomes Project or 1KGP). For the MHC region, "mapping to GRCh37" is typically equivalent to "mapping to PGF". To avoid confusion and to make it easily identifiable as the de facto current human reference sequence for the MHC, henceforth I will refer to the PGF sequence as GRCh37. I will refer to the seven alternative haplotypes by their specific names and refer to the group of the PGF and the seven haplotypes as the "GRCh37 haplotypes".

2.2.3 Background of the Pan-genome Reference Problem

Starting from a set of genomes in a genome alignment [Paten et al., 2011b], which partitions the genomes' subsequences into homology sets termed blocks, the problem is to find an ordering of the blocks that as closely as possible reflects the ordering of the underlying genome sequences. Such an ordering is called a pan-genome reference, in that it indexes every block, something that any individual genome within the population almost certainly does not.

Closely analogous to the problem of building a pan-genome reference of aligned input genomes, a great deal of previous work has focused on methods for ancestral reconstruction. Most relevant to this work is the (rearrangement) median problem. Informally, the median problem is, given a set of genomes and an edit operation, to find a median genome whose total pairwise edit distance from each of the other sequences is minimal [Tannier et al., 2009]. Naively, it might be assumed that good solutions to the median problem might have utility for finding an intra-species pan-genome reference. However, in the median problem the edit operations are not necessarily restricted to maintain sequence colinearity, while during evolution complex selective pressures often work to achieve exactly this [Kirkpatrick, 2010]. For example, consider the three signed permutations: (A, d, B, e, C), $(A, -e, B, -d, C)$ and $(A, B, e, -d, C)$. Assume that the capital letters, A, B and C represent very large subsequences of the genome and the lower case letters, d and e, represent short subsequences. In each of the sequences the large subsequences maintain their colinearity with respect to one another. When ignoring the short subsequences, no edits appear to have occurred. However, when incorporating the short subsequences, the optimal median sequence under either the double-cut-and-join (DCJ) or reversal edit operations is $(A, -e, -B, -d, C)$; the other sequences are each one operation away. This optimal median sequence contains an inversion of the large sequence B, which may make it biologically implausible to be a common ancestor, e.g. if there is a single gene with exons spanning A, B and C. This tendency to lose colinearity has led to the study of 'perfect' rearrangement scenarios, in which common intervals of ordered subsequences present in the input are conserved

[Berard et al., 2009].

However, current algorithms for finding perfect rearrangement scenarios require the common intervals to be pre-specified, do not allow copy number variation and require the common intervals to exist in all the inputs. This makes them inappropriate when there is no prior expert knowledge to define the intervals, or when representing large populations, where copy number variation is present and missing data and unusual variants break many intervals that would otherwise be common.

Methods to derive consensus orderings of sets of total and partial orders have been extensively considered, particularly in the domain of social choice [Fagin et al., 2002, Kendall, 1938]. In general, the inputs to such problems are sequences or structures equivalent (in their most general form) to directed acyclic graphs (DAG), and the output is a consensus (partial) ordering. In such work, algorithms often work to minimize the consensus' (weighted) symmetric difference distance or Kemeny tau distance [Kendall, 1938] (informally, the number of out of order (discordant) pairs). Recently, such consensus ordering procedures have been adapted to create consensus genetic maps from sets of individual subpopulation maps [Bertrand et al., 2009]. The problem formalized here has similarity to such approaches, with the important difference be that it explicitly models the double stranded nature of DNA, allowing us to account for the cost of sequences being inverted with respect to one another.

What follow are the formalization of the basic problem, proof of its NP-hardness, description of a principled heuristic decomposition of the problem using cactus graphs [Paten et al., 2011a], heuristic algorithms for the problem's solution, demonstra-

tion of the algorithms performance using simulation and lastly, description of the MHC pan-genome reference and its utilities in comparison with the human reference GRCh37.

2.3 Results

2.3.1 The Pan-genome Reference Problem

2.3.1.1 Genome Sequences

Let $S = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ be the input DNA sequences, with lengths $(n_1, n_2, \ldots, n_k)$. For simplicity it is assumed here that the DNA sequences are linear, though extensions to allow additional circular sequences are straightforward. Due to the double stranded nature of DNA, the $5'$ and $3'$ ends of each sequence element are distinguished. A tuple $(x \in \{1, 2, \ldots, k\}, i \in (1, 2, \ldots, n_x), a \in \{5', 3'\})$ is denoted as x_i^a, giving the coordinate of the a end of the ith element in σ_x. For any DNA sequence σ_x the ends are oriented consistently, so that for all $i > 1$ the $x_i^{5'}$ end is adjacent (contiguous) in the sequence to the $x_{i-1}^{3'}$ end and, for all $i < n_x$ the $x_i^{3'}$ end is adjacent in the sequence to the $x_{i+1}^{5'}$ end. Signed notation is used to distinguish ends, hence $-x_i^{5'} = x_i^{3'}$ and $x_i^{5'} = -x_i^{3'}$. The set of all end coordinates is $\mathbf{S}$.

2.3.1.2 Alignment

The end coordinates in $\mathbf{S}$ are partitioned by their alignment relationships. The alignment relation is defined as $\sim \subset \mathbf{S}^2$. The alignment relation is an equivalence relation, i.e. one that is transitive, symmetric and reflexive. The equivalence classes

for $\sim$ are denoted as $\mathbf{S}/\sim$, and $[x_i^a]$ represents an equivalence class containing x_i^a. The alignment relation is constrained to force the pairing of opposite ends. First, it is assumed that if $x_i^a \sim y_j^b$ then $-x_i^a \sim -y_j^b$, termed *strand consistency*. Second, it is assumed that if $x_i^a \sim y_j^b$ then neither $-x_i^a \sim y_j^b$ or $x_i^a \sim -y_j^b$, termed *strand exclusivity*. Due to *strand consistency*, for all $[x_i^a]$ in $\mathbf{S}/\sim$ there exists $[-x_i^a] = \{-y_j^b : y_j^b \in [x_i^a]\}$, the reverse complement of $[x_i^a]$. Due to *strand exclusivity*, for all x_i^a, $[x_i^a] \neq [-x_i^a]$. Combining these two statements it follows that $|\mathbf{S}/\sim|$ is even. The set $[-x_i^a]$ can be equivalently denoted $-[x_i^a]$, so that the reverse complement of X in $\mathbf{S}/\sim$ is $-X$. Each member of $\mathbf{S}/\sim$ is a *side*, and each pair set of forward and reverse complement sides is a *block*. Note that the alignment relation allows for copy number variation, i.e. arbitrary numbers of coordinates from sequences in the same genome can be present in a block.

2.3.1.3 Sequence Graphs

Let $G = (V, E)$ be a *(bidirected) sequence graph*. A bidirected graph is a graph in which each edge is given an independent orientation for each of its endpoints [Medvedev and Brudno, 2009]. The vertices are the set of blocks, $V = \{\{X, -X\} : X \in \mathbf{S}/\sim\}$. The edges, $E = \{\{[x_i^{3'}], [x_{i+1}^{5'}]\} : \sigma_x \in S \wedge i \in (1, 2, \ldots, n_x - 1)\}$, encode the adjacencies (biologically the covalent bonds) between contiguous ends of sequence elements. Each edge is a pair set of sides rather than a pair set of vertices, therefore giving each endpoint its orientation, see Fig. 2.1(A). The cardinality and size of G are clearly at most linear in the size of $\mathbf{S}$.

A sequence of sides $(X_1, X_2, \ldots, X_n)$ is a *thread*. If the elements in

$\{-X_1, X_2\}, \{-X_2, X_3\}, \ldots, \{-X_{n-1}, X_n\}$ are edges in the graph then the thread is a *thread path*. A sequence of sides rather than vertices is used because the sides orient the vertices, distinguishing forward and reverse complement orientations. For example, for each sequence $\sigma^x \in S$, $[x_1^{5'}], [x_2^{5'}], \ldots, [x_{n_x}^{5'}]$ is a thread path in G, because for all $i \in 1, 2, \ldots, n_{x-1}$, $\{[x_i^{3'}], [x_{i+1}^{5'}]\}$ (equivalently $\{-[x_i^{5'}], [x_{i+1}^{5'}]\}$) is an edge in G.

A *transitive sequence graph*, $\hat{G} = (V, \hat{E} = \{\{[x_i^{3'}], [x_j^{5'}]\} : \sigma_x \in S \wedge i < j\})$, includes the sequence graph G as a subgraph but additionally includes edges defined by *transitive adjacencies*, that is pairs of ends connected by a thread path. The cardinality (vertex number) of $\hat{G}$ is the same as G, but the size (edge number) of $\hat{G}$ is worst-case quadratic in the size of $\mathbf{S}$. A sequence graph encodes input sequences and an alignment, a transitive sequence graph models the complete set of ordering and orientation relationships between the blocks implied by the input sequences (Figure 2.2).

2.3.1.4 Pan-genome References

A *pan-genome reference* F is a set of non-empty threads such that each block is visited exactly once, see Fig. 2.1(B). Intuitively, not all pan-genome references are equally reasonable as a way of summarizing S, because they will not all be equally "consistent" with the set of adjacencies, $\hat{E}$. An edge $\{X, Y\}$ is *consistent* with a pan-genome reference F if and only if there exists a thread in F containing the subsequence $-X, \ldots, Y$, see Fig. 2.1(B). Given a weight function $z : \hat{E} \to \mathbb{R}_+$, which maps edges to positive real valued weights, the *pan-genome reference problem* is to find a pan-genome reference in $\mathbf{F} = \arg\max_F \sum_{e \in \hat{E}_F} z(e)$, where $\hat{E}_F$ is the subset of $\hat{E}$ consistent with F.

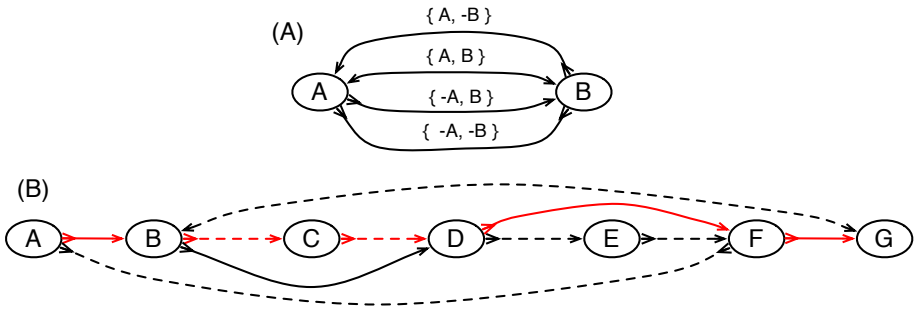

Figure 2.1: (A) A bidirected graph representing the four ways two blocks can be connected. The arrowheads on the edges indicate their endpoints: the sides of the vertices. (B) An example pan-genome reference on a sequence graph. There are two sequences, indicated by the color of the edges. The red sequence, represented by the thread A, B, C, D, F, G and the black sequence, represented by the thread $A, -F, -E, -D, -B, G$. The red thread visits the edges $\{-A, B\}$, $\{-B, C\}$, $\{-C, D\}$, $\{-D, F\}$ and $\{-F, G\}$ and the black thread visits the edges $\{-A, -F\}$, $\{F, -E\}$, $\{E, -D\}$, $\{D, -B\}$ and $\{B, G\}$. Neither thread includes all the blocks. A pan-genome reference, indicated by the dotted edges, is $A, -F, -E, -D, -C, -B, G$. The dotted edges and the edges $\{-B, D\}$ and $\{-D, F\}$ are the edges consistent with the given pan-genome reference.

2.3.1.5 Exponential Weight Function

Although many possible weight functions exist, inspired by the nature of genetic linkage, z is defined as $z(\{X, Y\}) = z'(X, Y) + z'(Y, X)$, where $z'(X, Y) = \sum_{\sigma_x \in S} \sum_{x_i^{3'} \in X} \sum_{x_j^{5'} \in Y} (1-\theta)^{j-i} I_{\{i<j\}}$, in which $I_{\{i<j\}}$ is the indicator function that is 1 for pairs of i and j for which $i < j$ else 0, and the parameter θ is a real number in the interval $[0, 1)$. The θ parameter intuitively represents the likelihood that an adjacency between two directly abutting sequence elements is broken or absent in any other randomly chosen sequence, and is defined analogously to its use in the LOD score [Griffiths et al., 1999] used in genetics. For $\theta > 0$, the score given to keeping elements in a sequence in the same order and orientation in the pan-genome reference declines

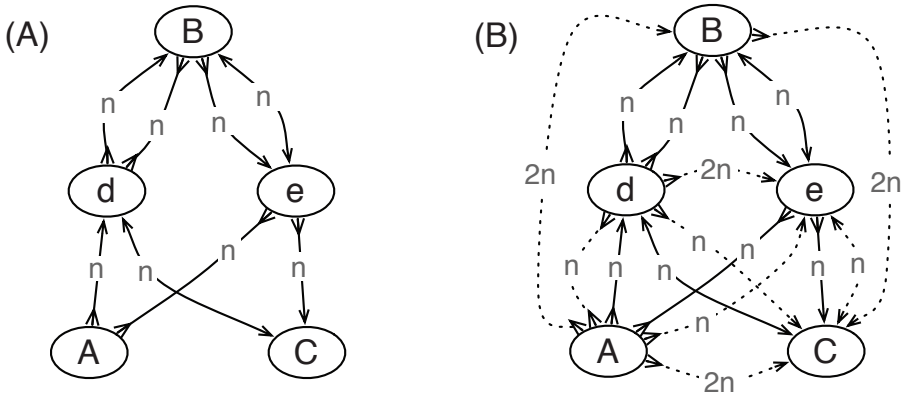

Figure 2.2: An illustration of a transitive sequence graph. (A) A sequence graph of n sequences of "A d B e C" and n sequences of "A -e B -d C". (B) A transitive sequence graph of the same sequences in (A).

exponentially with distance separating them.

To make it clear that an intermediate value of θ is desirable one can look at what happens at extreme values of the parameter. As θ approaches 1 the weight function become dependent only on edges in the sequence graph. Fig. 2.3 demonstrates a limitation with considering only these edges, which is similar to that described for edit operations in the introduction. At $\theta = 0$ all transitive adjacencies are equally weighted, however this can lead to longer sequences having undue influence on the solution; Fig. 2.3 also gives an example of this limitation when weighting all adjacencies equally. One issue not dealt with by the definition of z are the evolutionary interdependencies between the input sequences. It is possible to adjust the weights given to adjacencies given a phylogenetic tree that relates the input sequences (or the genomes they derive from). However, where homologous recombination is present a weighting based upon a phylogenetic tree is insufficient and yet more complex strategies are needed.

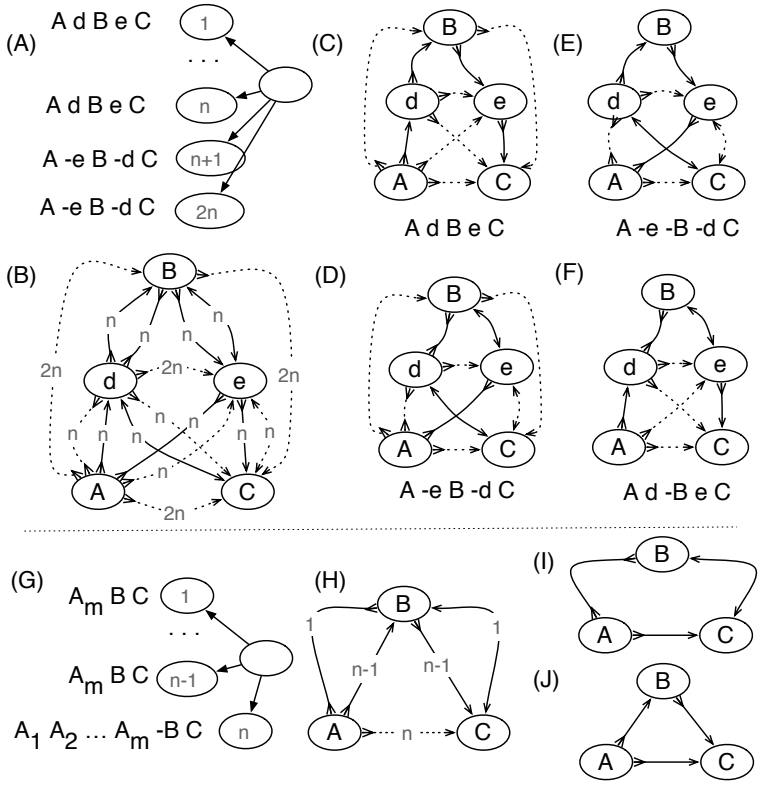

Figure 2.3: **Top**: An illustration of why it is not always sufficient to consider only abutting adjacencies. (A) There are five blocks, A, B, C, d and e, reprising their roles from the example given in the introduction. The input contains n copies of the sequence A, d, B, e, C and n copies of the sequence $A, -e, B, -d, C$. (B) The bidirected graph representation of this problem, with the number of adjacencies supporting each edge labeled, the abutting adjacencies shown as solid lines and the non-abutting adjacencies shown as dotted lines. If only solutions that start with A and end with C are of interest, there are 4 maximal solutions, shown in (C,D,E,F). Solutions (C) and (D) each have $4n$ abutting adjacencies and $10n$ non-abutting adjacencies. Solutions (E) and (F) each also have $4n$ abutting adjacencies but only $6n$ non-abutting ones. For θ <1 the (C) and (D) solutions are optimal. As θ approaches 1, the weight of non-abutting adjacencies approaches 0 and all four solutions become equally weighted, despite (E) and (F) having B in the reverse orientation. **Bottom**: An illustration of why θ should be greater than 0. (G) There are $m + 2$, blocks, the input contains $n - 1$ copies of the sequence A_m, B, C and 1 copy of the sequence $A_1, A_2, \ldots, A_m, -B, C$. (H) The bidirected graph representation of the problem, where the sequence of $A_1, A_2, \ldots, A_m$ blocks has been reduced to just a single vertex for convenience. The two maximal solutions are shown in (I,J), corresponding to the two distinct input sequences. If $m > n$ and θ is 0 then the solution with B in the reverse orientation (I) is optimal, despite this orientation being observed only once. By increasing θ the alternative solution with B in the forward orientation becomes optimal.

2.3.2 NP-hardness of the Pan-genome Reference Problem

The pan-genome reference problem is NP-hard and can be projected onto the problem of finding maximum weight subgraphs of a bidirected graph that do not contain characteristic classes of simple cycle. See Appendix Section A.1 for a full proof of the problem's NP-hardness.

2.3.3 Algorithms for the Pan-genome Reference Problem

Having established that the pan-genome reference problem is NP-hard, the following subsections describe a principled and novel heuristic to decompose the problem using cactus graphs, and briefly describe two straightforward algorithms to build and refine a pan-genome reference.

2.3.3.1 Cactus Decomposition of the Pan-genome Reference Problem

A cactus graph of the type introduced in [Paten et al., 2011a] describes a sequence graph in a hierarchical form. For a sequence graph G, a pair of sides X and Y form a *chain interval* if there exists one or more thread paths of the form $-X, \ldots, Y$, but no thread paths of the form $-X, \ldots, -Y$ or $X, \ldots, Y$. Chain intervals represent intervals that are "fundamental", in the sense that all the simple threads for all the sequences in S follow the traversal rules defined above. It is reasonable therefore to search for reference sequences that preserve all such intervals.

The chain interval relation defines a partition of the vertices into a set of disjoint *chains*. A *chain* is a thread $(X_1, X_2, \ldots, X_n)$ such that all and only pairs of

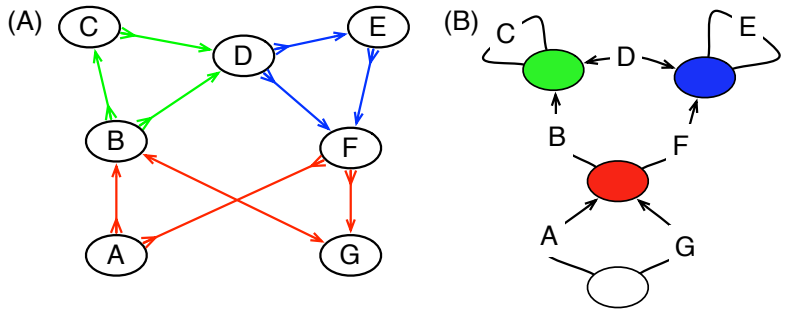

Figure 2.4: (A) The bidirected graph from Fig. 2.1(B) redrawn to show the nets as colored side subgraphs. (B) The cactus graph representation of the blocks and nets in (A), with the white net containing the highest level chains. The edges represent the blocks, the vertices the nets. The arrow heads on the edges indicate endpoints that are links.

form $(-X_i, X_j)$ for which $j-i \geq 1$ define a chain interval; each chain interval of the form $(-X_i, X_{i+1})$ is called a *link*. Chains can be arranged hierarchically, because one *child* chain may be contained within the link of a *parent* chain. Two chains are called *siblings* if either they are both children of the same parent chain link or both are not contained within any parent chain link (i.e. they are at the highest level of the hierarchy). For a thread $(X_1, X_2, \ldots, X_n)$ the two sides X_1 and $-X_n$ are *stubs*. A *net* is an induced subgraph of G defined by the set of stubs for a maximal set of sibling chains and (if they exist) the pair of sides that define the containing parent link, see Fig. 2.4(A). A graph in which the nodes are the nets and the edges are the oriented vertices of a sequence graph forms a cactus graph, see Fig. 2.4(B).

To construct a reference that respects all chain intervals, a pan-genome reference is created independently for each net, each pair of chain stubs treated as equivalent to blocks in the previous exposition. Additionally, the pan-genome reference for each

child net with a parent link must be composed of a single thread whose stubs are the sides of the parent link. This reduces the maximum size of the pan-genome reference problem to that of the largest net in the sequence graph, which as the sequence graph for alignments of variation data is often relatively sparse, has (in experience and in accordance to elementary random graph theory [Erdos and Rényi, 1960]) size only approximately logarithmically proportional to the number of vertices in the graph. It also facilitates parallel execution, because each net can be computed in parallel.

2.3.3.2 Greedy and Iterative Sampling Algorithms for the Pan-genome Reference Problem

Given the decomposition, a pan-genome reference for each subproblem is built using an initial greedy algorithm, before iterative refinement that employs simulated annealing.

In overview (see the source-code for more details), a pan-genome reference F is composed, starting from the empty set, by greedily adding one member of V to F at a time, each time picking the combination of insertion point and member of V that maximizes consistency with elements already in the F. The algorithm is naively $|V|^3$ time (as each insertion is $|V|^2$ time), though by heuristically ignoring weights less than a specified threshold (the weight declines exponentially with sequence separation), and using a priority queue to decide which member of V to add next, it can be improved $|V|log(|V|)$ in practice.

Given an initial reference F the procedure progressively searches through a

sequence of neighboring permutations, where for a reference F a *neighboring permutation* is created by removing an element from F and then inserting it either in the positive or negative orientation as a prefix, suffix or coordinate between elements in the reduced F, potentially including the elements original coordinate. The algorithm incorporates simulated annealing by using a monotonically decreasing temperature function to control the likelihood of choosing neighboring, lower scoring permutations. As the temperature tends to zero the algorithm becomes greedy and a local minima can be searched for, while as the temperature tends to positive infinity all permutations become equally probable and the search becomes a random walk. Each iteration of sampling, in which the repositioning of every block is considered once, is naively $|V|^2$ time, but is improved to $|V|log(|V|)$ in practice.

2.3.4 Simulation Experiments

To test the algorithms described we use a simple simulation of a rearrangement median problem. We start with a single linear chromosome, represented as a signed permutation of 250 elements, which we call the original median. We then simulate either 3, 5 or 10 leaves, treating each leaf with a set number of random edits. For convenience we simulate only translocations and inversions, which results in each leaf remaining a single contiguous chromosome, and apply an equal number of translocations and inversions. Note, for simplicity, we did not assess copy number changes (e.g. duplicative rearrangements), but doing so would be interesting.

We performed two sets of simulations, in the first we did not constrain the

length of the subsequence of elements inverted or translocated. In such a scenario only a few edits are sufficient to radically reorder the genome and break many resulting ordering relationships. In the second scenario we constrained the lengths of inverted subsequences to 2 or 1, and constrained the length of translocated subsequences to just 1. In this scenario relatively large numbers of rearrangements are required to breakup the ordering of the original median.

To find solutions to the pan-genome reference problem we use a combination of the algorithms described above, first using the greedy algorithm, then refining it with iterative sampling, performing 1000 iterations of improvement and setting $\theta = 0.1$ (values of theta between 0.5 and 0.001 made little difference to the result). We call this combination Ref. Alg. in the results that follow. To compare performance of our solutions we compare them to the original median, and to a median genome inferred using the AsMedian program [Xu, 2009] (using default parameters), which finds optimal solutions to the DCJ median problem with three leaves. We assess performance by looking at three metrics. First, the DCJ distance, which gives the minimum number of edits needed to translate one genome into another by DCJ edits. Second, viewing the medians as two signed, partial order relations A and B on the blocks, the symmetric difference distance, defined as $\frac{|A \triangle B|}{|(A \cup B)|}$. This gives the proportion of order plus orientation relationships not common to the two medians. Last, we compare a weighted form of the symmetric difference distance, in which each ordered pair present in the symmetric difference of the two order relations is weighted by $(1-\theta)^i$, where i is the length of the elements separation in the sequence the pair is present in. This distance (in common

with the weight function used for the pan-genome reference problem) therefore gives exponentially more weight to pairs of elements close together in one of the medians but not in the same order and orientation in the other median.

The top panel of Fig. 2.5 shows the results of simulating unconstrained, arbitrary translocations and inversions. Unsurprisingly, Ref. Alg. constructs medians that are substantially farther from the leaves or the original median in terms of DCJ distance than the results of AsMedian (avg. 44% and 103% more overall than Ref. Alg. with 3 leaves, respectively, from the leaves and original median). Furthermore, in terms of weighted and unweighted symmetric difference distance, the AsMedian solutions are closer to the original median (though not the leaves) than those constructed using Ref. Alg. for moderate numbers of simulated edits (avg. 34% and 52%, respectively, in terms of symmetric difference and weighted symmetric difference compared to Ref. Alg. with 3 leaves). This clearly demonstrates that using the Ref. Alg. for sequences whose ordering have been turned over by large rearrangements produces poor results, and that ancestral reconstruction algorithms can be used more effectively for moderate numbers of edits in this scenario, with the caveat that they may construct a multi-chromosomal ordering of the data.

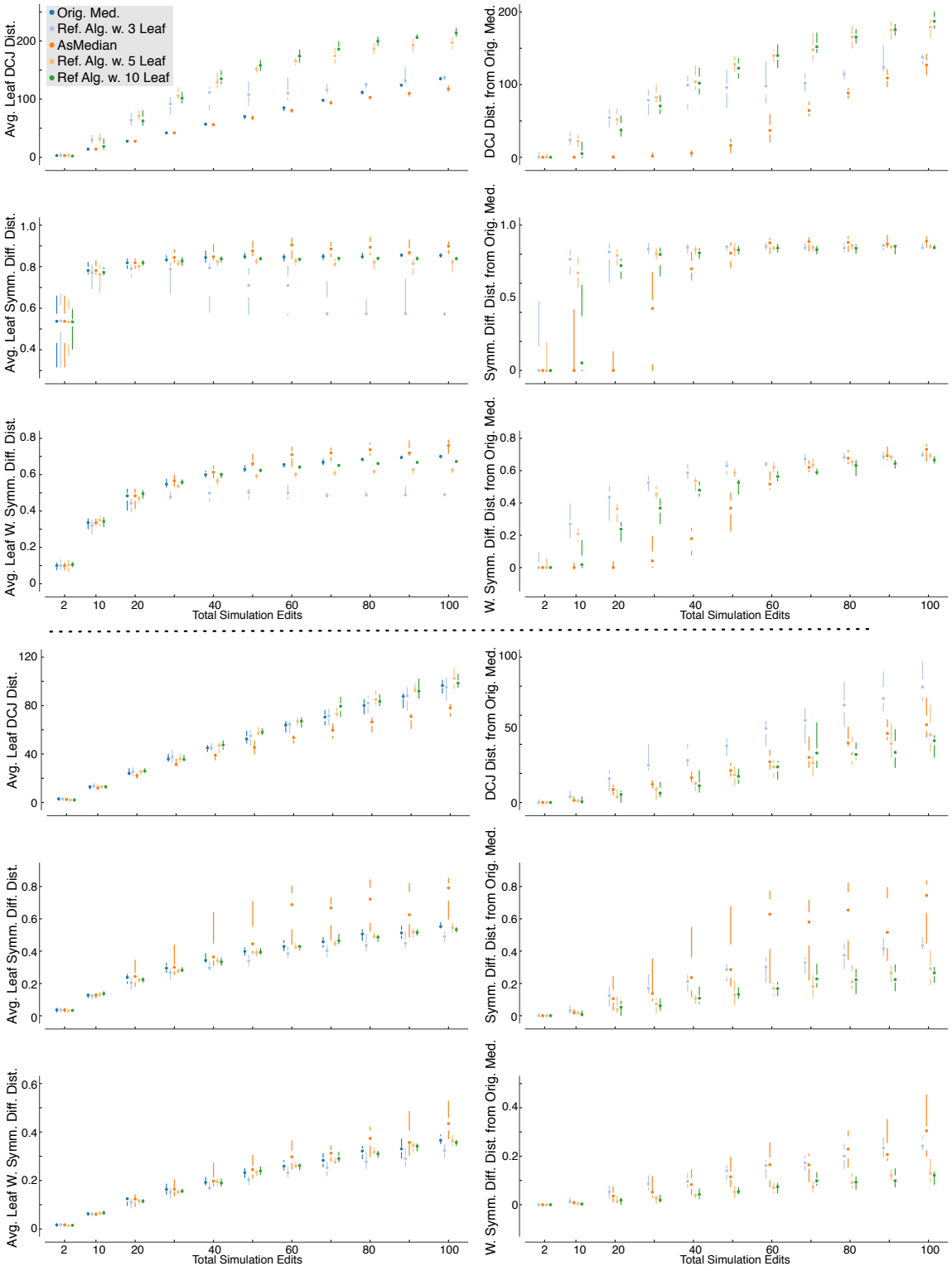
Orig. Med.
Ref. Alg. w. 3 Leaf
AsMedian
Ref. Alg. w. 5 Leaf
Ref Alg. w. 10 Leaf
Avg. Leaf DCJ Dist.
DCJ Dist. from Orig. Med.
Avg. Leaf Symm. Diff. Dist.
Symm. Diff. Dist. from Orig. Med.
Avg. Leaf W. Symm. Diff. Dist.
W. Symm. Diff. Dist. from Orig. Med.
Total Simulation Edits
Total Simulation Edits
Avg. Leaf DCJ Dist.
DCJ Dist. from Orig. Med.
Avg. Leaf Symm. Diff. Dist.
Symm. Diff. Dist. from Orig. Med.
Avg. Leaf W. Symm. Diff. Dist.
W. Symm. Diff. Dist. from Orig. Med.
Total Simulation Edits
Total Simulation Edits

Figure 2.5: **Top**: Simulation results using arbitrary inversion and translocation operations. Each plot shows the total number of operations (a mixture of 50% inversions and 50% translocations) vs. either the DCJ distance (top two plots), symmetric difference distance (middle two plots) or weighted symmetric difference distance (bottom two plots). The left plots give the average distance from the leaf genomes and the right plots give the distance from the original "true" median genome. Series shown include the original median genome (left plots only), the inferred median genome from the AsMedian program [Xu, 2009] using three leaves, and the inferred median genomes using our combined reference algorithms, using, separately, three, five and ten leaf genomes as input. Simulations used ten replicates for each fixed number of edits, points give median result, lines show max and min quartiles. **Bottom**: Simulation results using short inversion and translocation operations, laid out as in top panel.

The bottom panel of Fig. 2.5 shows the results of simulating short edits, demonstrating a striking converse to the unconstrained case. In terms of DCJ distance, the Ref. Alg. with 5 and 10 leaves is actually able to outperform the AsMedian program in terms of distance to the original median (Ref. Alg. with 5 leaves requires 20% on avg. fewer DCJ edits than AsMedian), while in terms of weighted and unweighted symmetric distance Ref. Alg. with 3 leaves is able to find solutions that are as close to the leaves as the original median and substantially closer to the original median than the AsMedian results (Ref. Alg. with 3 leaves is 31% closer on avg. than AsMedian in terms of symmetric difference distance to the original median). Furthermore, adding more leaves improves the results substantially (Ref. Alg. with 10 leaves is 52%, 44% and 55% closer to the original median in terms of avg. DCJ, symmetric difference and weighted symmetric difference distance than Ref. Alg. with 3 leaves). These results demonstrate that if edits have largely maintained the linear ordering of the sequences then, even when the sequences have been subject to substantial numbers of edits, Ref. Alg. is competitive with an ancestral reconstruction method in terms of DCJ, while

ensuring that all elements appear in an ordering that is closer, in terms of ordering and orientation, than an optimal ancestral reconstruction method.

2.3.5 Creating a Pan-genome Reference for the Human Major Histocompatibility Complex (MHC)

First, I describe in overview the construction of a pan-genome reference sequence for the MHC, which I refer to as C. Ref. (consensus reference) henceforth. I then detail a series of comprehensive evaluations designed to assess C. Ref., in particular in contrast to the existing reference genome for this region.

2.3.5.1 Samples and Assemblies

The input data consist of 16 human samples, for full details see Subsection 2.5.1. Eight of these (pgf, apd, cox, dbb, mann, mcf, qbl and ssto) were the GRCh37 haplotypes, all of which were previously assembled [Horton et al., 2008]. Five of the samples derived from two deeply sequenced trios (parents and child) from the 1000 Genomes Project (1KGP), one trio of African descent and one trio of European descent [1000-Genomes-Project-Consortium, 2010]. The paternal sample of the European trio, named NA12891, was excluded from the analysis for data quality reasons. Of these samples, I assembled four (NA12892, NA19238, NA19239, NA19240) using the Velvet assembly program [Zerbino and Birney, 2008] and extracted one (NA12878) from a recent human assembly made using the ALLPATHS-LG program [Gnerre et al., 2011]. Additionally, I used the recent Asian (Yh1) [Wang et al., 2008] and African (NA18507)

de novo genome assemblies [Bentley et al., 2008] and the Venter genome assembly [Levy et al., 2007]. Finally, as an outgroup I included the chimpanzee reference genome sequence (panTro3) [Chimpanzee-Sequencing-Analysis-Consortium, 2005], using it to break ties in decisions about constructing C. Ref.

Apart from the reference mapping step used to isolate scaffolds and reads that are specific to MHC, all the assemblies may otherwise be considered de novo assemblies, in that they were assembled without assistance of a reference genome. Although this is not a requirement of the methods, it helped to avoid reference allele bias, a tendency to assemble reference alleles in preference to alternatives, influencing the construction of C. Ref. Another important consideration of the genome assemblies used is that they are haploid. In the case of the GRCh37 haplotypes this is because the underlying samples are haploid in the MHC region. The remaining samples are diploid artificially made haploid by the assembly process. For this reason it is expected that around half the variants present in these samples will be missed.

2.3.5.2 Alignment and Pan-genome Reference Construction

Given the input assemblies, I construct an MSA using Cactus [Paten et al., 2011b], partitioning the individual bases (bp) in the input sequences into sets of homologous bases, called *homology sets*. A homology set that contains two or more bases is called a *column*, in analogy to the representation used in a traditional 2-dimensional MSA, where the rows are the sequences and the columns are the homology sets representing the alignment of individual bases. The bases within

a homology set are oriented, thus each homology set has a forward and a reverse complement orientation.

Each input sequence defines a sequence of oriented columns, with some potentially interstitial unaligned subsequences. The software constructs C. Ref. by picking a sequence of these oriented columns, such that each column is present exactly once in either its forward or reverse complement orientation (this sequence is therefore a signed permutation), and using the most frequently occurring base in each column to define the actual base representing the column in C. Ref.

There are two major challenges in the approach. The first challenge is finding the optimal ordering of the columns, the solution to which is presented in previous Section. The second challenge is finding an appropriate set of columns. This problem is approached by modifying the Cactus [Paten et al., 2011b] alignment program: First, to avoid aligning ancient paralogies, the software works to exclude all homologies that significantly predate the speciation of humans and the outgroup species chimpanzee. Second, in cases where copy number variation is present, to ensure that C. Ref contains all recurrent copies, the aligner first undoes relevant homologies in an initial alignment ("melting") and then reanneals the homologies using synteny partition, so that in the resulting alignment, each column contains at most one base from each sample, i.e. no self-alignments (Figures 2.6, 2.7). Last, as stated, the columns are required to include at least two bases. By the previous modification, these two bases must come from different samples. This prevents rare or erroneously assembled subsequences from being included in C. Ref. I call the bases within columns *recurrent* because they have

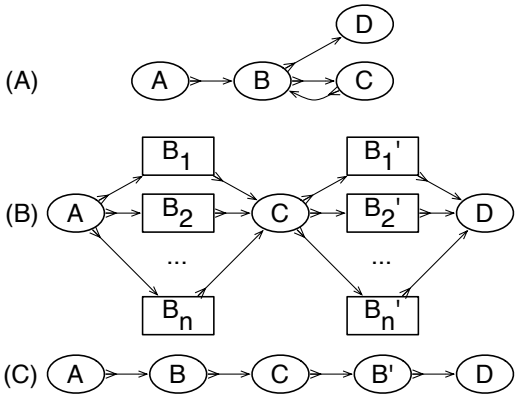

Figure 2.6: An example of a synteny partition, used to ensure that when copy number variation is present, the pan-genome reference contains all recurrent copies. There are n identical sequences $A; B; C; B'; D$, such that B and B' are homologous by a segmental duplication. (A) A bidirected graph for the example including only direct adjacencies. (B) After melting the B vertex in (A) ("melting" is undoing the alignment relationship, or homologies, in the vertex). The rectangles represent interstitial sequence that is unaligned (not within a column). There are two groups, one containing an end of A and an end of C, and the other containing an end of C and an end of D. (C) After reannealing the homologies, but only between positions in the same group. The graph now has two copies of B, separated into homologies groups by the segmental duplication.

homologs. Inversely, bases not contained within columns are called *non-recurrent*.

2.3.5.3 C. Ref. Sample Composition: C. Ref Contains ∼6% of Recurrent Bases That Are Absent in GRCh37

To understand the contribution of each sample to C. Ref. I analyze the alignment of each sample in the MSA. In this work each sample can be considered a set of contigs. A contig and the bases it contains are *covered* by the alignment if one or more bases in the contig is recurrent, and therefore included in C. Ref. Figure 2.8(A) (also Appendix Table A.1) shows the total length of covered contigs for each sample (grey color bars), the number of recurrent bases (red color bars, equivalent with the number

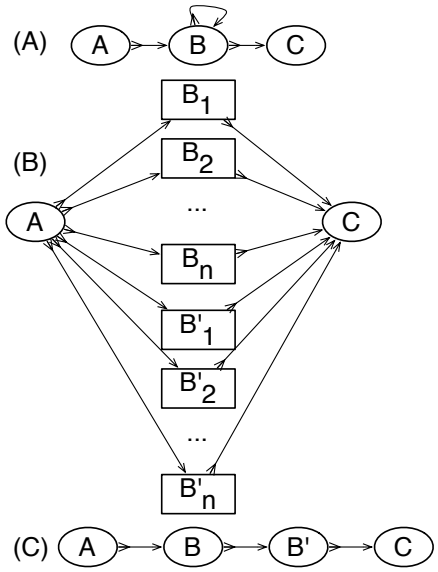

Figure 2.7: An example of the synteny partitioning of a tandem duplication. There are n identical sequences $A; B; B'; C$, such that B and B' are homologous by a tandem duplication. (A) A bidirected graph for the example including only direct adjacencies. (B) After melting the B vertex in (A). The rectangles represent interstitial sequence that is unaligned (not within a column). There is only one group, hence reannealing will return the graph to its state in (A). Instead, the BAR algorithm is used, which creates an alignment in which only substitutions and indels are allowed, hence B and B' end up in separate columns.

of bases aligning to C. Ref.) and the number of bases in each sample that align to bases in GRCh37 (blue color bars).

There are substantial differences in the numbers of covered bases between the samples. These differences are largely explained by the use of different sequencing technologies: the Venter and the haploid GRCh37 samples were generated using Sanger sequencing and have larger average numbers of covered bases (avg. 4,162,770 bp) than the remaining samples (avg. 3,168,268 bp), which all used Illumina short read (avg. 47bp in this study) sequencing technology [Bentley et al., 2008]. However, there are

two clear exceptions to this pattern. The Sanger sequenced APD sample has noticeably fewer such bases (2,320,747 total bp). Conversely, the Illumina sequenced NA12878 sample (4,192,579 total bp) has similar coverage with the Sanger sequenced samples, probably because it had higher sequencing coverage and a greater variety of paired end libraries than the other Illumina samples [Gnerre et al., 2011]. Related to this, Figure 2.8(B) (see also Appendix Figure *A*.2) shows the number of bases in selected samples aligned within homology sets of a given cardinality; the cardinality of a homology set is the number of different samples having bases aligned in that set. The relationships for each sample shown are complex and subtly different, but, comparing with Figure 2.8(A), all seem affected by the differing coverage of subsets of the samples.

For a sample, the difference between the number of covered bases and the number of recurrent bases is the number of non-recurrent bases in the MSA. These are bases which are either part of relatively rare segregating polymorphisms, or erroneous due to mis-alignment or mis-assembly. Encouragingly, the Chimpanzee sample has by far the largest number of non-recurrent bases in the MSA, a total of 185,330 bp (3.31% of covered bp); in contrast the average human sample has only 19,362 such bases (0.52% of covered bp). Summing across the human samples, a total of 309,896 bp are non recurrent (5.54% of all columns and non-recurrent bp).

Given the relatively small number of samples, all recurrent bases are likely segregating at a reasonable frequency in the population. An important category of recurrent (segregating) bases are those that GRCh37 fails to represent (i.e, not in GRCh 37). On average each human sample has 68,525 such bases (1.84% of covered bp).

Summing across the samples, 329,190 recurrent bases in the MSA (5.88% of columns and non-recurrent bp) do not contain bases in GRCh37.

2.3.5.4 A Comparison of the MSA's Variation Predictions to the dbSNP/1KGP Data Confirms the High Accuracy of the MSA

To assess the accuracy of the MSA I compared its SNV and short ($\leq$ 10 bp) insertion and deletion (collectively indels) predictions made with respect to GRCh37 to the intersection of the dbSNP database [Sherry et al., 2001] and the 1KGP data [1000-Genomes-Project-Consortium, 2010]. In overview, the MSA made 56,080 distinct SNV predictions relative to GRCh37, of which 42,584 (76%) were confirmed by dbSNP. This accounts for 28% of all SNVs currently in the dbSNP/1KGP data. Given that there were only 15 samples other than GRCh37 used in this study, observing 28% of the population variation is significant.

One important set of predicted SNVs are those that are present in C. Ref, as these reflect differences between the pan-genome reference sequence and GRCh37. Approximately 97% of such SNVs are contained in the dbSNP/1KGP data, leaving 264 total possible "false positives" (false positive with respect to the dbSNP/1KGP). The majority of these (91% or 241 SNVs) occurred in bases that were labeled as either being repetitive or proximal to a breakpoint in one or more of the samples. Repetitive regions and breakpoint vicinities are challenging cases and often result in multiple equivalent solutions in alignment. Therefore, it is expected to observe disagreements between the MSA and the dbSNP/1KGP data in these regions. A careful manual analysis of

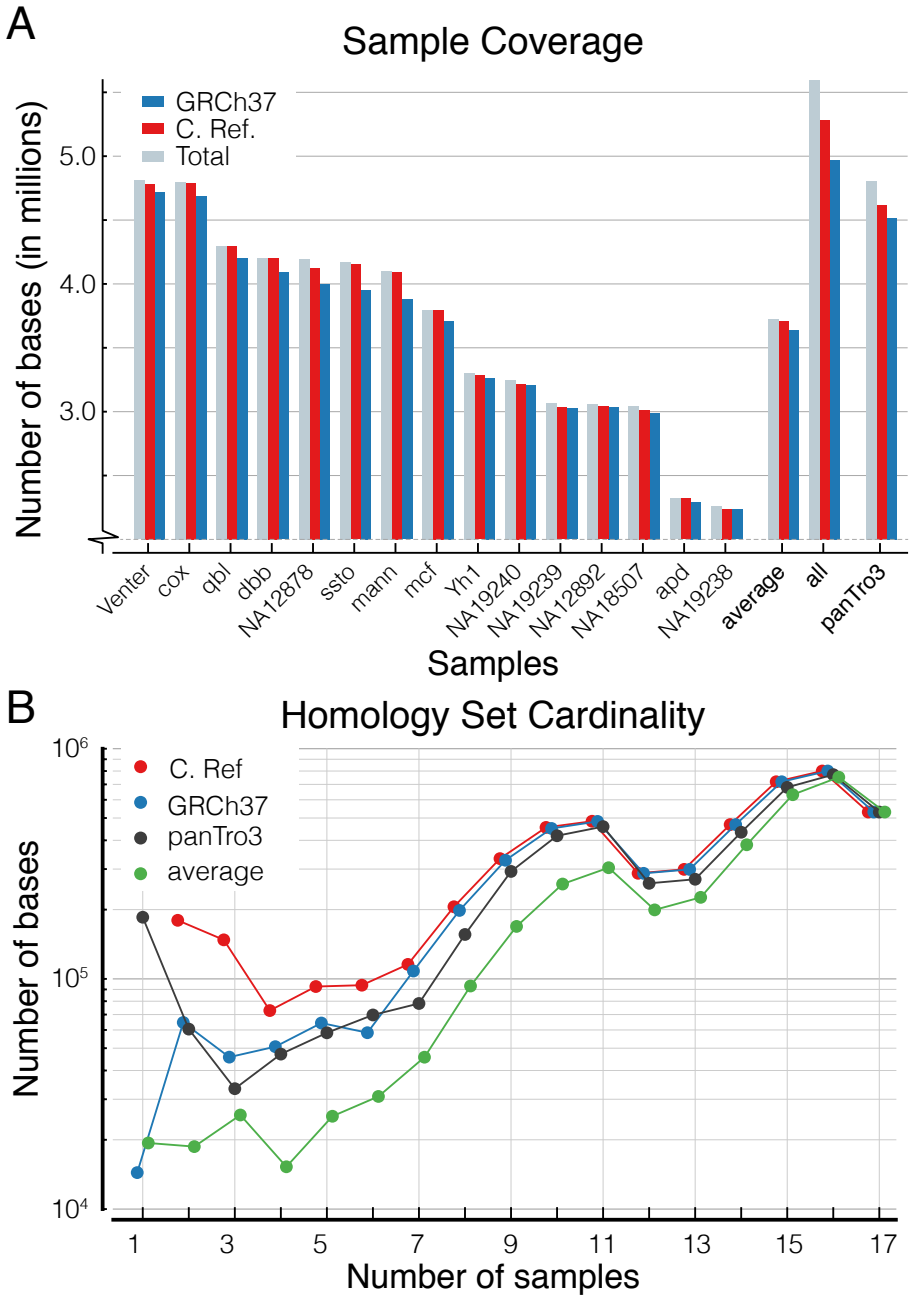

Figure 2.8: C. Ref contains ∼6% of recurrent bases that are not represented in GRCh37, and Sanger sequenced samples have higher coverage than Illumina sequenced samples. (A) The number of bases from each sample mapped to the Cactus MSA, to C. Ref. and to GRCh37. The 'average' category gives the average over all human samples. The 'all' category considers all columns and unaligned bases in all the human samples, i.e. as 1 base per homology set. (B) The number of bases within columns of a given cardinality. 'Average' is average over all human samples.

the other 23 non-repetitive, non-breakpoint-proximal cases of the 264 false positives revealed supporting evidence for 21 of them (Method Subsection 2.5.6 and Appendix Table A.6). Overall, the results confirm the SNVs between C.Ref. and GRCh37 and validate C.Ref.'s quality.

I see a similar picture with short indels to that with SNVs, but a generally higher level of disagreement between the MSA predictions and the dbSNP/1KGP data. Overall, the MSA made 22,360 indel predictions of which 14,575 (65%) were confirmed, accounting for 34% of all short indels currently in the dbSNP/1KGP data.

False Positive SNVs Likely Resulted From the Assembly Quality of the Illumina Sequenced Samples and Reduced by the Recurrence Requirement

With 76% of the total predicted SNVs confirmed by dbSNP, there were 24% of false positives. A large proportion of these false positives came from the Illumina sequenced samples (Figure 2.9 and Supplementary Table *A*.2). On average, the Sanger sequenced samples had a 98% true positive rate in comparison to only a 78% rate in the Illumnia sequenced ones.

To further investigate these false positive SNVs I subdivided the MSA predictions into categories based upon: First, their presence outside of a sequence of GRCh37 annotated as repetitive, and therefore hard to correctly assign homology to; second, their distance from a breakpoint within the MSA, which might result in misalignment, and third, whether they were recurrent, i.e. predicted by multiple samples and therefore

were less likely to be erroneous (Figure 2.9 and Supplementary Tables $A.2$, $A.3$, $A.4$ and $A.5$).

Being outside of repetitive sequence and more than 5 bp from a breakpoint results in a small positive increase in the average true positive rate (2.7% and 3.1% increase, respectively); combined this effect is even stronger (4.9% increase). SNVs within repeats and near breakpoints are therefore likely to be genuine candidates for misalignment and consequent false SNV prediction.

Being recurrent had a small effect on the Sanger samples (avg. 0.6% increase in true positive rate), but a huge effect on the Illumina sequenced samples (avg. 17% increase in true positive rate). Looking at only recurrent SNVs, the overall true positive rate is 95% and is similar between Sanger (avg. 98% per sample) and Illumina sequenced samples (avg. 96% per sample). Looking only at recurrent SNVs also only leads to a 13% average reduction in the total number of SNVs called, a much smaller reduction than that of looking at SNVs only in non-repeat regions (54%) or SNVs not proximal to a breakpoint (19%). The high accuracy in the SNV predictions of the Sanger sequenced samples, together with the significant improvement in accuracy of the Illunima sequenced samples' SNV predictions when the "recurrent" condition was required, suggest that most of the false positives may be attributed to sample's assembly quality (sequencing or assembling errors) and not alignment errors.

Assessing false negative rates is harder as SNV and indel calls for the individual GRCh37 haplotypes were not available. However, given the high true positive rate, I estimate the false negative rate for each of these samples by assuming the MSA

predictions are correct and using the total number of previously reported SNV predictions [Horton et al., 2008]. Given this caveat, in the haploid (Sanger) samples I see an average false negative rate of 2% per sample. In the diploid samples I see an average false negative rate of 59.5% per sample, which is reasonable given that, as mentioned, I expect to miss half of all their variants.

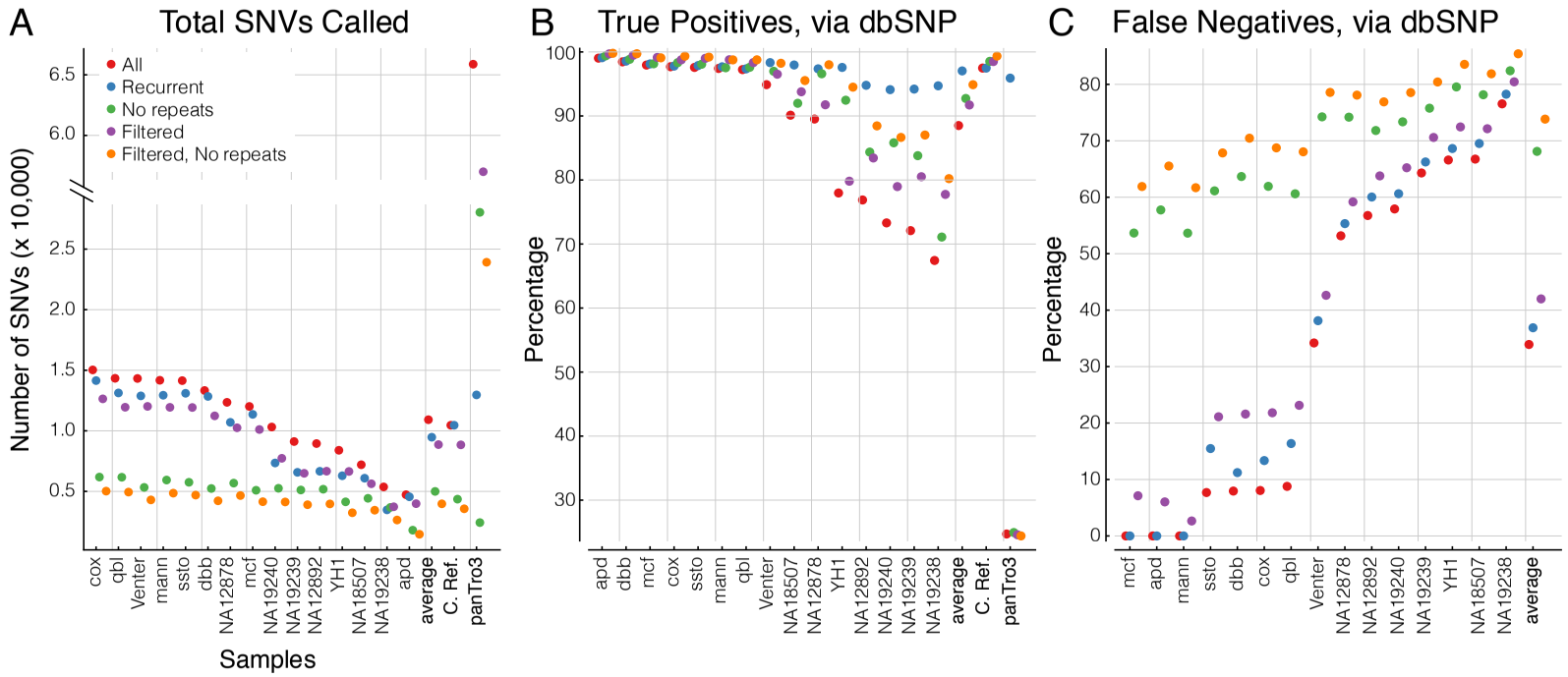

Figure 2.9: A detailed comparison of the SNVs predicted by the Cactus MSA to those in dbSNP shows high accuracy in variation predictions of the Sanger sequenced samples and in recurrent variation predictions of the Illumina sequenced samples. (A) The total number of SNVs predicted by the MSA with respect to GRCh37. (B) The proportion of SNVs predicted by the MSA with respect to GRCh37 already present in dbSNP. (C) The proportion of (previously reported) SNVs for a given sample in dbSNP not predicted by the MSA. Key gives categories of SNVs predicted by the MSA; 'All': All SNVs, 'Recurrent': All SNVs present in at least two samples, 'No repeats': Excluding SNVs at bases labeled repetitive in GRCh37, 'Filtered': Excluding SNVs within columns that are within 5 bp of a breakpoint in any sequence, 'Filtered, No Repeats': Intersection of previous two categories.

2.3.5.5 Variation Rate Comparisons

Having verified many of the MSA's SNV and short indel predictions I compare the rates of these events in each sample with respect to GRCh37 and with respect to C. Ref; Figure 2.10 shows the rates for SNVs, insertions and deletions.

SNVs: C. Ref. Is Closer to the Input Samples than Is GRCh37

With respect to GRCh37 the SNV rate differs between samples from between 0.0021 to 0.0036 SNVs per bp. With respect to C. Ref. the SNV rate differs between samples from between 0.0014 to 0.0027 SNVs per bp. In every sample there are fewer SNVs with respect to C. Ref. than to GRCh37, the difference being 29% on average. To confirm that the reduction in SNVs was not caused only by a bias in the MSA, I constructed versions of C. Ref. excluding a 1KGP sample and then compared SNV calls made with this held out sample using a short read pipeline (see Subsection 2.3.5.9) mapping alternatively to GRCh37 and the held out version of C. Ref. The results of this experiment are shown in Figure 2.10(B), in every case C. Ref. is closer to the sample than GRCh37, the difference being 16% on average, despite more reads mapping to the modified C. Ref. in every case (see Subsection 2.3.5.9).

Indels: C. Ref. Is Inclusive of All Recurrent Segregating Bases and Has More Deletions as a Trade-off

With respect to GRCh37 the rate of insertions averages $2.8\text{x}10^{-4}$ per bp overall,

and is similar to the rate of deletions, averaging 2.6x10^{-4} per bp overall. Conversely, the rate of insertions with respect to C. Ref. is much lower, averaging only 0.5x10^{-4} per bp overall, while the rate of deletions is much higher, at 5.3x10^{-4} per bp overall. This demonstrates a key difference between GRCh37, or likely any existing biological sample, and C. Ref., in that C. Ref. includes all recurrent segregating bases, while any biological sample is likely to have approximately equal numbers of insertions and deletions with respect to any other sample.

2.3.5.6 Contiguity and Non-linear Breakpoints

To create C. Ref. the software chooses an ordering and orientation for the columns in the MSA. To assess this oriented ordering I investigate two complementary metrics: first, analyzing the ordering and orientation in C. Ref. of pairs of ordered and oriented bases within the contigs, and second, by discovering breakpoints implied by the alignment that lead to nonlinear orderings of the contigs.

Correct Contiguity: Both C. Ref. and GRCh37 Maintain the Order and Orientation of the Bases of the Input Samples

A pair of bases within a contig of a sample are *correctly contiguous* [Earl et al., 2011] in a given reference sample if the pair aligns to a single contig within the reference that maintains their order and orientation (Figure 2.11, see Appendix Section $A.3$ for a formal definition). Appendix Table $A.11$ shows the results of this metric for each sample, comparing C. Ref. to GRCh37. I find that in GRCh37 on average

3.2% of selected pairs are not correctly contiguous, but in C. Ref. only 1% of selected pairs are not correctly contiguous. This large difference is accounted for mostly by the first two requirements of correct contiguity, that the bases in the contigs align to bases in the reference. Looking only at pairs which do align to the reference, 176 pairs per million (0.00018%) on average are not correctly contiguous in GRCh37, while 128 pairs per million (0.00013%) on average are not correctly contiguous in C. Ref. Both C. Ref. and GRCh37 maintain the order and orientation of the bases in the input samples. This observation is expected because there are few structural rearrangements present in the input dataset (as will be discussed in the following section).

In Appendix Figures *A*.4 and *A*.5 I analyze correct contiguity, and correct contiguity given that the pairs align to the reference as a function of the separation distance between each pair of bases, but find little evidence for a consistent trend. This is perhaps to be expected given the apparently very small number of pairs that do align but are not correctly oriented.

Nonlinear Breakpoints: A Segregating Inversion Accounts for All the Recurrent Nonlinear Breakpoints in the Sample Set

To assess nonlinear breakpoints I analyze subgraphs of the alignment (see Methods Subsection 2.5.4). In concordance with the correct contiguity analysis, I find relatively few such breakpoints, the median being 2 with respect to GRCh37 and 1 with respect to C. Ref. (Figure 2.12). Analyzing these breakpoints, I find one recurrent inversion segregating in the population that explains all the recurrent nonlinear

breakpoints (7 across all samples) observed with respect to C. Ref. With respect to Chimpanzee I find evidence that the segregating inversion is present in GRCh37, the MANN, the MCF and NA19239 samples, and not present in the remaining samples, except for NA19238, where missing information precludes us from knowing. Figures 2.13 and 2.14 show the inversion in the UCSC genome browser, using the MSA as well as independently generated alignments that confirm it. The inversion was missed by previous studies of the eight GRCh37 haplotypes [Traherne et al., 2006, Horton et al., 2008], but a nearby inversion described in the Database of Genomic Variants [Zhang et al., 2006] may be related (see Appendix Figure 2.14). Figures 2.13 and 2.14 also demonstrate that for this region, C. Ref. provides a better comparative genomic visualization than does GRCh37, not only because the majority of the samples do not have the inversion but also because GRCh37 does not contain the $\sim$3000 bases surrounding it that the majority of the samples do.

The remaining breakpoints with respect to C. Ref. (38 across all samples) are, apart from being non-recurrent, mostly present (35 of them) in the Illumina assemblies, making it more likely that these are technology related misassemblies.

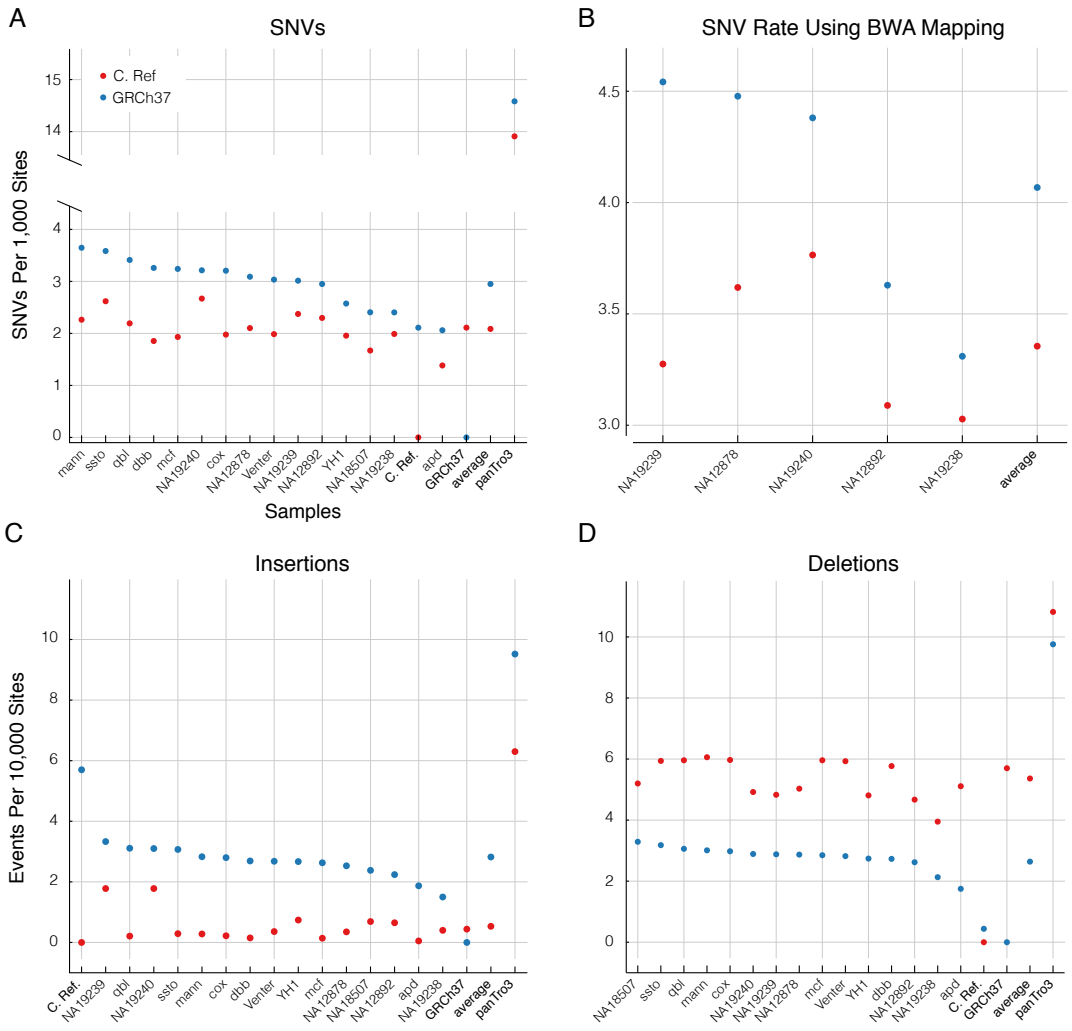

Figure 2.10: A comparison of SNV and indel rates between C. Ref. and GRCh37 shows that C.Ref contains less SNVs and is more inclusive, and has more deletions as a trade-off. (A) The number of SNVs per site (position) of each sample, as predicted by the Cactus MSA with respect to GRCh37 and C. Ref. (B) As in (A), but as predicted by BWA/pileup [Li and Durbin, 2009] and only for the Thousand Genomes Project samples. (C) As in (A), but for insertions. (D) As in (A), but for deletions.

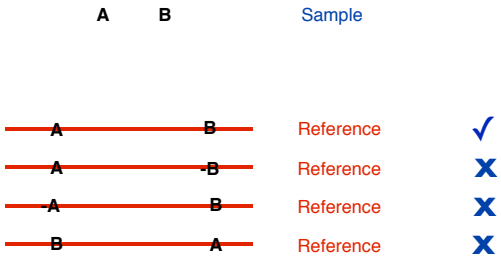

Figure 2.11: An illustration of correct contiguity. The blue and red horizontal lines represent the sample and the reference sequences, respectively. A and B represent two bases on the sample that get mapped to the reference. Four examples of the mapping scenario are shown, only the first one has a correct contiguity.

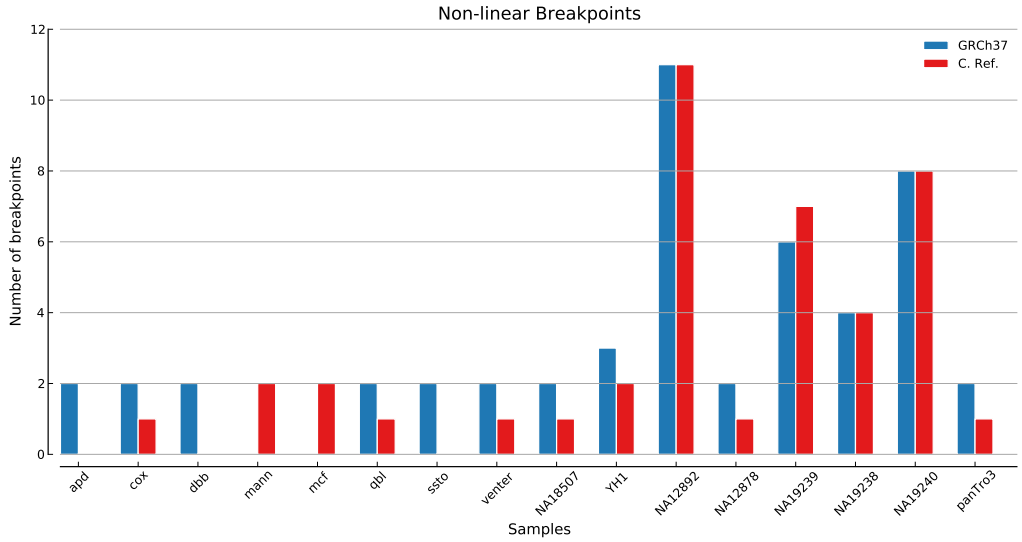

Figure 2.12: A histogram showing the number of non-linear breakpoints per sample with respect to GRCh37 and separately with respect to C. Ref.

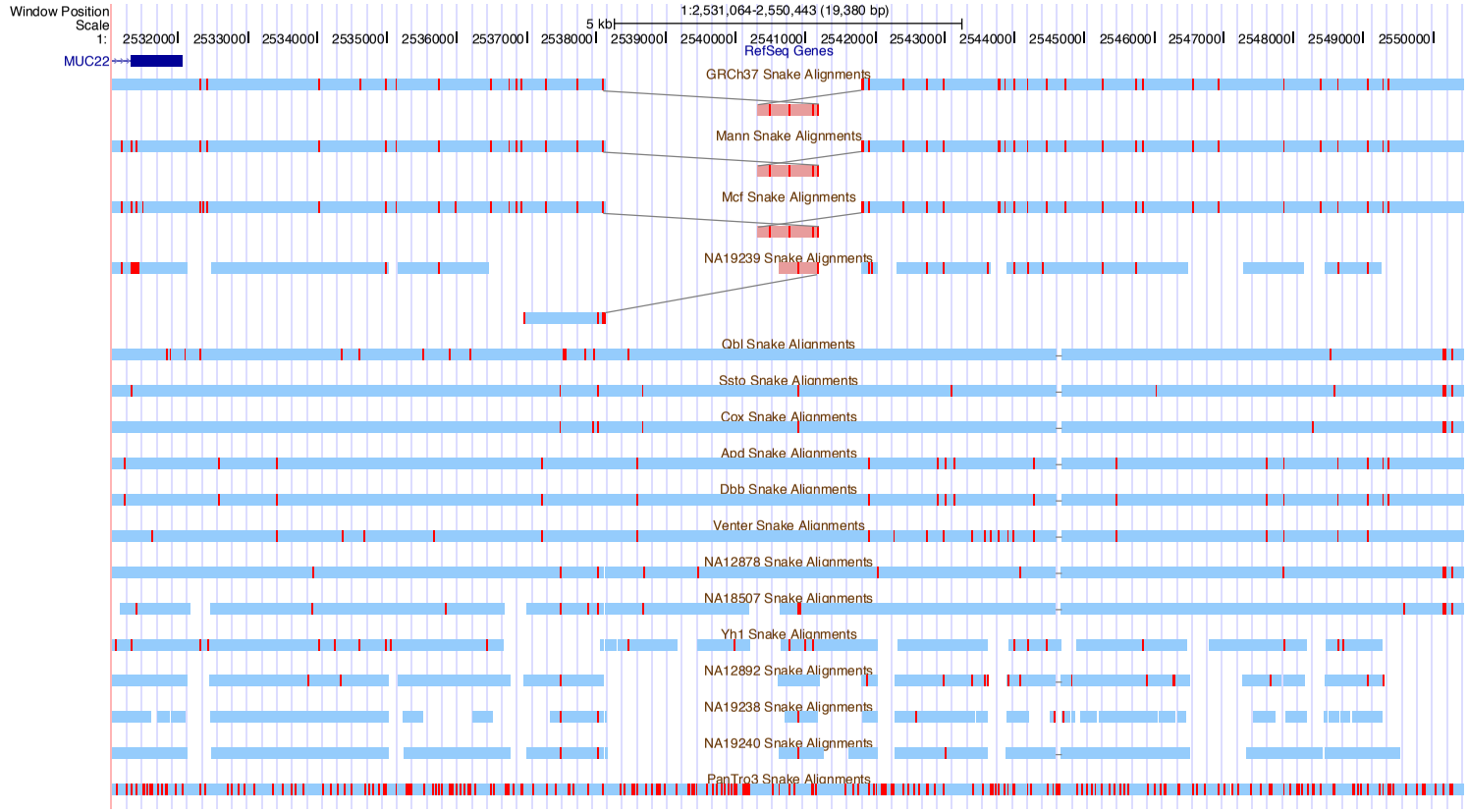

Figure 2.13: A UCSC Browser screenshot showing a segregating inversion in a prototype C. Ref. MHC reference browser. With respect to Chimpanzee I find evidence that the segregating inversion is present in GRCh37, the Mann, the MCF and NA19239 samples, and not present in the remaining samples, except for NA19238, where missing information precludes us from knowing. The figure is arranged similarly to Figure A.14

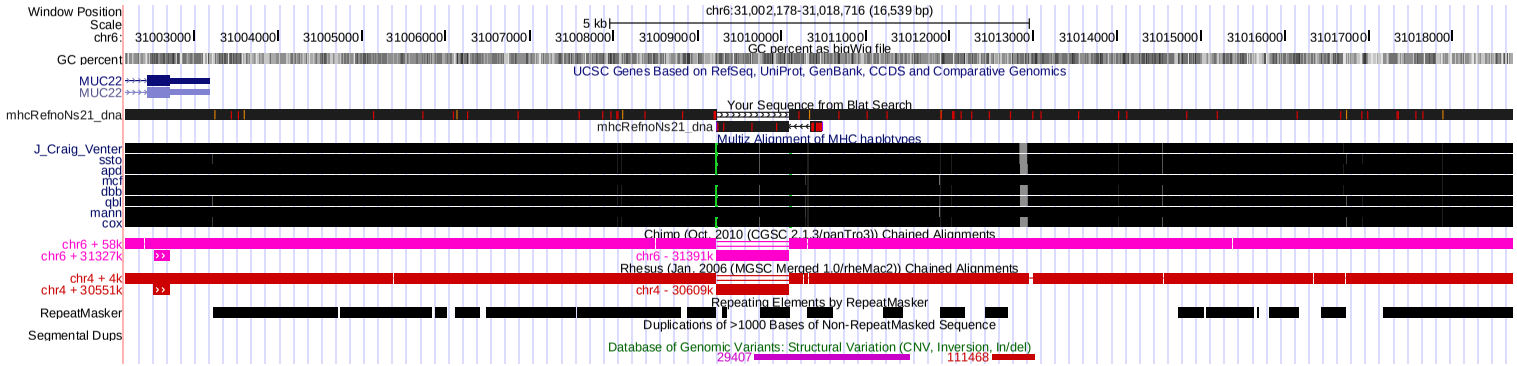

Figure 2.14: A UCSC Browser screenshot showing the homologous region in GRCh37 of that shown in Figure 2.13. The screenshot shows the MULTIZ alignment of the 8 GRCh37 MHC haplotypes, generated independently to the Cactus alignment, which also implies the same inversion. BLAT alignments (shown) of the inverted sequence also demonstrate the inversion, they additionally show that the inverted sequence does not match with greater than 90% identity anywhere else in GRCh37 (not shown). UCSC Chain and Net alignments [Kent et al., 2003] of Chimpanzee and Rhesus Macaque show the same inversion with respect to the GRCh37 assembly. The inversion maybe related to an inversion identified in the DGV [Zhang et al., 2006] as Variation_29407 at chr6:31009686-31011525 in the UCSC HG19 Browser.

2.3.5.7 Large Indels

As would be expected, the frequency of short indels of a given length approximately exponentially declines with length [Kent et al., 2003] (see Appendix Figures $A.6$ and $A.7$). Consequently, the number of bases on average affected by indels of a given length generally declines with this length (see Appendix Figures $A.8$ and $A.9$). However, I also find exceptions to this trend: a few very large indels (length $\geq$ 1000 bp) that contribute disproportionally (60% on average per sample) to the total number of bases affected by indels (see Appendix Figures $A.10$ and $A.11$).

Due to the presence of missing data and the use of only one outgroup, it is hard to determine if an insertion or deletion with respect to C. Ref. or GRCh37 is truly the result of the gain of new bases or the loss of previously present bases. However, for large indels I can search for close homologs to partially characterize them. In brief, I took the set of predicted insertions with respect to GRCh37 larger than 1000 bp and aligned them to the whole of GRCh37 (not just the MHC, see Appendix Section $A.4$). Table 2.1 shows these results. For a total of 76 insertions, 44 (58%) map best within the MHC, 9 (12%) apparently resulting in copy number change. I find 18 (24%) map best outside the MHC and that the remaining 14 (18%) were unmappable. A manual analysis of these unmappable sequences indicates they are likely true evolutionary deletions in GRCh37. A large proportion of sequences (68%) have $\geq$ 50% bases that were labeled repetitive and a large portion of sequences mapped well to multiple locations (29%), given their size, this suggests many of these events may have arisen by repeat-related

	Bases	% Total bases	Events	% Total Events
Mapped to MHC	431,154	63.9	44	57.9
Mapped Outside MHC	189,425	28.1	18	23.7
Unmapped	54,263	8.0	14	18.4
Multi-mapping	54,344	8.1	22	28.9
Copy Number Change	31,405	4.7	9	11.8
Tandem Duplications	1,910	0.3	1	1.3
Repeats	495,384	73.4	52	68.4
Total	674,842	100	76	100

Table 2.1: The origin of large ($\geq$ 1000 bp) insertions with respect to GRCh37. Columns: 'Bases': Number of bases in insertions. 'Events': Number of insertions. '% Total bases': Proportion of total bases. '% Total Events': Proportion of total events. Rows: 'Mapped to MHC': Events which mapped best within MHC (see Appendix Section A.4). 'Mapped Outside MHC': Events which mapped best outside MHC. 'Unmapped': Events which did not map well anywhere. 'Multi-mapping': Events which mapped well to multiple locations. 'Copy number change': Events which resulted in copy number change in the sample. 'Tandem duplications': Events that mapped within 1000 bases of their insertion location and resulted in copy number change. 'Repeats': Number of insertions with $\geq$ 50% bases classified by Repeat Masker as repetitive [Smit et al., 2010]. 'Total': Total numbers of bases and events.

mechanisms.

2.3.5.8 RNA Alignments: C. Ref. Contains All GRCh37's RefSeq Transcripts that Mapped to the MHC Region as well as Additional Transcripts Missing from GRCh37

Any reference genome must contain a set of representative gene structures for the species. I took the RefSeq human transcripts [Pruitt et al., 2012] and aligned them independently to GRCh37 and C. Ref, using stringent identity and coverage parameters and keeping only the best alignments for each transcript (See Appendix Section A.5). Requiring 95% of the transcript to be aligned at 95% identity, I find 210 RefSeq genes whose transcripts all align to both the MHC region of GRCh37 and C. Ref. better than the remainder of GRCh37 (Appendix Table A.13). I find no transcripts that align to the GRCh37 MHC region and not to C. Ref., but 3 genes with transcripts that align to C. Ref. but not to GRCh37. One of these genes, HLA-DQB1, has 3 RefSeq transcripts, of which only one (NM_001243962) did not map to GRCh37. By reducing the required identity for this transcript to 90%, I was able to obtain a reduced stringency match.

The HLA-DRB Hypervariable Region: C. Ref. Includes Segregating HLA-DRB Genes Not in GRCh37

The remaining two genes with transcripts that mapped to C. Ref. but not to GRCh37 (HLA-DRB3 and HLA-DRB4) both were entirely missing from GRCh37. Interestingly, both mapped to C. Ref. within large indels contained within a region that corresponds to the HLA-DRB Hypervariable Region. In C. Ref. the HLA-

DRB Hypervariable Region [Traherne, 2008] contains a large number (29 events) of large insertions with respect to GRCh37; Appendix Figure A.14 shows this region in a prototype C. Ref. genome browser, while Appendix Figure A.12 shows the entire MHC for C. Ref. All the expected HLA genes in this region are present (known genes HLA-DRB5, HLA-DRB1 and pseudogenes HLA-DRB9, HLA-DRB6), as well as the two extra HLA genes (HLA-DRB3, HLA-DRB4) described, and also pseudogenes (HLA-DRB2, HLA-DRB7, HLA-DRB8) that are recurrent in the input samples and known to be segregating in humans but not present in GRCh37 [Stewart et al., 2004, Traherne et al., 2006, Horton et al., 2008].

One issue with the use of RefSeq transcripts is that they have been constructed and curated using the existing reference genome, and therefore are potentially biased against a novel reference based upon a more comprehensive set of samples. To address this possibility I repeated the described alignment process using all the GenBank human RNAs. Though these sequences could not all be associated with individual genes, at a 95% identity and coverage level I found that 236 RNAs (7.6% of all RNAs that mapped either to C. Ref. or the MHC region of GRCh37) mapped to C. Ref. but not to GRCh37 and only 22 RNAs (0.7%) mapped to the MHC region of GRCh37 but not to C. Ref. This large difference indicates that there are potentially further gene structures missing from GRCh37 that are contained in C. Ref.

The RCCX Module: C. Ref. Is Inclusive of All Recurrent Copies

Converse to a region containing genes within large insertions with respect to

GRCh37, an interesting region with significant large deletions with respect to GRCh37 is the RCCX module. The RCCX module, typically containing the genes STK19 (RP), C4A/B, CYP21 and TNXB, may be duplicated or triplicated, resulting in additional copies of the complement component C4 genes, as well as the additional pseudogenes CYP21A1P, TNXA, and STK19P [Horton et al., 2008]. The C4 genes of individual haplotypes can be in either or both of two versions, C4A and C4B, and each gene can be in either long (C4AL, C4BL) or short (C4AS, C4BS) forms. Among the GRCh37 haplotypes, GRCh37 (C4AL, C4BL), SSTO (C4BL, C4BL) and DBB (C4AL, C4BS) have been previously reported as bimodular; COX (C4BS) and QBL (C4AS) as monomodular; and the MCF (with evidence of being bimodular), APD and MANN haplotypes as incomplete in this region, due to sequence gaps [Horton et al., 2008]. C. Ref. contains the consensus copy number of the input samples, i.e a duplicated RCCX module (see Appendix Figure A.13). The Cactus MSA confirms previously reported annotations [Horton et al., 2008], except for the COX and SSTO samples. The MSA predicts that COX has C4AS instead of the C4BS, with the alignment between the COX sequence and the C4A gene mapped perfectly without any substitution or indel. The previously reported annotation appears less parsimonious, as it creates two large deletions instead of one. The MSA predicts that SSTO has C4AL and C4BL instead of two C4BL genes.

2.3.5.9 Short Read Mapping: More Reads Map to C. Ref. Than to GRCh37 But Less Reads Map Uniquely to C. Ref.

In Section 2.3.5.5 I demonstrated that C. Ref. is inclusive, having relatively few insertions with respect to the input samples. However, the cost of including segregating but potentially rare subsequences in a C. Ref. is the inclusion of potentially rare breakpoints. Such rare breakpoints could disrupt more common subsequences and potentially make mapping and annotation more difficult. To assess the tradeoffs made, I test C. Ref. as a target for mapping experiments. In brief, I constructed versions of C. Ref. excluding a held out 1KGP sample and then compared mappings made with BWA [Li and Durbin, 2009] of the held out sample to GRCh37 and the held out C. Ref. (see Appendix Section *A*.6). I mapped Illumina unpaired and paired reads. For each paired read I required that both its ends mapped in the proper orientation given the pairing constraint and were separated by at most 1000 bases, calling such reads *properly paired*.

I find that on average 5,958 (0.6%) more unpaired reads and 13,828 (0.5%) more paired reads map to C. Ref. than to GRCh37 (Figure 2.15 and Appendix Table A.14). Of note, this is in reasonable agreement with the average proportion of bases in these samples (0.89%) that are recurrent but which do not map to GRCh37.

Converse to an increase in the numbers of mapping reads, I find on average per sample 9,656 (0.26%) fewer unpaired reads and 6,413 (0.25%) fewer paired reads map uniquely to C. Ref. than map uniquely to GRCh37. To analyze this reduction I investigated reads that map uniquely to GRCh37 but non-uniquely to C. Ref., calling

such reads *GRCh37 mapping discordant*. Table 2.2 shows characteristics of these reads; I find that 73% of the bases to which such reads map are labeled repetitive, which is 1.4 times the GRCh37 average. I also find that there are 3 fold more SNVs in the dbSNP/1KGP data called at these bases than on average. I hypothesize this enrichment for SNVs is due to the absence of a orthologous sequence in GRCh37 that is present in the sample being mapped (Figure 2.16). This missing ortholog then results in the appearance of unique mapping to its paralog. In Appendix Table *A*.16 I turn the experiment around and look at reads which map uniquely to C. Ref. but non-uniquely to GRCh37. I see similar effects, but 2.9 fold fewer mapping discordant reads than when doing the experiment with the references reversed.

In Appendix Figures *A*.15 and *A*.16 I compare mapping to C. Ref. vs. the other GRCh37 haplotype samples, both individually and in combination. Individually they are all poorer than GRCh37 and therefore substantially poorer than C. Ref. Combined they could not be used for paired or unique mapping, as a large proportion of reads then map nonuniquely or between the samples in the case of paired reads. However, for mapping reads combined they were 1% better than C. Ref, which is expected as even a consensus sequence is unlikely for this purpose to be superior to a substantial collection of individual samples. However, mapping to multiple individual references are computationally costly and are not commonly practiced.

In Appendix Figure *A*.17 I compare mapping to versions of C. Ref. in which columns and non-recurrent bases were included or excluded according to a parameter α, which determines the minimum cardinality of homology sets included in C. Ref. I find

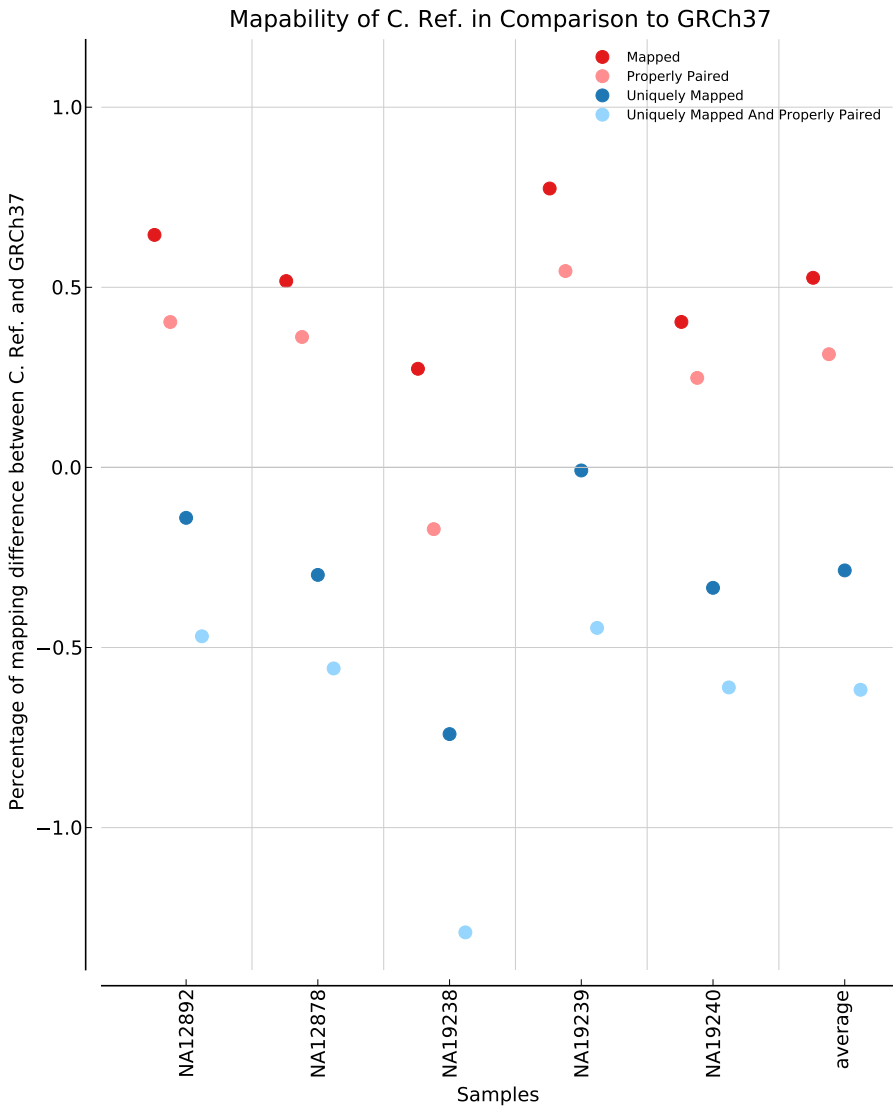

Figure 2.15: A comparison of short read mapping to C. Ref. and to GRCh37 shows that C. Ref. has slightly more mapped and properly paired reads and slightly less uniquely mapped reads. The y-axis is the ratio of the difference between the number of reads of a sample mapping to a C. Ref. constructed without the sample in question and the number of reads of the sample mapping to GRCh37 over the number of reads of the sample mapping to GRCh37. 'Mapped': For all reads. 'Properly Paired': For paired reads (see methods for definition of 'proper pairing'). 'Uniquely Mapped': As 'Mapped', but ignoring reads that map to multiple region equally well. 'Uniquely Mapped and Properly Paired': As 'Properly Paired', but ignoring reads that map to multiple region equally well.

Sample	MD Reads	Total MD bases	% Repeats	SNV Rate	dbSNP ESR	bcftools ESR
NA12878	52,731	354,575	74.08	0.0144	2.6	3.2
NA12892	30,542	264,912	73.63	0.0100	2.7	2.8
NA19238	31,335	249,473	71.34	0.0088	3.0	2.7
NA19239	40,308	266,904	71.77	0.0161	2.8	3.5
NA19240	50,928	327,527	72.71	0.0128	2.7	2.9
average	41,168	292,678	72.70	0.0124	2.7	3.1

Table 2.2: An analysis of GRCh37 mapping discordant reads show that these reads map mostly to repetitive regions that have an enrich in SNVs called by dbSNP/1KGP. 'MD Reads': Total GRCh37 mapping discordant reads. 'Total MD bases': Total mapping discordant bases in GRCh37. '% Repeats': Proportion of mapping discordant bases in GRCh37 classified as repetitive. 'SNV Rate': Number of SNVs predicted by dbSNP/1KGP per mapping discordant base in GRCh37. 'dbSNP ESR': dbSNP 'Enriched SNV Ratio', ratio of mapping discordant SNV rate (previous column) over overall SNV rate predicted by dbSNP/1KGP. 'bcftools ESR': bcftools 'Enriched SNV Ratio', ratio of mapping discordant SNV rate over overall SNV rate predicted by bcftools.

that including non-recurrent bases in C. Ref. (e.g. $\alpha = 1$) actually substantially reduces the number of reads that map, and that $\alpha = 2$ is optimal for non-unique mapping for all samples.

2.4 Discussion

In this chapter, I defined a problem useful for creating a pan-genome reference between closely related genomes, described proofs of its NP-hardness and showed principled heuristics to find approximate solutions. Simulations showed the tradeoffs between optimizing for conserved order relationships and minimizing DCJ operations.

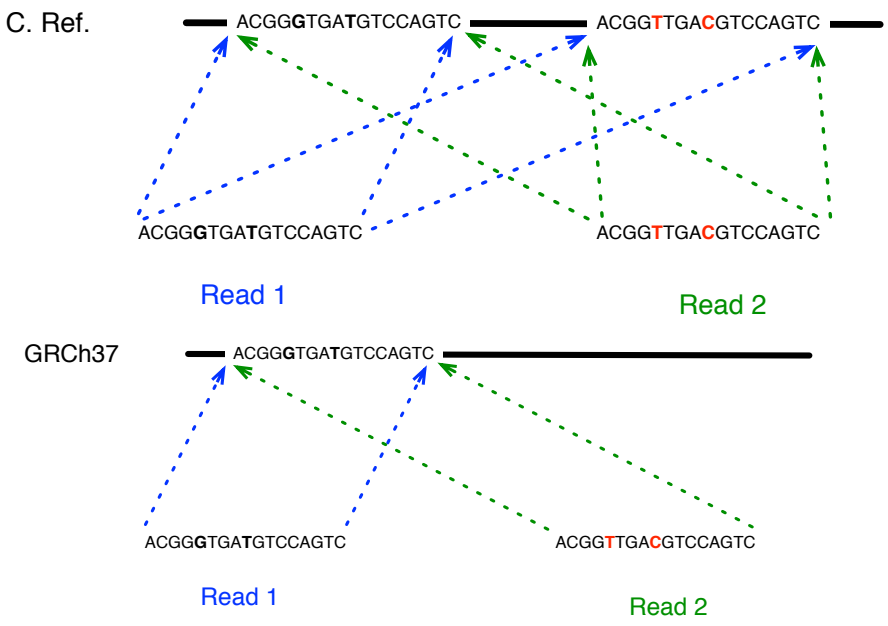

Figure 2.16: An example scenario of reads mapping to a paralog when the true ortholog is missing. Here, C. Ref. contains the orthologous sequence of Read 2 and therefore results in non-unique mapping. GRCh37 does not have orthologous sequence of Read 2 and results in the mis-mapping of Read 2 to the orthologous sequence of Read 1, resulting in unique-mapping, but higher SNPs.

I applied the methods to construct a pan-genome reference for a limited set of human MHC sequences. In light of the results, I now try to summarize the arguments for and against this approach, both in comparison to the existing human reference genome, GRCh37, and in comparison to any potential genome from a single individual.

The strongest argument for a pan-genome reference is that it has the potential to be a better 'median' genome for a population as a whole than any existing individual genome. For example, by picking the most frequently occurring nucleotide in each column it is guaranteed to be as close as possible on aggregate to the genomes in the sample set in terms of nucleotide substitutions. Here I observed an average 29% reduc-

tion in SNVs per sample when comparing against C. Ref. rather than GRCh37. I have demonstrated by contiguity and breakpoint analysis a small but significant improvement in the order of C. Ref. vs. GRCh37.

Another argument for a computationally derived reference is that it can be made 'inclusive', i.e. missing few subsequences that are segregating within the population. In this study, where I have accounted for around 34% of known MHC indels, I find GRCh37 is missing 2819 deletions that are recurrent in the other samples and therefore included in C. Ref. These missing sequences, which account for around 6% of bases in the region, contain important variants that are also segregating within the population. For example, I find that within the HLA-DRB hypervariable region entire genes are missing from GRCh37 that are segregating recurrently in the population.

Critics may argue that a pan-genome reference is 'artificial', in that it does not represent a haploid genome in any one individual. However, this is also true of the current GRCh37 reference, which is a chimera of haplotypes [International-Human-Genome-Sequencing-Consortium, 2004]. Furthermore, the technology for completely sequencing and assembling the haplotypes of a human genome (or a genome of a member of a similarly outbred population of any large mammal) to the 'finished' quality of the GRCh37 genome is not yet readily available [Salzberg et al., 2011, Earl et al., 2011, Gnerre et al., 2011].

Instead of relying on one reference genome for a species, multiple reference genomes could instead be used. Indeed this approach is being taken by the Genome Reference Consortium [Church et al., 2011] in identifying and sequencing al-

ternative haplotypes in highly polymorphic regions, the MHC being one such example [Horton et al., 2008]. Given such a set of reference genomes there are two obvious approaches.

The first approach would identify the best reference genome for a particular sample being studied from a set of multiple reference genomes and then use it exclusively. For the limited MHC region, I have shown no sample in the set of GRCh37 haplotypes and Venter is as good as C.Ref. for mapping paired or unpaired short reads for any of the 1KGP samples studied. The mapping experiments also identified a subtle paralogous mapping issue present in GRCh37, and likely any single sample, that is much less apparent in C. Ref due to its inclusiveness. For each 1KGP sample a substantial number of reads map uniquely to GRCh37 and non-uniquely to C. Ref. I demonstrated that the sites to which these reads map are substantially enriched for SNVs and repeats, suggesting that substantial numbers of such SNVs are likely paralogous mappings as a result of the true ortholog being missing from GRCh37.

The second approach would use a set of multiple reference genomes in combination. This combined approach necessitates a homology map between the multiple references, to avoid chronic ambiguities in the multiple mapping of both the reads and annotations. In fact, such a homology map can be fully represented by the genome MSA described here, but the combination of the MSA and samples leads a more complex data structure than a consensus reference sequence, and is naturally described as a graph structure. Such a graph is conceptually described in the pan-genome reference problem above, and in a more sophisticated form elsewhere [Paten et al., 2011a]. It

is likely that such graphs will play an increasingly important part in reasoning about population variation, but algorithms that use the reference, such as mappers and those that reason about functional annotations, will need to be largely rewritten to take advantage of such structures. This is a laudable long term direction, but for algorithms in the medium term and probably always for human consumption, there will be a need to produce an ordering through such graphs that can therefore be considered a consensus reference.

Ancestral reconstructions that more strongly preserve order relationships may be preferable, all other things being equal, due to the selective pressure to maintain chromosomal and reproductive compatibility. Being somewhat analogous to methods for ancestral reconstruction, this work is also concordant with methods that pursue perfect rearrangement scenarios. The use of the cactus graph as a novel principled heuristic method of decomposition for the problem is in this spirit, and may well also be useful for breaking down rearrangement median problems.

The pan-genome reference problem also has close similarities with sequence assembly problems, which have variants explicitly described on bidirected graphs [Medvedev and Brudno, 2009]. In particular, the scaffolding problem given paired reads involves arranging a set of "scaffold" sequences in a partial order to essentially maximize the numbers of consistently ordered, oriented and spaced paired reads. Apart from the additional constraint on spacing, the scaffolding problem with paired reads can be defined equivalently to the pan-genome reference problem.

Another utility of the pan-genome reference is in visualization of variation data

as it provides a view of the alignment not typically possible from any input genome. In addition, pan-genome references, being comprehensive and consensus orderings, are likely to prove useful for other purposes. Where reference genomes are currently used for computational convenience, for example in read compression, and are not integral for biological interpretation, a pan-genome reference may present a useful alternative to current reference genomes. Additionally, given that (sequence) graphs do not have a implicit linear decomposition, having a pan-genome coordinate system on such graphs could prove useful in processing multiple alignments.

2.5 Materials and Methods

2.5.1 Sequence Assemblies

The input samples include 16 human MHC haploid assemblies. In each case sets of scaffolds were obtained as described below and then converted into sets of contigs. This was done by splitting the scaffolds at scaffold gaps, defined as contiguous subsequences of 10 or more 'N' or 'n' characters, resulting in the removal of the scaffold gaps and the replacement of previously contiguous scaffolds with multiple separate contigs.

Of the assemblies, 8 were the GRCh37 haplotypes [Horton et al., 2008, Church et al., 2011]. These sequences (chr6:28477754-33448354, chr6_apd_hap1, chr6_cox_hap2, chr6_dbb_hap3, chr6_mann_hap4, chr6_mcf_hap5, chr6_qbl_hap6, chr6_ssto_hap7) were obtained from the UCSC genome browser GRCh37/hg19 database.

One assembly was the Venter MHC sequence (chr6:28284180-33170530), which was extracted by mapping the Venter assembly (September 2007 release, ftp://ftp.jcvi.org/pub/data/huref/) to the GRCh37 MHC loci. The mapping was done using LASTZ (http://www.bx.psu.edu/~rsharris/lastz/, version 1.02.00), with the minimum identity set to 97% and default parameters otherwise.

Two other assemblies came from the 'African' (NA18057) [Bentley et al., 2008] and the 'Asian' (Yh1) [Wang et al., 2008] genomes. For these two genomes, I obtained the scaffolds from the BGI de novo assemblies (see ftp://public.genomics.org.cn/BGI/yanhuang/Genomeassembly/african2.scafSeq.closure.gz and ftp://public.genomics.org.cn/BGI/yanhuang/Genomeassembly/asm_yanh.scafSeq.closure.gz, respectively) and extracted out the sequences that mapped to the GRCh37 MHC loci as done for the Venter assembly.

The last 5 assemblies were from the 1000 Genomes pilot project trios, including NA12878, NA12892, NA19238, NA19239 and NA19240. The NA12878 MHC assembly was extracted from the recently de novo assembled NA12878 genome [Gnerre et al., 2011] (see http://www.ncbi.nlm.nih.gov/nuccore?term=GL582980:GL586310), again by mapping the scaffolds as described above.

The other 4 assemblies were made using the Velvet de novo assembly program [Zerbino and Birney, 2008], version 1.1.06. For each assembly, I only used the Illumina reads http://www.illumina.com/systems/hiseq_2000.ilmn that were mapped to the GRCh37 MHC main region, using the 1KGP alignments (downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NAxxxxx/alignment/, where NAxxxxx

is replaced with the sample name, e.g. NA12892). The Velvet parameters used were as follows: *kmer 25, exp_cov auto*, *ins_length* and *ins_length_sd* obtained using the perl script *observed-insert-length.pl*, which was included in the Velvet package (`http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf`).

In addition, I also added the MHC sequence of the chimpanzee reference assembly as an outgroup. This sequence was extracted from the UCSC genome browser chimpanzee (assembly panTro3) by mapping the GRCh37 MHC main region to this assembly using the 'Covert' function of the UCSC browser.

2.5.2 Creating Human Haplotype Alignments

To create the alignments (homology relation) I use an adapted version of the Cactus alignment program [Paten et al., 2011b], with the default parameters and $\theta = 10^{-4}$. The θ values between 10^{-2} and 10^{-6} were also tested and produced similar results.

The Cactus program's CAF algorithm starts by using the LASTZ pairwise alignment program (`http://www.bx.psu.edu/~rsharris/lastz/`, version 1.02.00) to generate a set of pairwise alignments between all the input sequences. It then filters these pairwise alignments to create a consistent multiple sequence alignment.

In the adapted version of Cactus used for this work the following parameters are passed to LASTZ: *–hspthresh=1800 –identity=X*, where $X = 95 = \lfloor 100 - \frac{300}{4}(1 - e^{-d\beta\frac{4}{3}}) \rfloor$ is the maximum likelihood identity expected by the Jukes Cantor model [Cantor and Cantor, 1969], given a liberal estimate of the maximum evolutionary distance between the human and the Chimpanzee outgroup of $d = 0.015$

[Patterson et al., 2006], and a conservative factor of $\beta = 3$ to allow for regional accelerations in the substitution rate.

Even with this reasonably high identity threshold the resulting Cactus alignment contains many intra-haplotype homologies, resulting from recent duplications. The aim is to construct a "consensus" reference sequence, such that if a column contains n positions within each reference sample, the resulting reference sequence will also contain n copies of the column. To achieve this, the Cactus CAF algorithm's annealing and melting cycles were modified to further partition the columns containing duplications into multiple columns, such that each of the resulting columns contains at most one position from each sample.

For a set of sequences and a homology relation, let G' be a graph whose nodes are the ends of the columns and whose edges are the direct adjacencies. A *group* is a connected component of G'. For a set of columns C a *synteny partition* is the removal (melting step in the CAF algorithm) of homologies between positions in C and then their selective re-addition (annealing step in the CAF algorithm), allowing only homologies between positions within the same group of the modified graph; Appendix Figure 2.6 illustrates this process. The process of synteny partitioning is repeated during each round of annealing and melting in the CAF algorithm, which otherwise remains unchanged to that described in [Paten et al., 2011b]. To remove tandem duplications, which survive synteny partitioning, at the end of the CAF algorithm, any columns containing multiple positions within any single sample are melt (removing homologies between positions). The Cactus BAR algorithm [Paten et al., 2011b], which models

only substitutions, insertions and deletions, is then able to rediscover a subset of these homologies, Figure 2.7 illustrates this. Using the α parameter (by default $\alpha = 2$), which only includes columns with a minimum number of positions in the reference, the process of synteny partitioning can be used to heuristically achieve an approximation to a consensus copy number in the reference sequence.

2.5.3 MSA Post Processing

Having constructed an initial MSA and C. Ref. for the samples above the alignment was "trimmed" so that the GRCh37 sample, which is a single contig, was present in the first and last columns of C. Ref. This trimming resulted in a MSA and C. Ref. that could then be fairly compared with GRCh37, as no sequence included in the MSA mapped to before or after the interval defined by GRCh37.

The trimming was achieved by recomputing the MSA and C. Ref. with exactly the same parameters as in the initial run, but with suffixes and prefixes of contigs that mapped to columns in C. Ref. that preceded the first column containing positions from GRCh37 or proceeded the last column containing positions of GRCh37 removed.

2.5.4 Identifying SNVs, Indels and Nonlinear Breakpoints in the MSA

Given the MSA and a designated reference sample, generally either C. Ref. or GRCh37, variation predictions for each of the other input samples are made. This was achieved by analyzing columns and subgraphs of the MSA.

2.5.4.1 SNVs

For each column containing a position from an input sample and a position from a chosen reference a SNV is predicted for the sample with respect to the reference if the oriented bases of the two positions differ. The set of SNVs for a given input sample and reference is then the set of all such SNVs.

2.5.4.2 Indels

Let $G = (V, E)$ be the bidirected graph for the MSA constructed such that $\alpha = 1$. Let $G' = (V' \subset V, E' \subset E)$ be the subgraph of G containing only nodes and direct adjacency edges representing a chosen reference sample R and input sample T. Due to the synteny partitioning process, any column in G' represents at most one position from each of R and T, while because $\alpha = 1$ every unaligned position is also represented by a node.

Let C be an $M, 2$ cycle in G'. C has one positive node, A, and one negative node, B, and both must contain positions from R and T. C can be subdivided into two paths P_1 and P_2 that both include A and B, but are otherwise disjoint. Let x_i and x_j be a pair of positions such that $[x_i] = A$, $[x_j] = B$ and P_1 represents $x_i <_S x_j$. Similarly, let y_k and y_l be a pair of positions such that $[y_k] = A$, $[y_l] = B$ and P_2 represents $y_k <_S y_l$. Without loss of generality assume that x is the sequence in R and y is the sequence in T. If $j - i > 1$ then a deletion is counted in T with respect to R of length $j - i$. Similarly if $l - k > 1$ then an insertion is counted in T with respect to R of length $l - k$.

2.5.4.3 Nonlinear Breakpoints

Adapting the definition given for indels, let C now be an $M, 1$-cycle such that $M > 1$. Let A be the non-balanced node. C represents one non-linear breakpoint if there exists another node B in C such that the two paths P_1 and P_2, defined for A, B and C as before, both represent one or more adjacencies.

2.5.5 dbSNP/1000 Genomes Project Comparisons

I compared the MSA's SNV and short ($\leq$ 10 bp) indel predictions made with respect to GRCh37 to the intersection of the dbSNP database (build 134) [Sherry et al., 2001] and the 1KGP data (release 20110521) [1000-Genomes-Project-Consortium, 2010].

The dbSNP data for the GRCh37 MHC was obtained from the UCSC Genome Browser 'snp134' table (assembly GRCh37/hg19) for the region chr6:28477754-33448354. The SNVs included were all records classified as 'single' or 'MNP' (Multiple Nucleotide Polymorphism). I considered each base of the MNPs as equivalent to one SNV. The short insertions include records classified as 'insertion' or 'in-del' with length $\leq$ 10 bases. Similarly, the short deletions include records classified as 'deletion' or 'in-del' with length $\leq$ 10 bp.

The 1KGP data was obtained from `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/`, again including all records within the region chr6:28477754-33448354.

A variant predicted by the Cactus MSA and present in the intersection of the

dbSNP and the 1KGP datasets was called *true positive*. In contrast, the *false negatives* were defined as variants previously reported for a given sample but not predicted by the Cactus MSA.

Except for the GRCh37 haplotypes whose sample-specific variants could not be located, the variants of each other sample were obtained from the corresponding UCSC Genome Browser's unpublished Personal Genome Variants tables, assembly GRCh37/hg19. The sample specific data from the 1KGP came from the March 2010 release `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_03/`, `ftp://ftp.jcvi.org/pub/data/huref/HuRef.InternalHuRef-NCBI.gff`, `http://yh.genomics.org.cn/do.downServlet?file=data/snps/yhsnp_add.gff`, using variants called using MAQ with the sequences from `http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP000239`.

2.5.6 Manual Analysis of False Positive SNVs

I manually analyzed C. Ref. false positives (not in dbSNP or the 1000 Genomes project data) SNVs with respect to GRCh37 using the UCSC Genome Browser [Fujita et al., 2011] and the Browser's unpublished MULTIZ [Miller et al., 2007] multiple sequence alignment of the GRCh37 haplotypes.

I separated the SNVs into five categories: 'confirmed', 'dbSNP bug', 'alignment disagreement', 'recurrent', and 'single'. SNVs were labeled 'confirmed' if they were observed in the MULTIZ MSA. SNVs that were not in dbSNP build 134 due to a bug in dbSNP were labeled 'dbSNP bug'; I reported this bug and it has now been fixed

in dbSNP build 135. SNVs were labeled 'alignment disagreement' when the MULTIZ alignment opted for indels and the Cactus MSA opted for substitutions. SNVs were 'recurrent' if there was no evidence from the MULTIZ MSA (often due to missing data or non-aligned sequences), but the SNVs were present in two or more samples from the Cactus MSA. SNVs were 'single' if there was no evidence from the MULTIZ MSA and the SNVs were present in only one sample.

2.5.7 Gene Mapping

To assess how genes mapped to C. Ref. in comparison to GRCh37, I aligned RefSeq [Pruitt et al., 2012]) transcripts and Genbank RNAs [Benson et al., 2012] to the GRCh37 assembly (excluding the alternative loci) and to a hybrid GRCh37/C. Ref. assembly, which was the GRCh37 assembly with the MHC region replaced by the C. Ref. sequence. The alignments were done using Blat [Kent, 2002], version 34x10. For each sequence the best alignments were chosen using the *pslCDnaFilter* program available in the UCSC Genome Browswer source code. Alignments with less than 95% base identity were discarded. In addition, two different coverage filterings, 90% and 95%, were applied, which respectively required that the alignments covered at least 90% and 95% of the RNA's/transcript's bases to be kept. RefSeq and Genbank RNAs were obtained from the UCSC Genome Browser tables, *refGene* and *all_mrna*, respectively, GRCh37 assembly.

HLA-DRB pseudogenes genes that were not in the RefSeq database (HLA-DRB2, HLA-DRB7, HLA-DRB8, HLA-DRB9) were mapped to C. Ref. using Blat with

identity $\geq$ 95%. Sequences for these genes were obtained from the UCSC Genome Browser, using the predicted mRNA sequences generated by Ensembl (HLA-DRB2: ENST00000419200, HLA-DRB7:ENST00000422566, HLA-DRB8:ENST00000436297, HLA-DRB9: ENST00000449413).

2.5.8 Short Read Mapping

To test C. Ref. as a target for mapping experiments, I constructed versions of C. Ref. excluding a held out 1KGP sample and then compared mappings made with BWA of the held out sample to GRCh37 and the held out C. Ref. The mappings include Illumina reads that were mapped to the GRCh37 MHC main region and Illumina unmapped reads (see Appendix Subsection 2.5.1 for the data source). Unpaired and paired reads were mapped using *bwa samse* (*-n 10000*) and *bwa sampe* (*-n 10000 -N 10000 -a 1000*), respectively. The *bwa* version was 0.5.9-r16.

2.5.9 Code Availability

The core code used to build C. Ref. and perform all the evaluations described can be retrieved from `https://github.com/benedictpaten/referenceScripts` and `https://github.com/ngannguyen/referenceViz`. Dependencies for this project are described in its documentation and can be retrieved from `https://github.com/benedictpaten/foo`, where *foo* is the name of the dependency. The version of the project and its dependencies used in this paper have all be tagged with the "referenceMHCProject" tag. For each project the "master" branch of the Git repository

should be used. Once installed all the evaluations described in the paper may be repeated using a simple *make* based pipeline.

2.5.10 Data Availability

The generated C. Ref sequence, the MSA and the assembled sequences used to build it can be downloaded from `http://compbio.soe.ucsc.edu/reconstruction/mhcReference/supplement.tar.bz2`.

Chapter 3

Comparative Assembly Hubs: Web Accessible Browsers for Comparative Genomics[1]

[1]This chapter expands on results from a recent paper that I worked on with Benedict Paten, Glenn Hickey and Brian Raney [Nguyen et al., 2014a].

3.1 Overview

Researchers now have access to large volumes of genome sequences for comparative analysis, some generated by the plethora of public sequencing projects and, increasingly, from individual efforts. It is not feasible nor is it most efficient for public genome browsers attempt to curate all this data. Instead, a wealth of powerful tools is emerging to empower users to create their own visualizations and browsers.

We develop a pipeline to easily generate collections of web accessible UCSC genome browsers interrelated by an alignment. The pipeline, named the comparative assembly hubs (CAH) pipeline, is intended to democratize UCSC comparative genomic browser resources, serving the broad and growing community of evolutionary genomicists and facilitating easy public sharing via the internet. Using the alignment, all annotations and the alignment itself can be symmetrically and efficiently viewed with reference to any genome in the collection. A new, intelligently scaled alignment display makes it simple to view all changes between the genomes at all levels of resolution, from substitutions to complex structural rearrangements, including duplications.

To demonstrate this work I create a comparative assembly hub for 57 *Escherichia coli* and 9 *Shigella* complete genomes. I report here comparative analyses of the *E. coli*/ *Shigella* genomes, including core genome, pan-genome and phylogenetic relationship analyses. The results recapitulate previous works and show that information can be gained and/or easily updated utilizing the CAH framework.

The *E. coli*/ *Shigella* genome hubs are now public hubs listed on the UCSC

Browser Public hubs webpage. The hubs contain many important annotation tracks, with the annotations obtained from both external databases (genes, non-coding RNAs, genomic islands, antibiotic resistance, pathogenicity) and automatic computations (conservation, alignability, repetitive elements, GC content). This is the first and only *E. coli* comparative browser resource that has base level resolution and in which the core genome and pan genome are incorporated and all genomes are interconnected by one consistent multiple sequence alignment (MSA).

3.2 Introduction

3.2.1 Motivation

Visualization is key to understanding functional and comparative genomic information. Genome browsers are therefore critical to the study of biology, providing accessible resources for displaying annotations and alignments. The UCSC Genome Browser [Karolchik et al., 2014] is one of the most popular, but creating a browser within it previously required significant resources, since it was necessary to create a mirror site to separately host the browser or to work with the staff of the genome browser to create a browser within the main site; this is a process that does not naturally scale due to limited resources.

Assembly hubs [Karolchik et al., 2014], which build on the successful track hub model [Raney et al., 2013], make it easy to generate an individual UCSC browser simply by hosting the data in the form of flat-files on any publicly addressable URL. This frees

up users from having to install and configure the substantial browser code base on their machines and using user hosting simplifies the updating process.

Increasingly, users, with access to low cost sequencing technology, and the wealth of genomes available, will want to be able to create not just a single custom browser, but sets of genome browsers. Partly this is being driven by large-scale projects [10k-Community-of Scientists, 2009, i5K Consortium, 2013], and partly by the growth in individual lab sequencing. This work is intended to meet this growing need. It extends assembly hubs to allow users to quickly create "comparative assembly hubs", a framework of multiple genome browsers and annotations interrelated by an alignment. Included in this development is a series of novel features intended to improve visualization, exploration and community sharing of novel comparative genomics data.

3.2.2 Challenges in Multiple Genome Alignment Visualizations

Displaying multiple genome alignments is extremely challenging due to both the high dimensionality and volume of the underlying data (see [Nielsen et al., 2010] for a review). Extensive efforts have been invested in addressing these challenges. There are three main ways to visualize multiple genome alignments: dot plots, circle plots and linear, row-oriented representations. For each method, there are a vast number of softwares available, examples include DAGChainer [Haas et al., 2004], VISTA-Dot [Mayor et al., 2000], MUMmer [S et al., 2004], GenomeMatcher [Ohtsubo, 2008] and MEDEA [Jen et al., 2009] for dot plots, Circos [Krzywinski et al., 2009], GenomeRing [Herbig et al., 2012],

CGView [Grant et al., 2012], GenomeViz [Ghai et al., 2004] for circle plots, and IGV [Waterhouse et al., 2009], Jalview [Thorvaldsdóttir et al., 2013], GenomeView [Abeel et al., 2012], UCSC [Karolchik et al., 2014] and VISTA [Mayor et al., 2000] browser conservation tracks for linear, row-oriented representations.

Each of these visualization methods has its clear benefits and weaknesses. Dots plots, having two dimensions, provide equivalently powerful representations of two genomes in one graphic, however, they are pairwise and therefore unsuitable for the display of multiple genome alignments. Circular genomes plots, being organized around a circle, are typically less visually cluttered than are the linear representations for globally viewing genomic rearrangements, but are less useful for viewing local multiple alignments, with a consideration at the gene or base level. In addition, they do not scale well in displaying structural variations to the number of genomes involved.

Linear representations have the advantage that they fit neatly with the genome browser displays and tracks, and are flexible, in that they work reasonably at multiple levels of resolution. However, similar to circular representations, they are limited in displaying structural variations. Typically in an MSA display, each genome is represented by a row (track) and homologous segments between two genomes (rows) are color coded and connected by lines from one row to the other (Figure 3.1). With this representation, structural rearrangements are visible but the display gets crowded quickly as the complexity increases and becomes incomprehensive. Moreover, the display is dependent on the transitiveness of the rows. Only genomes that are in consecutive rows have their homologous segments connected. As such, the comparisons of genomes in non-consecutive

Figure 3.1: An example MSA linear display obtained from Mauve User Guide [Darling et al., 2004] (`http://asap.ahabs.wisc.edu/mauve-aligner/mauve-user-guide/mauve-screenshots.html`). The accompanied caption is as followed: The screenshot was taken with Mauve 2.0 visualizing an alignment of nine Yersinia genomes. The screenshot visualizes the global rearrangement structure of the chromosomes.Each genome is laid out horizontally and homologous segments are shown as colored blocks that are connected across genomes. Blocks that are shifted downward in any genome represent segments that are inverted relative to the reference genome (Yersinia pestis KIM). Clicking in the display will vertically align the homologous segment in each genome.

rows rely on intermediate comparisons with genomes in the intermediate rows. The indirect comparisons unavoidably result in the dropping of information.

The popular alternative for displaying structural variations is the UCSC genome browser *chains* and *nets* representation [Kent et al., 2003] (Figure 3.2). A *chain* is a sequence of non-overlapping gapless alignment blocks, in which both target and query coordinates are either increasing or flat. The *chains* display consists of all possible chains within the alignment and reflects the many-to-many alignment relationship,

i.e a query position can be aligned to many target positions and a target position can be aligned to many query positions. A *net* is a hierarchial display of chains, with the highest-scoring non-overlapping *chains* at the top level and lower scoring chains at the lower levels, filling in the higher levels' gaps where possible. The *nets* display reflects the one-to-many relationship: a query position can be aligned to many target positions but each target position can only be aligned to at most one query position. The *chains* and *nets* representation allows users to deduce rearrangement and duplication events. With query locations shown by mousing-over the display, there are no lines that cross multiple rows and line congestion ceases to be a problem. The disadvantage of the *chains* and *nets* representation is the loss of visual cues seen in the overall image of rearrangement events that such lines provide.

In addition to limitations in visualizing possible structural variations, the key issue for current implementations of MSA displays is that they are reference-dependent. Restricting visualization to any single reference can restrict the viewer from observing visualizations of regions in other genomes that do not map to the reference, or regions that are mapped to an alternate rearrangement with respect to a reference. Furthermore, separate MSA constructions are often required if different references are considered. Such additional constructions may be computationally expensive and may lead to inconsistencies.

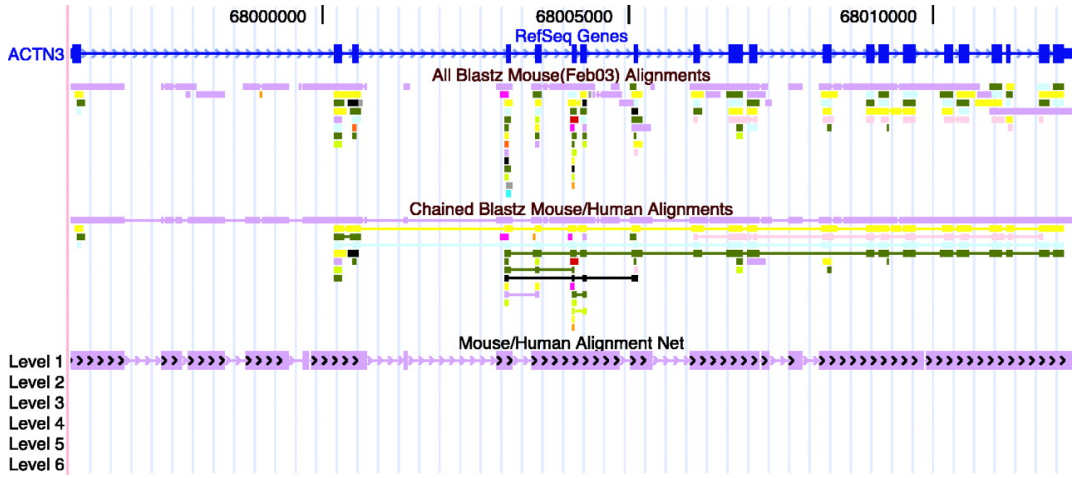

Figure 3.2: An example of the UCSC chain and net display. Figure is Figure 1 extracted from Kent *et. al* [Kent et al., 2003]. The accompanied caption is: "Mouse/human alignments at Actinin α-3 before and after chaining and netting, as displayed at the genome browser at http://genome.ucsc.edu. The RefSeq genes track shows the exon/intron structure of this human gene, which has an ortholog as well as several paralogs and pseudogenes in the mouse. The all BLASTZ Mouse track shows BLASTZ alignments colored by mouse chromosome. The orthologous gene is on mouse chromosome 19, which is colored purple. Although BLASTZ finds the homology in a very sensitive manner, it is fragmented. The chained BLASTZ track shows the alignments after chaining. The chaining links related fragments. The orthologous genes and paralogs are each in a single piece. The chaining also merges some redundant alignments and eliminates a few very low-scoring isolated alignments. The Mouse/Human Alignment Net track is designed to show only the orthologous alignments. In this case, there has been no rearrangement other than moderate-sized insertions and deletions, so the net track is quite simple. Clicking on a chain or net track allows the user to open a new browser on the corresponding region in the other species."

3.2.3 Innovative Features of Comparative Assembly Hub

The CAH framework introduces a new linear display representation - the *snake* track; I believe this track style to be the first linear representation which gives a representation for all variations, including structural rearrangements, duplications, substitutions, insertions and deletions in a single, visually appealing, interactive visualization. The *snake* track provides visual cues to possible structural variation events while avoid-

ing lines that cross multiple genomes. In addition to giving a more complete picture of the genome of interest, the *snake* track contains a series of innovative features. Because the *snake* employs a novel algorithm which generates procedural levels of detail, the user is able to zoom in or out of every level of resolution, from complete chromosomes to individual residues. Being based upon a symmetric, reference independent alignment format [Hickey et al., 2013], for the first time in an alignment view using the UCSC Browser, *snake* tracks are viewable between any set of genomes in the hub and from any chosen genome, as all genomes in a hub have a generated reference browser. Additionally, to overcome the limitations of viewing the alignment from the perspective of single reference genome, along with each generated comparative assembly hub, a pan-genome reference browser is also given via the algorithm reported in previous chapter [Nguyen et al., 2014b].

The comparative assembly hubs are also novel from a genome browser perspective: The underlying alignment may be used, by a process of "lift-over" (coordinate conversion between assemblies) [Zhu et al., 2007], to automatically project annotations to any genome in the hub, even if the annotations were originally mapped to just one genome. Previously lift-over was employed on a case-by-case basis within the UCSC genome browser to project tracks between assemblies, for example, when moving to an updated assembly. Here it is a default, integral feature, making it easy to view putative genes and functional annotations on any novel genomes by a process of translation using the underlying alignment. To my knowledge, the CAH framework is also the first web-based genome browser that permits easy public sharing of comparative data with-

out hosting or requiring other users to download data - only a web browser is required. Finally, separately to the novel features introduced, these user generated browsers are integrated with many of the existing (powerful) tools of the UCSC browser such as the Table Browser.

3.2.4 *E. coli* Comparative Genomic Resources

E. coli is one of the most studied organisms and consequently has been one of the most sequenced bacterial species. At the time of writing, *E. coli* has the second highest number of complete sequence genomes publicly available among all bacteria (after *Salmonella enterica*, source: `http://www.ebi.ac.uk/genomes/bacteria.html`). *E. coli* contains substantial intra-species genomic diversity, which allows for its high versatility including various pathotypes, different antibiotic resistances as well as different lifestyles (see a review at [Leimbach et al., 2013]). Comparative genomic analyses of multiple strains of *E. coli* have proven useful in understanding the molecular basis of their phenotypic differences [Leimbach et al., 2013, Ogura et al., 2009, Lukjancenko et al., 2010], and these analyses have been beneficial to promote practical applications such as diagnostic and antibiotic developments for bacterial infectious diseases [Didelot et al., 2012, Rasko et al., 2011, Mellmann et al., 2011, Rohde et al., 2011].

E. coli comparative genomics paints a familiar picture of the current challenges faced by the bioinformatics driven scientific community and is a great illustration of the pressing needs for a robust comparative genomics framework. The rapid speed of data

generation has challenged scientists to provide the necessary well-tuned data analysis, as in the case of *E. coli* comparative genomics. As new data becomes available, not only are novel analyses required to answer new questions which will inevitably arise, but previous analyses will also most likely need to be redone to update existing knowledge. While the prior step is usually accomplished and published in the paper accompanied each sequencing effort, the latter typically falls to the side. This is a common case in current *E. coli* research: each individual sequencing effort usually targets only specific questions of interest. Meanwhile, the accumulation of many individual efforts results in a tremendous amount of data, which are much valuable for comparative genomics opportunities. Unfortunately, this source of information remain untapped and dormant because considerable efforts are required to carry out the comparative genomics analyses.

Some examples illustrating the importance of having a constant update of existing knowledge are the analyses which surround the interplay between the core genome, pan genome, and phylogenetics for *E. coli*. Over the years, numerous studies have been written on these topics, yet the results have a tendency to be inconsistent from one study to another (see Sections 3.3.3.2, 3.3.3.3 and 3.3.4). The inconsistencies are mostly due to two reasons: different data sets and different methodologies (e.g gene-based versus genomic-based approaches). Both reasons may be traced to the differences in both the availability and nonavailability of sequencing data.

I now show the CAH framework offers a procedural, automated solution for updating such analyses with minimal efforts which helps research to keep pace with the current deluge of data.

3.2.4.1 *E. coli* Comparative Genomic Visualization Resources

Despite *E. coli*'s popularity as a model organism, there is currently no central resource for visualizations of *E. coli* comparative genomics. Ensembl (`http://bacteria.ensembl.org/`) offers individual browser for each *E. coli* genome yet does not provide comparisons with other genomes. Similarly, the UCSC Archaeal Genome Browser `http://archaea.ucsc.edu/genomes/bacteria/` contains browsers for a small number of individual strains (eight *E. coli* and one *Shigella*) but no comparative information is provided. Also, the UCSC Archaeal Genome Browser is not kept up to date. The Microbial Genome Viewer (MGV) [Kerkhoven et al., 2004] does offer individual browsers equipped with limited comparative functions, such as links to orthologous genes and gene-context of COGs (cluster of orthologous groups), but these browsers lack MSA views (Figure 3.3). Various *E. coli* comparative genomic studies have provided static visualizations (typically in form of figures) of global MSAs for *E. coli* sets specific to each studies with limited annotations [Lukjancenko et al., 2010, Rasko et al., 2011, Grant et al., 2012]. And, these static visualizations do not display structural variations (Figure 3.4).

Using the CAH pipeline, I built the *E. coli* comparative assembly hubs for all *E. coli* and *Shigella spp.* complete genomes that were publicly available at the time of writing. The resulting hubs contain one browser for each input genome plus one browser for the pan-genome and each browser is accompanied by alignment and annotation tracks. This resource provides researchers with the ability to explore the

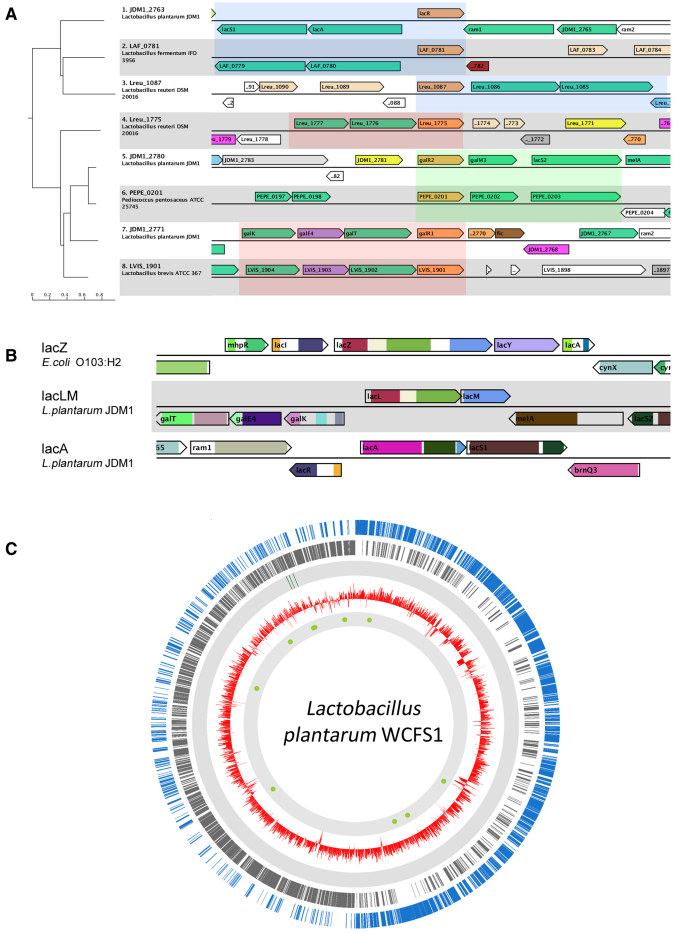

Figure 3.3: An example of MGV displays. Figure is Figure 2 extracted from Overmars *et al.* [Overmars et al., 2013]. A shorten version of the accompanied caption is: "EbgR-like transcription factors in *L. plantarum* and other *lactobacilli*. A) MGcV visualization of a phylogenetic tree of EbgR-type regulators in some Lactobacilli. B) Comparative context map of the beta-galactosidase encoding genes lacZ (E. coli), lacLM (L. plantarum) and lacA (L. plantarum). C) A circular genome map of L. plantarum in which the ORFs on the plus strand (blue), on the minus stand (grey), the locations of regulator encoding genes lacR, rafR and galR (green), the GC% (red) and putative binding sites (similarity to motif >90%; represented by the green dots) are included."

data and facilitates data analysis tasks. More importantly, such resource can be easily generated or updated as new data become available.

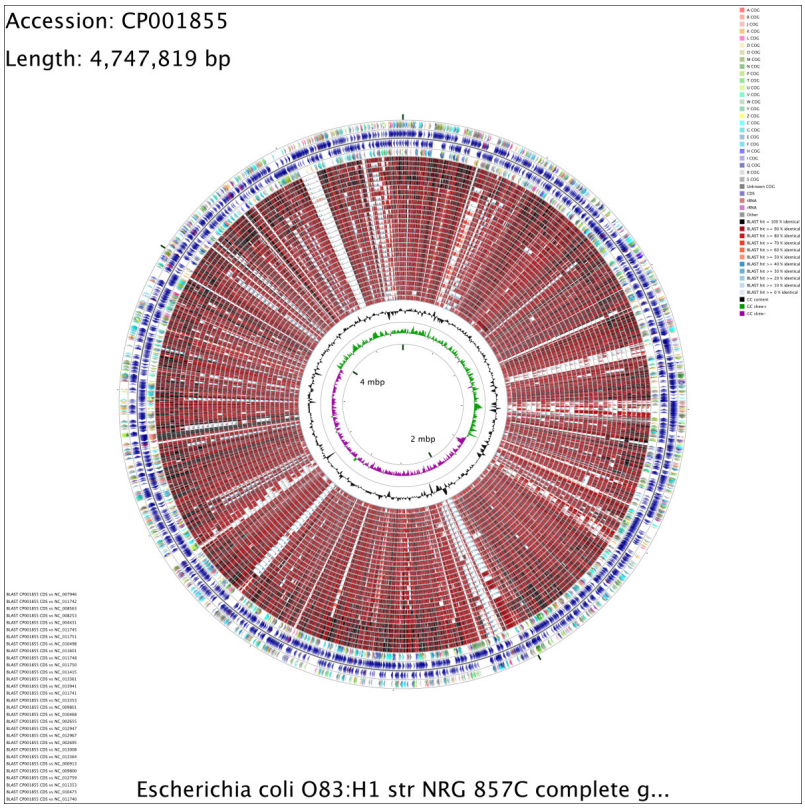

Figure 3.4: An example of CGView display. Figure is Figure 1 extracted from Grant *et al.* [Grant et al., 2012]. The shorten accompanied caption is: "CCT map comparing an *E. coli* reference sequence to other *E. coli* genomes. Starting from the outermost ring the feature rings depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features; 3. Reverse strand sequence features; 4. COG functional categories for reverse strand coding sequences. The next 30 rings show regions of sequence similarity detected by BLAST comparisons conducted between CDS translations from the reference genome and 30 *E. coli* comparison genomes. The last two rings display the GC content and GC skew."

3.3 Results

In the following section, I first describe the software pipeline for building comparative assembly hubs, following which, I describe the hubs by example using an alignment of *E. coli* genomes. Last, I illustrate an application of the comparative genomics framework with the *E. coli* core genome, pan genome and phylogenetic analyses.

3.3.1 The Comparative Assembly Hubs (CAH) Pipeline

The open-source pipeline, composed three modular components, has been written to work on UNIX distributions. The Cactus alignment program [Paten et al., 2011b], its first component, takes as input the user's set of genome sequences and outputs a genome multiple sequence alignment in Hierarchical Alignment (HAL) format [Hickey et al., 2013] in return. The second component is HAL tools, to which there is a series of command line tools and C/C++ APIs for manipulating HAL files and building comparative assembly hubs. The final component is the *snake* track display, which is now part of the UCSC Genome Browser code base [Karolchik et al., 2014], and which provides visualization of alignments directly from HAL files.

The pipeline is run in three steps (see Methods). Firstly, either Cactus is run to generate the HAL alignment file directly, or a MAF file, generated separately by an aligner such as Multiz [Miller et al., 2007], is converted into a HAL file by way of the maf2hal tool (in the HAL tools package). Secondly, the hal2AssemblyHub script (in the HAL tools package) builds the comparative assembly hub using the HAL file and any set of annotation files provided, either in *bed* or *wig* format (`http://genome.ucsc.edu/FAQ/FAQformat.html`). This script takes care of converting the base annotation files into the display scaleable *bigBed* and *bigWig* formats, and optionally translates these annotations using a process of alignment lift-over [Zhu et al., 2007], to all the other genomes. In addition to the provided annotation tracks, for each genome, the script

is able to compute a number of other useful annotation tracks, such as the gap track, alignability track, GC-content track, and evolutionary conservation track (using the phyloP program [Cooper et al., 2005, Siepel et al., 2006]). Finally, a directory is created containing the necessary files, using compressed formats for minimal space usage. In the final step, the location of the 'hub.txt' file, addressable as a public URL, is pasted into the UCSC browser hub page to view the browsers.

The pipeline builds one browser for each input genome, and, in addition, any ancestral or pan-genomes that were imputed by Cactus (if used) during the alignment process [Nguyen et al., 2014b].

3.3.2 *E. coli* Comparative Assembly Hub

To demonstrate this work I use a collection of 57 *E. coli* and 9 *Shigella* complete genomes and various accompanying annotations, including repetitive elements, genomic islands, pathogenic genes, non-coding RNAs and antibiotic resistance genes. A total of 67 browsers are built, one browser for each input genome and one pan-genome browser. Each browser consists of a set of annotation tracks: one for each input annotation file, one *snake* track for each of the other 66 genome browsers, and three additional tracks that are computed automatically by the pipeline, including a conservation track, a GC-content track, and an alignability track.

As an illustration of a hub browser, Figure 3.5 displays a region of one of the *E. coli* reference genomes, K12_MG1655, with *snake* tracks, *bed* and *wig* annotations and lifted-over *bed* annotations. The top most tracks are K12_MG1655 annotations, includ-

ing Alignability (number of genomes mapped to each position), GC%, Antibiotic Resistance Database (ARDB), Genes, Genomic Islands (GI) and non-coding RNAs (rRNA and tRNA). Below these are *snake* tracks, showing the alignment of the genome to a subset of the other genomes and lifted-over ncRNA annotation track (track K12_W3110 RNA) of *E. coli* K12_W3110.

3.3.2.1 The *Snake* Track

Each *snake* track shows the relationship between the chosen browser genome, termed the reference (genome), and another genome, termed the query (genome). The *snake* display is capable of showing all possible types of structural rearrangement. Stacked together, *snake* tracks allow flexible view of the multiple genomes.

In *full* display mode (*snake* tracks in Figure 3.5), it can be decomposed into two primitive drawing elements, segments, which are the colored rectangles, and adjacencies, which are the lines connecting the segments. Segments represent subsequences of the query genome aligned to the given portion of the reference genome. Adjacencies represent the covalent bonds between the aligned subsequences of the query genome. Segments can be configured to be colored by chromosome, strand (as shown) or kept a single color. Layout of the segments is described in the methods.

Red tick-marks within segments represent base substitutions as compared with the reference and by default (user configurable), are displayed up to a 50 kilo-base resolution (Figure 3.5(b-c)). Zoomed in to the base-level resolution, such substitutions are labeled by the non-reference base (Figure 3.5(d)). An insertion in the reference

relative to the query creates a gap between abutting segment sides that is connected by an adjacency. An insertion in the query relative to the reference is represented by an orange tick mark that splits a segment at the location where the extra bases may be inserted, or by coloring an adjacency orange, indicating that there are unaligned bases between the two segment ends it connects.

More complex structural rearrangements create adjacencies that connect the sides of non-abutting segments in a natural fashion. An example is shown in Figure 3.5a, visualizing a known, large inversion in the closely related strain K12_W3110 with respect to the reference strain K12_MG1655 [Hill and Harnish, 1981, Hayashi et al., 2006]. The inversion is flanked by the ribosomal RNA operons *rrnD* and *rrnE* (RNA tracks in green), and is the result of homologous recombination between them. Operon *rrnD*, consisting of *rrsD*, *ileU*, *alaU*, *rrlD*, *rrfD*, *thrV*, *rrfF*, and operon *rrnE*, consisting of *rrsE*, *gltV*, *rrlE*, *rrfE*, are homologous segments with opposite directions, as can be seen by zooming in (Figure 3.5(b, c)). Also shown in Figure 3.5a are two smaller inversions in KO11FL [Turner et al., 2012] and O26 H111 1368 [Ogura et al., 2009] and a relatively much smaller inversion in HS.

Duplications within the query genome create extra segments that overlap along the reference genome axis. For example, Figure 3.6 shows a tandem repeat region of *E. coli* KO11FL_162099 displayed along the genome of *E. coli* KO11FL_52593. *E. coli* 52593 was engineered by chromosomal insertion of the *Zymomonas mobilis pdc*, *adhB* and *cat* genes into the parental strain *E. coli* W for ethanol production purpose [Ohta et al., 1991]; 162099 is a derivative of 52593 (after 20 years of serial transfers) and

contains 20 tandem copies of the inserted *pdc-adhB-cat* genes[Turner et al., 2012]). To show regions where the query segments align to multiple locations within the reference, at the top of each *snake* track there are colored coded sets of lines along the reference genome axis that indicate self homologies (intervals of the reference genome that align to other intervals of the reference genome), and to maintain the semantics of the *snake*, query segments that align to these regions are aligned arbitrarily to just one copy of the reference (Figure 3.6).

There is a large deletion in W as this parent strain does not have the inserted region. The figure shows 20 tandem copies of a 10kb unit spanning genes (*pflA*, *pflB-L*, *cat*, *adhB*, *pdc* and *pdfB-S*) in KO11FL_162099. KO11FL_52593 has two copies of genes *pflA*, *pflB-L*, and *pflB-S*, i.e. self-alignments, as shown in the figure by the colored lines on top of each query *snake* track.

The above examples demonstrate that the browser linear representation with *full*-mode *snake* tracks, annotation tracks and the ability to zoom in and out to any resolution provide an intuitive way of viewing structural variations and examining and exploring biological information.

The *pack* display option can be used to display a *snake* track in more limited vertical space. It eliminates the adjacencies from the display and forces the segments onto as few rows as possible, given the constraint of still showing duplications in the query sequence (e.g. track W_162099 of Figure 3.6). The *dense* display further eliminates these duplications so that a *snake* track is compactly represented along just one row (e.g. tracks SE11 and IAI1 of Figure 3.6). The *dense* display is equivalent to

the popular view generated by many existing comparative genomic visualization tools [Grant et al., 2012].

Clicking on a segment translates the browser view from the present reference genome to the corresponding region in the query genome, making it simple to navigate between references, all of which have equivalent displays. This symmetry frees the user from investigating the alignment from just one perspective. Various mouseovers are implemented to display the sizes of display elements, and the *snakes* and annotations can be reordered by dragging them.

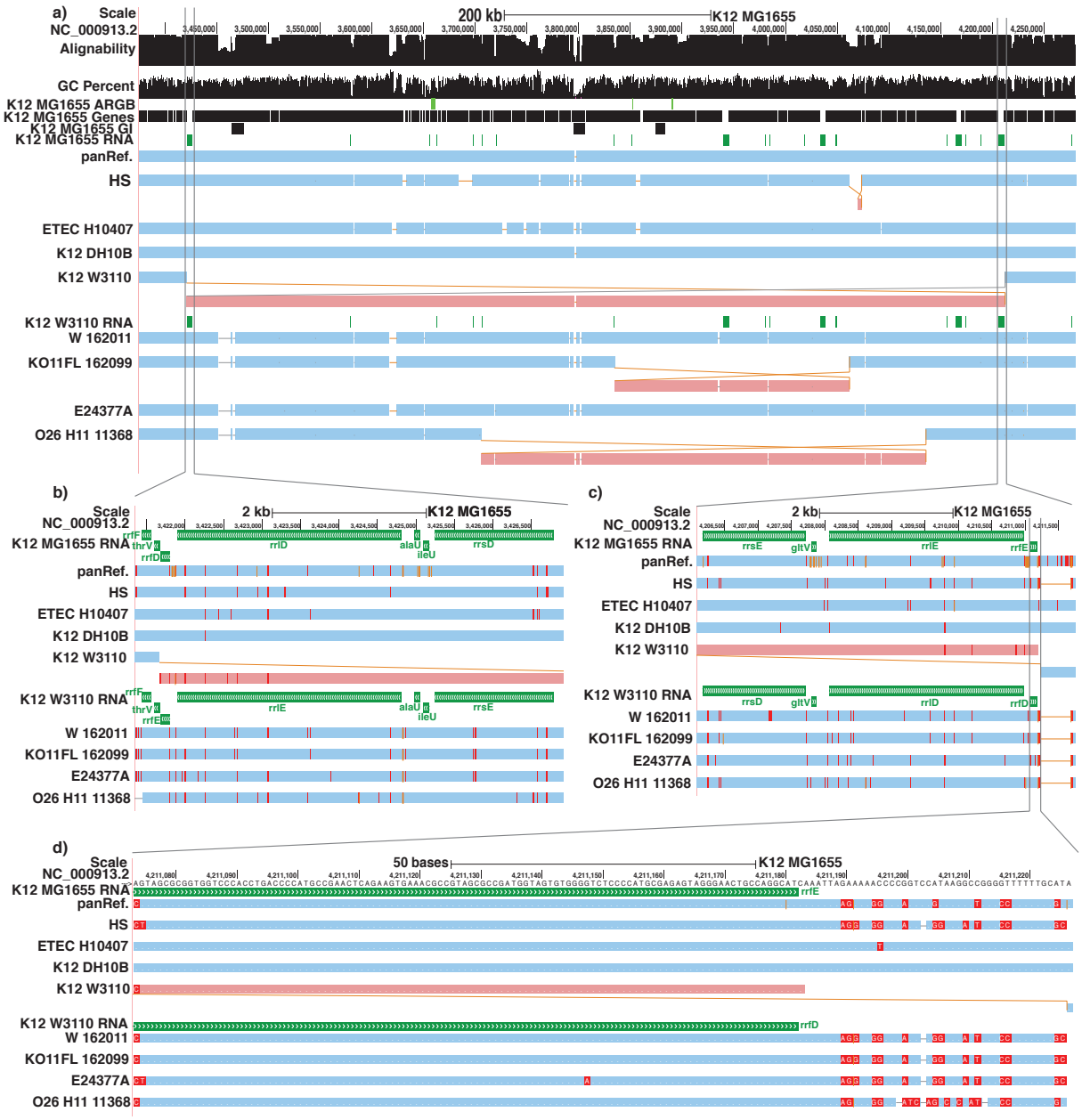
a)
Scale
NC_000913.2
Alignability
GC Percent
K12 MG1655 ARGB
K12 MG1655 Genes
K12 MG1655 GI
K12 MG1655 RNA
panRef.
HS
ETEC H10407
K12 DH10B
K12 W3110
K12 W3110 RNA
W 162011
KO11FL 162099
E24377A
O26 H11 11368
200 kb
K12 MG1655
b)
2 kb
rrfF
thrV
rrfD
rrlD
alaU
ileU
rrsD
rrlE
rrsE
c)
rrsE
gltV
rrlE
rrfE
rrsD
rrlD
rrfD
d)
50 bases
rrfE
rrfD

Figure 3.5

Figure 3.5: An example view of the *E. coli* comparative assembly hub illustrating the CAH pipeline's innovative features, which include: a novel MSA visualization that is able of displaying all types of variation, multiple levels of resolution, and various annotation tracks generated from input annotations, built-in computations, or lifting-over (mapping) of input annotations. The *E. coli* strain K12_MG1655 is the reference browser. Shown is a subset of genomes of group A (from HS to K12 W3110) and B1 (the rest). The top browser screenshot (**a**) shows a 900kb region with a known, large inversion (light red) in the closely related strain K12_W3110. The inversion is flanked by homologous (with opposite orientations) ribosomal RNA operons *rrnD* and *rrnE* [Hill and Harnish, 1981, Hayashi et al., 2006], and is the result of homologous recombination between them. Besides *snake* tracks, from top to bottom: Tracks "Alignability" and "GC Percent" are computed by the pipeline using the input sequence information. Tracks "K12_MG1655 ARGB", "K12_MG1655 Genes", "K12_MG1655 GI" and "K12_MG1655 RNA" are generated from K12_MG1655 input annotations. Track "panRef." is the computed pan-genome of *E. coli* and *Shigella*. Track "K12_W3110 RNA" is K12_W3110 RNA annotation mapped onto K12_MG1655. (**b-c**) Zoom-in of the K12_W3110 inversion left and right boundaries, respectively, showing operon *rrnE* of K12_W3110 ('K12_W3110 RNA' track, in green, which is K12_W3110 ncRNA annotation track lifted-over to K12_MG1655) aligned to operon *rrnD* of K12_MG1655 ('K12_MG1655 RNA' track, also in green) on the left and operon *rrnD* of K12_W3110 aligned to operon *rrnE* of K12_MG1655 on the right. Further zoomed in (**d**), SNPs and query insertions are visible. The text on the screenshots was adjusted for better readability.

3.3.2.2 Procedural Levels of Alignment Detail

The different levels of detail displayed in Figure 3.5(a-d) show the alignment at megabase, kilobase and base level. To achieve this in a web browser, serving data across the internet (generally still a relatively slow and high latency connection), we developed a novel solution. For instance, a chromosome is typically decomposed into millions of segments in a HAL graph. In Nguyen *et al.* 2014 [Nguyen et al., 2014a], we describe detailed methods of pre-generating interpolated HAL graphs that store only as much information as visible on the screen at different zoom levels, and demonstrate that we achieve constant load times for webpages at all levels of resolution using the method.

3.3.2.3 UCSC browser integration

A key benefit of comparative assembly hubs is their integration with the popular UCSC browser and the tools it provides. For example, export of subregions of the alignment and track intersections can be made via the UCSC table browser [Karolchik et al., 2004], and using user sessions, individual browser displays can be shared (see Methods for links to examples).

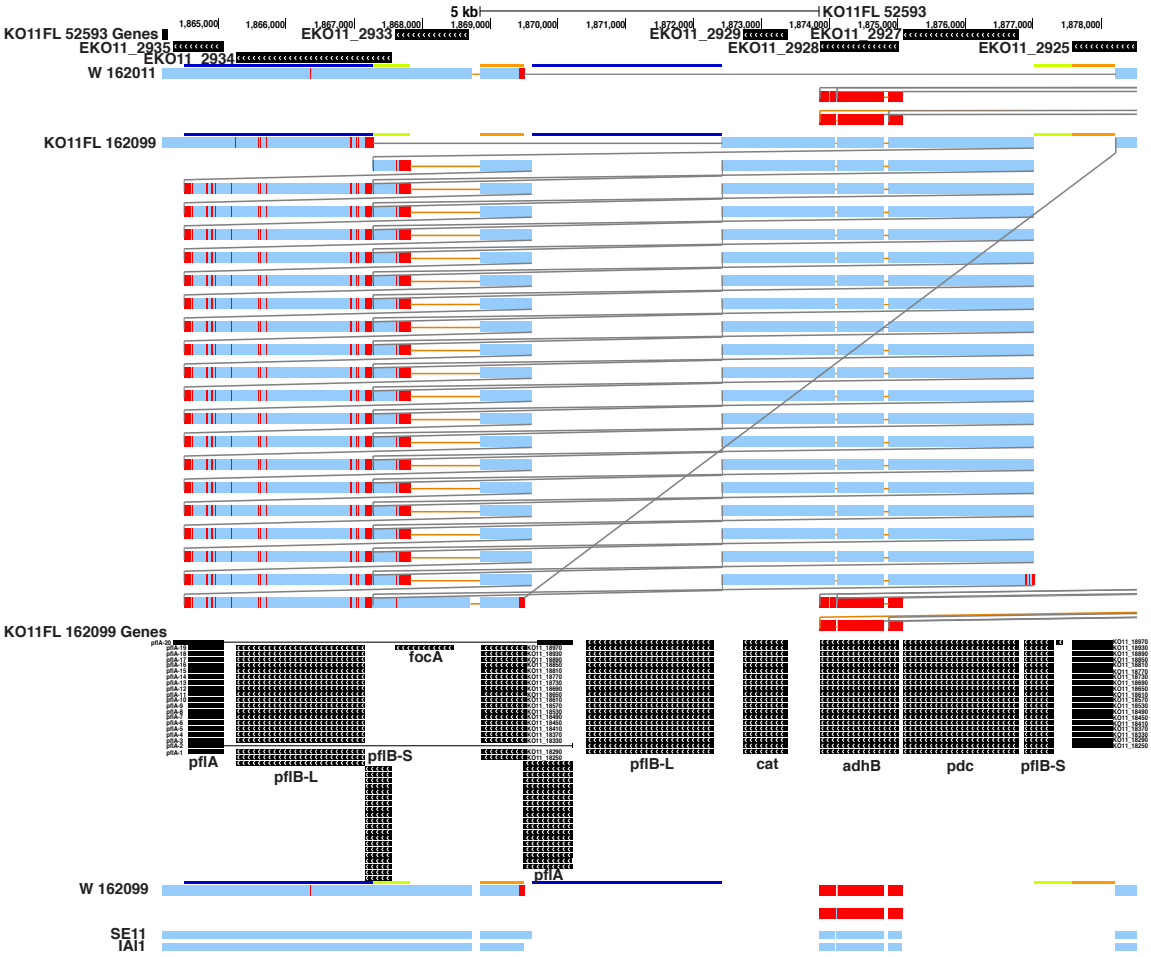

Figure 3.6: A browser screenshot demonstrates the *snake* visualization of duplications and the different display modes of the *snake* track, including *full* mode (tracks "W 162011" and "KO11FL 162099"), *pack* mode (track "W 162099") and *squish* mode (tracks "SE11" and "IAI1"). The browser shows the *pdc-adhB-cat* tandem repeat region of *E. coli* KO11FL_162099 [Turner et al., 2012] displayed along the genome of *E. coli* KO11FL_52593. The colored horizontal bars on top of each *snake* track indicate duplications in KO11FL_52593 (two copies of each gene *pflA* (dark blue), *pflB-L* (dark blue), *pflB-S* (light green) and KO11_18*** (orange)). There is a large deletion in the parental strain W_162011 as this strain does not contain the *pdc-adhB-cat* insert. Following the *snake* track of KO11FL_162099, there are 20 copies of (*pflA*, *pflB-L*, *cat*, *adhB*, *pdc*, *pdfB-S* and *KO11-18****, where "*KO11-18****" may be replaced by specific IDs shown in the figure and listed in Appendix Section B.1). Note that since KO11FL 52593 has two copies of *pflA*, *pflB-L*, *pflB-S* and *KO11_18****, the display arbitrarily picks one copy of each to map corresponding KO11FL 162099 orthologous genes to. The text on the screenshot was adjusted for better readability.

3.3.2.4 Managing Alignments and Lifted Annotations

A unique feature of comparative assembly hubs is that all annotations can be viewed from any genome through the alignment. To make managing the large number of possible *snake* and lifted-over annotations easy for each browser, a central configuration page is provided that uses a grid layout as its basis (Figure 3.7). This configuration page layout is adapted from the UCSC Encode Browser [Rosenbloom et al., 2010], where, instead of using it to select tracks from combinations of cell-line and assay types, it is instead used to select from the available combination of genomes and (lifted-over) tracks, laid out phylogenetically (if a tree is provided). As with the Encode Browser, the grid is sufficiently compact to display hundreds of tracks on one page, without moving to a hierarchical layout that would involve greater user navigation.

3.3.2.5 Clade-exclusive Genomic Regions

The hub browser display with the default phylogenetically ordering of the *snake* tracks allows for easy identification of clade-specific or clade-dominant variations. An example is the small deletions (relative to K12_MG1655) near position 3,450,000 that is observed in genomes of all other groups except of the A and B2 (except for genome O127_H6_E234869) groups (Figure 3.5a). Similar clade(s)-specific variations can be easily spotted out when looking at the browser.

In addition, if a phylogenetic tree is specified, the pipeline has an option to compute genomic regions that are specific to each clade of the tree. A clade can contain only an individual leaf node, in which case the regions computed are leaf-specific or

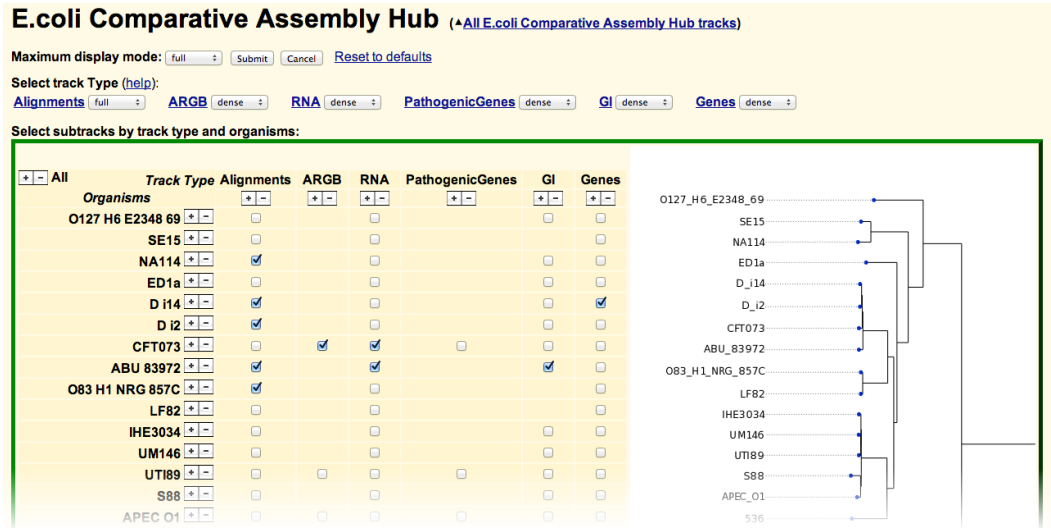

Figure 3.7: An example portion of a comparative assembly hub configuration webpage, each browser in the hub has its own such equivalent configuration page. Using the grid layout (rows represent the genomes, columns the track types) alignments and annotations can be selected regardless of which genome they were originally described upon. The inset phylogenetic tree is generated automatically by the CAH pipeline. The track controls above the grid allow quick, overall configuration. Fine grained track controls (not shown) are provided at the bottom of the page, in a list of the selected tracks. Tracks not potentially lifted through the alignment, such as GC content, repeat masking and conservation tracks are configured using the standard drop down menus on each browser page (not shown).

genome-specific regions, or a group of leaf nodes with a common ancestor. The result is one annotation track for each node (both ancestral and leaf) in the tree, displaying the corresponding computed regions. The minimum number of in-group and the maximum number of out-group genomes can be adjusted by the users.

For *E. coli*, which has substantial horizontal gene transfers and constant emergence of new (pathogenic, or disease-causing) strains, such tracks are useful for the identification of new (pathogenic) genetic materials that the new strains acquired. Figure 3.8 shows a region that is specific to two pathogenic groups of *E. coli* strains:

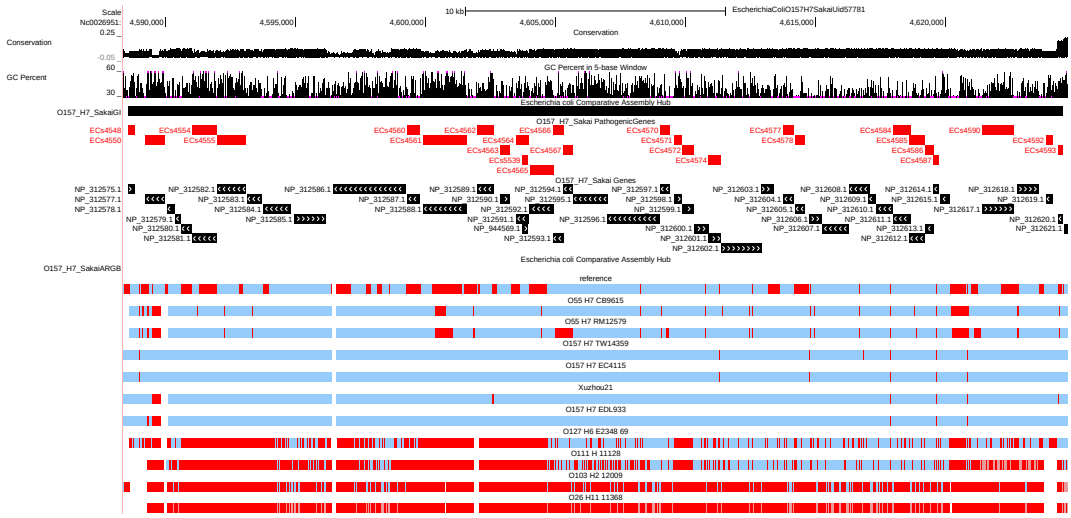

Figure 3.8: An example of an EHEC/EPEC-specific region displayed along the genome of *E. coli* O157H7 Sakai. The browser shows all the genomes that have bases aligned to the region. Except the two O55 H7 strains and the O127 H6 E2348 69 strain which are EPEC, all other strains are EHEC . This region corresponds to the LEE pathogenetic genomic island that contains many pathogenic genes [Ogura et al., 2009].

EHEC and EPEC. This region corresponds to the LEE pathogenetic genomic island that contains many pathogenic genes [Ogura et al., 2009], as shown by the 'O157_H7_Sakai PathogenicGenes' track. Hypothetically, this region is transferred to a commensal *E. coli* strain and results in a new pathogenic strain. The new strain, having most of its genome identical to the parental commensal strain, will likely be phylogenetically grouped close to that strain. Nevertheless, its pathogenicity can be quickly identified as one can observe that it aligns to an EHEC/EPEC-specific region, which is accompanied by the pathogenic gene annotations.

3.3.3 *E. coli* Comparative Genomics Analyses

In this section, I first show assessments of the *E. coli*/*Shigella* MSA. I then demonstrate that the CAH pipeline offers a procedural, automated solution for updating comparative analyses with minimal efforts by recapitulating the pan-genomic, core genomic and phylogenetic analyses of *E. coli*.

3.3.3.1 Assessing The Genome Alignment

In a comparative assembly hub all genome comparison and lifted track displays are driven consistently by a single underlying genome alignment and summaries of it. This provides great consistency and is likely to lead to less confusion when interpreting the visualization. For example, high-level views can always be drilled down to reach the original base-level alignment, and lifted annotations can be easily interrogated via a *snake* track that shows the actual alignment used to do the lift-over. Certainly, the accuracy of the genome alignment is always key and must be kept in mind. Alignments for assembly hubs can be created by any aligner that can export a MAF file (a simple flat-file format), however currently the most general solution is the Cactus alignment program.

Cactus was used to align the *E. coli* genomes reported here. It has been tested and has been proved to be highly accurate elsewhere [Paten et al., 2011b]. Here, as an extra quality assurance step for the constructed hubs, I assessed the *E. coli* alignment to see how well orthologous genes of input genomes were aligned to each other. Gene annotations of each input genome obtained from NCBI and BLAT [Kent, 2002] pairwise

alignments were used to group genes into orthologous groups (see Methods). For each pair of genomes I computed the number of orthologous coding gene families that were aligned in the Cactus alignment. On average each genome contains 4751 gene families and shares 3374 gene families with another genome (Table 3.3.3.1). Across all possible pairs of genomes, the vast majority (99%, 3333/3374) of each pair's orthologous groups were aligned to each other in the multiple alignment.

Rearrangements and gene gain and gene loss are commonly observed in *E. coli* and subsequently result in the gain and loss of operons [Touchon et al., 2009]. However, if an operon of one genome has all its constituent genes each individually conserved in another genome, the order and orientation of these genes are often conserved as well [Rocha, 2008]. As another assessment of the alignment, I analyzed the conservation of *E. coli* K12_MG1655 operons when these operons were mapped by the alignment to other genomes (target genomes, 65 comparisons total). K12_MG1655 is one of the community selected *E. coli* reference genomes and its operons are well annotated.

A total of 535 K12_MG1655 operons, each comprised of two or more genes were analyzed (see Methods). Of these, on average 452 operons were *shared* with a target genome. An operon was defined to be *shared* with a target genome if all its constituent genes each were individually conserved in the target genome. Conserved was defined as being mapped by the alignment to the target genome with at least 90% coverage. Of the *shared* operons, I found only two cases of operons in which the gene orders and orientations were disrupted in the two target genomes. In both cases, the disruptions were due to rearrangements in the target genomes and not to alignment errors. Operon

Gene and Operon Alignment Assessment

Category	Total	Shared	Conserved	%
Gene Families	4751	3374	3333	98.80
Operons	535	452	452	100.00

Table 3.1: The vast majority of orthologous gene families are aligned in the MSA and on average, 100% of the K12 MG1655 operons that are present in other genomes have their gene order and orientation conserved. Together, the two assessments show that the MSA is accurate. 'Total' is defined by the average number of gene families each genome has or the total number of K12 MG1655 operons analyzed. 'Shared' is defined by the average number of gene families each genome shares with another genome (pairwise comparisons) or the average number of operons with all constituent genes conserved in another genome (pairwise comparisons). 'Conserved' is defined by the average number of shared gene families that are aligned by the MSA or average number of 'shared' operons with the gene order and orientation conserved. '%' is defined by the percentage of 'Shared' that are 'Conserved'.

envY-ompT was disrupted in five O157 genomes (EC4115, EDL933, Sakai, TW14359 and Xuzhou21) as a result of recombination. Operon *fumAC* was disrupted in *Shigella sonnei* 53G due to an inversion. Besides these two cases, 100% of the *shared* operons had their order and orientation conserved in the target genome. This observation, together with the previous observation of orthologous gene alignments, confirm the quality of the MSA.

3.3.3.2 Constructing the *E. coli/Shigella* Core Genome

To demonstrate the flexibility of comparative assembly hubs and the recently introduced pan-genome displays the software incorporates [Nguyen et al., 2014b], I created a comparative assembly hub that represented the *E. coli/Shigella* core genome.

The core genome of a collection of organisms (e.g. 66 *E. coli* and *Shigella* strains) is comprised of genomic regions shared by all the organisms. Here, the core genome was computed using the same algorithm used to impute the pan-genome with an additional requirement that every alignment block contained sequence from all input genomes (see Methods).

The core genome represents the genomic material that is essential to the species. Many studies have investigated the size and content of the *E. coli/Shigella* core genome, however the results have been inconsistent, with the size of the core genome ranging from 1,000 to 3,000 genes [Fukiya et al., 2004, Kaas et al., 2012, H et al., 2007, Lukjancenko et al., 2010, Chattopadhyay et al., 2009, Touchon et al., 2009, Vieira et al., 2011]. One obvious reason is the use of different sets of genomes. However, even when this difference is taken into account, inconsistencies abound. The main difficulties are the limitations intrinsic to traditional gene-based approaches employed to compute the core genome. Gene-based methods compute the core genome by finding the set of genes that are shared by all involved genomes. They depend heavily on clustering algorithms, different methods for prediction of "orthology", as well as gene annotation qualities.

These limitations may be circumvented by using the whole genome multiple alignment approach. This (genomic-based) approach computes the core genome from the MSA by selecting all genomic regions that are shared by every involved genome. The approach is becoming more popular as the availability of genomic data increases. Darling *et al.* [Darling et al., 2010] reported a 2.9 Mbp core genome for 16 *E.coli/Shigella*

strains and Sahl *et al.* [Sahl et al., 2011] reported a 2.7 Mbp core genome for 44 strains. In agreement with both studies, the core genome (for 66 strains) computed by the CAH pipeline is 2.7 Mbp in size. It is expected that the core genome size decreases as the number of genomes increases, until enough genomes are added, at which point the core genome size becomes stabilized [Touchon et al., 2009, Lukjancenko et al., 2010, Leimbach et al., 2013]. This observation is recapitulated here, as shown in Figure 3.9. The core genome sizes of 2.9 Mbp for 16 genomes and 2.7 Mbp for 44 and 66 genomes demonstrate a great consistency of the genomic-based approach.

The average size of an *E.coli/Shigella* genome is about 5 Mbp, of which 86% code for genes (corresponding to 5000 genes). Assuming that the genes are evenly distributed across the genome, 2.3 Mbp (86% of 2.7 Mbp) of the core genome is expected to be genic, and this quantity corresponds to about 2,300 genes. This is consistent with the average number of genes of each genome I observed to be overlapped with the core genome (2348 and 2507 genes for 98% and 90% minimum coverage cutoffs, respectively).

For comparison, I have also computed the core genome using the gene-based approach, which resulted in only 1200 genes at a minimum coverage cutoff of 90%. Manual analyses of genes included in the core genome via genomic approach but excluded via the gene based approach revealed two main reasons. The first reason is the under-annotation of genes in a small number of genomes and consequently, their exclusion out of the core genome. The second reason is the different annotated gene lengths, resulting in orthologous genes in a small number of genomes not having enough coverage to pass the cutoff. Both reasons are likely results of low-quality gene annotations

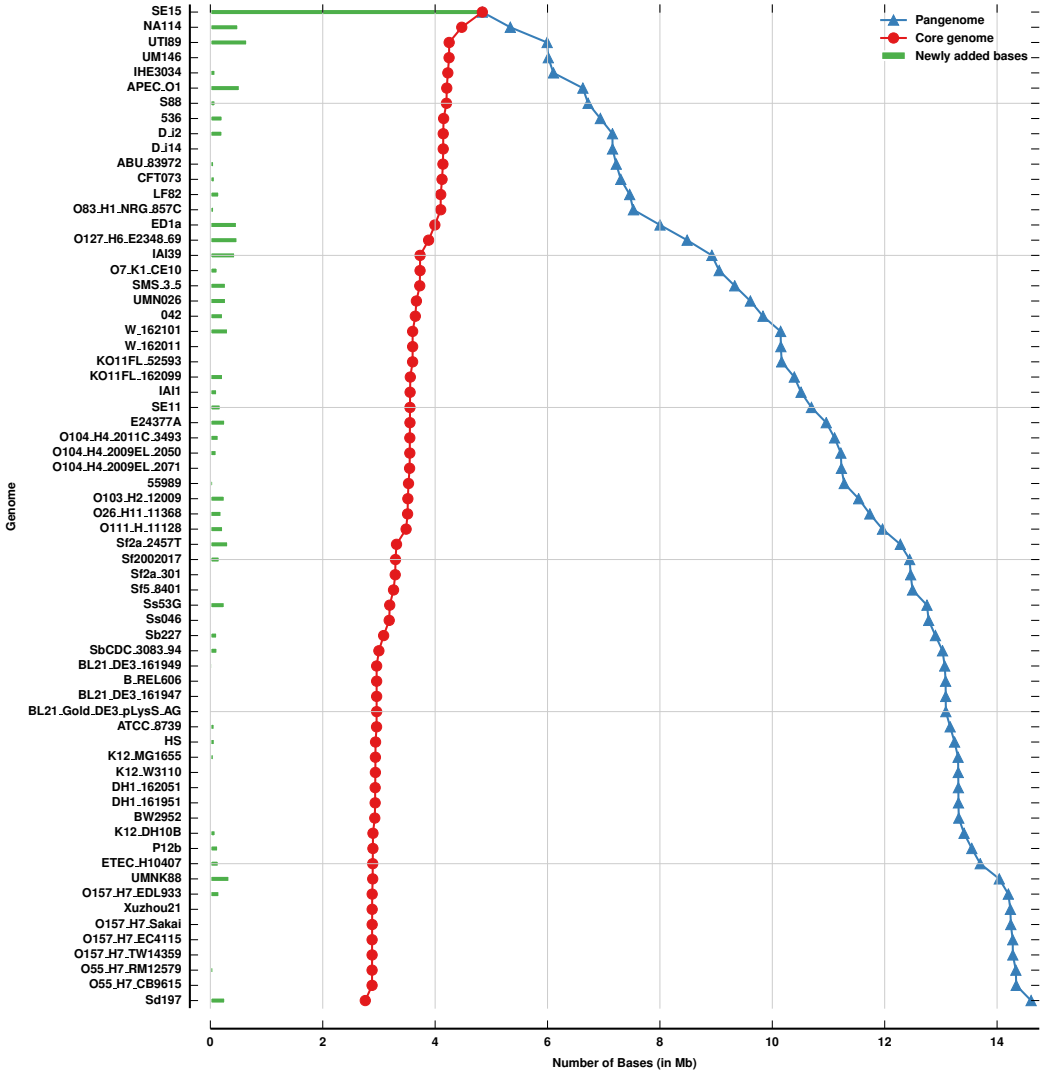

Figure 3.9: Pan-genome and core genome sizes, an adaptation of Figure 4 in Lukjancenko *et al.* [Lukjancenko et al., 2010], supporting the open pan-genome and the stable core genome models in *E. coli*. The x-axis shows the number of bases (in Mb) in the pan-genome (blue triangle), core genome (red circle) and the number of the new bases added to the pan-genome (horizontal green bar) as more genomes are added into the analysis. The y-axis shows the genome that is added each step. Genomes were added in the order guided by the phylogenetic tree in Figure 3.13.

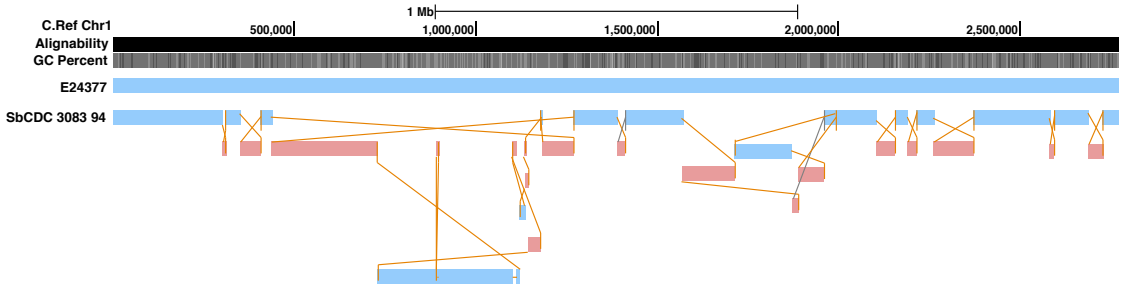

Figure 3.10: The *E. coli*/*Shigella* core genome browser, showing the highly conserved ordering relationships between blocks of the *E. coli* core genome and the less conserved ordering in *Shigella*. Most *E. coli* strains look like the first *snake* track, with no high-level rearrangements (for space only one is shown). In contrast, the Shigella strains have, with respect to Ecoli, a fragmented core genome (again, only one shown for lack of space).

of a number of genomes or of assembly errors and/or quality (that lead to the gene under-annotations/truncations).

One remarkable observation is that at a high level (approximately 10kb or greater block size), the core genome is entirely un-rearranged in the majority of *E. coli* genomes, despite the dramatic differences between them in their wider pan-genome. The consensus ordering generated by the pan-genome display algorithm (even if the blocks correspond to the core genome) is reflected in the core genome display (see Figures 3.10). The striking converse of the ordering conservation in *E. coli* is demonstrated by the *Shigella* genomes, which (as shown in the figure) are significantly reordered - though the summary allows a complete, clear tracing of this reordering.

3.3.3.3 Constructing the *E. coli*/*Shigella* Pan-Genome

The pan-genome display is created using the algorithm described in Chapter 2. Briefly, each set of homologous segments, called a block, is arranged into a set of intervals according to a consensus objective function that uses the weighted set of linkage relationships between the sides of the blocks. This consensus ordering of the blocks is then converted into a set of sequences by creating a consensus segment for each block and concatenating these consensus segments together according to the chosen ordering. The pan-genome has two attractive properties for visualization, first, it includes every block, which any single genome very likely does not, making it possible to get a complete picture of all variations present within a chromosome. Second, as it includes exactly one copy of each block, it contains no self-alignments, and thus all duplications are representable within the target genomes.

An example is in Figure 3.11, showing the Shiga toxin (*Stx1* and *Stx2*) gene family displayed along the pan-genome browser, with subunit A on the left and subunit B on the right. The number of *Stx* genes as well as which *Stx* groups (*Stx1* or *Stx2*) present vary among the genomes. The pan-genome allows for a clear presentation of this variation, showing that Sd197 has one *Stx1* copy, O157 H7 Sakai has one *Stx1* copy and one *Stx2* copy, O104 H4 2011 C3493 has one *Stx2* copy, and O157 H7 EC4115 has two *Stx2* copies. Variations (indels and substitutions) between *Stx1* and *Stx2* are also shown. (Appendix Figure B.1 shows the same region, with all genomes containing the *Stx* genes displayed.)

Figure 3.9 shows the core and pan-genome sizes of an increasing number of input genomes. The genomes were added in the order guided by the phylogenetic tree (see Section 3.3.4 below), with the y-axis in Figure 3.9 showing the genome that is added each step. The curves contain bumpy regions corresponding to a faster increase in the number of bases being added to the pan-genome and, conversely, a faster decrease in the number of bases being removed from the core genome. These regions correspond well to the boundaries of grouping in the phylogenetic tree, consistent with the idea that genomes within individual phylogroups (or subgroups) share more bases than genomes between phylogroups (as further confirmation see also Figure 3.12, which shows the number of genes shared by each pair of genomes).

High inter-phylogroup sharing may indicate some common phenotypes among the genomes involved, an example is the sharing between genomes Xuzhou21, O157 H7 Sakai, O157 H7 EDL933, O157 H7 EC4115, O157 H7 TW14359, O55 H7 CB9615 and O55 H7 RM12579 of group E and genomes O103 H2 12009 and O26 H11 11368 of group B1. The two O55 strains are enteropathogenic *E. coli* (EPEC) and the others are Enterohaemorrhagic *E. coli* (EHEC) and their high inter-group sharing reflects the EHEC/EPEC-specific genes [Ogura et al., 2009], that can be visualized with the comparative assembly hub, as in Figure 3.8.

3.3.4 Constructing the *E. coli*/*Shigella* Phylogenetic Tree

Based upon the core genome provided by the MSA, I built a maximum-likelihood based phylogenetic tree for the 66 *E. coli*/*Shigella* strains using

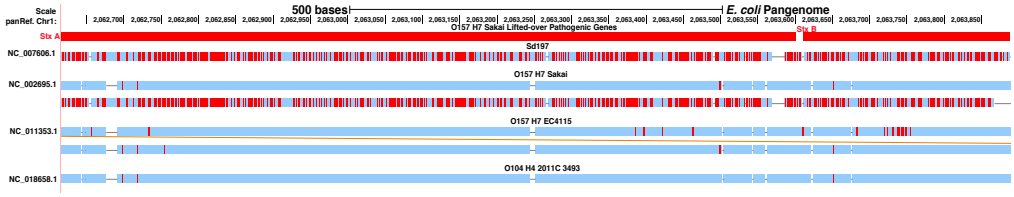

Figure 3.11: The Shiga toxin region displayed along the pangenome browser, showing a subset of genomes containing the *Stx* genes. The "O157 H7 Sakai Lifted-over Pathogenic Genes" track shows the lifted-over pathogenic gene annotations of strain O157 H7 Sakai, with the *Stx* subunit A on the left and subunit B on the right. There are two major groups of *Stx*, *Stx1* and *Stx2*. Different genomes contain different numbers of *Stx* genes as well as different *Stx* groups. The pangenome view allows for the presentation of these variations, showing that Sd197 has one copy of *Stx1*, O157 H7 Sakai has one copy of *Stx1* and one copy of *Stx2*, O104 H4 2011 C3493 has one copy of *Stx2*, and O157 H7 EC4115 has two copies of *Stx2*. As there are more copies of *Stx2* than *Stx1* (12 versus 7, see Appendix Figure B.1 for the complete browser display of all *Stx* carrying genomes), the pangenome, which is a consensus sequence, is more similar to *Stx2* than *Stx1*, visibly by many SNPS on the *Stx1* copies. Variations (SNPs and indels) between the two groups *Stx1* and *Stx2* and among different genomes are shown. The texts (the labels) on the screenshot were minorly adjusted for better readability.

RaXML[Stamatakis, 2006] (see Methods). The resulting tree (Figure 3.13) is consistent with the genomes' phylogroup annotations (Appendix Table B.2.2) as well as previously reported trees built using the core genome (for smaller sets of *E. coli* and *Shigella* genomes because there were not as many complete genomes available at the time of those publications as there are currently) [Touchon et al., 2009, Perna, 2011, Leimbach et al., 2013, Chaudhuri et al., 2010]. The tree reconfirms 1/ the monophylic of phylogroups B2, D2, D1, E, A, and B1 with group B2 as the basal lineage, 2/ the spread of similar pathotypes across distinct lineages and 3/ the multiple origins of *Shigella* spp.

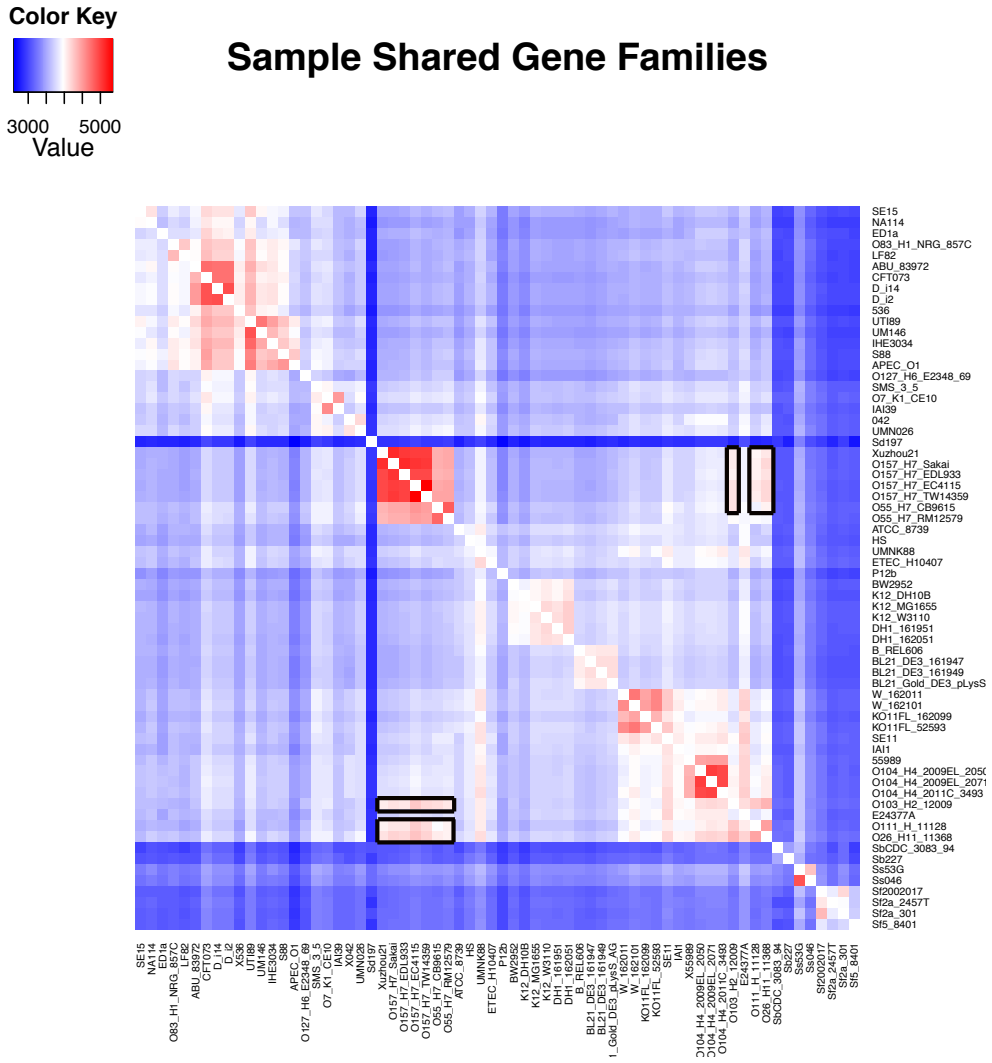

Figure 3.12: Heatmap of the number of orthologous gene families shared by all pairs of genomes. Rows and columns are genomes in phylogenetic order provided by Figure 3.13, showing higher intra-phylogroup sharing in comparison with inter-phylogroup sharing. There is, however, a visible relatively high inter-phylogroup sharing among genomes Xuzhou21, O157 H7 Sakai, O157 H7 EDL933, O157 H7 EC4115 and O157 H7 TW14359 of group E and genomes O103 H2 12009 and O26 H11 11368 of group B1 (boxed in black). These genomes are all Enterohaemorrhagic or Enteropathogenic *E. coli* (EHEC/EPEC), which suggests that their high inter-phylogroup sharing reflects previously reported EHEC/EPEC-specific genes [Ogura et al., 2009].

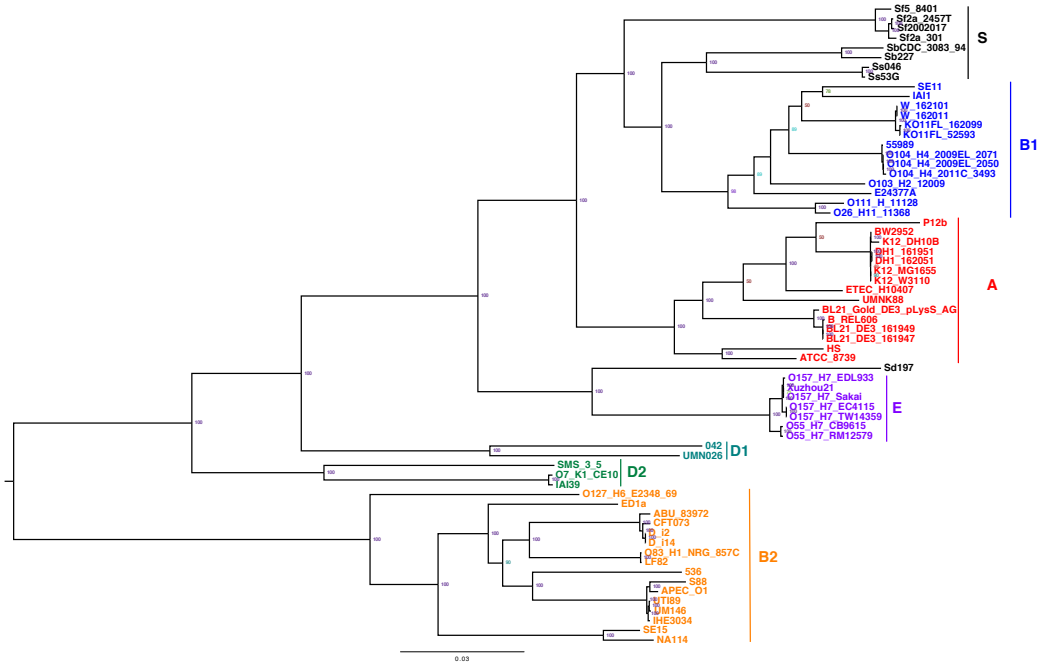

Figure 3.13: Maximum-likelihood based phylogenetic tree of 66 *E. coli* and *Shigella* spp. genomes, constructed from their core genome alignment using RAxML. The genomes are colored by their annotated phylogroups (Appendix Table B.2.2): orange: B2, green: D2, teal: D1, purple: E, red: A, blue: B1 and black: *Shigella*.

3.4 Discussion

In this work I have shown how UCSC comparative assembly hubs can be easily constructed to provide useful, extensible browsers for collections of evolutionarily related genomes.

The CAH framework is novel in several respects. All the alignments and lifted over annotations shown are mutually consistent with one another because for the first time, the annotation lift over and alignment display are symmetrically driven by one reference free alignment process, rather than a mixture of different pairwise and reference based multiple alignments. Being reference free, the multiple alignment

process and HAL format also allow us to make all the browsers equivalently powerful: all the annotations and alignments can be displayed from any vantage point. The *snake* tracks for the first time in the UCSC browser and a linear display format, fully express all the possible mutation types in one track, while the resolution scaling makes this useful at all resolution levels. The pan-genome display gives a new view of the data that for some purposes is more useful for display than any single genome.

I used *E. coli* and *Shigella* spp. genomes as a test, and demonstrated that the alignments were able to accurately align the vast majority of genes correctly, and automatically reconstruct a core genome of all *E. coli*/*Shigella* that recapitulates earlier results, and visually demonstrates the many rearrangements present in the core genome of the *Shigella* phylogroup. The *E. coli*/*Shigella* comparative assembly hubs are now available on the UCSC Browser public assembly hub listing, facilitating data exploration and analyses. Such resources can be easily generated and updated as new data become available.

Comparative assembly hubs have been tested with clades of mammalian genomes (a forthcoming Reptile/Bird comparative of 23-genomes is in the process of being made public). It is feasible to use for large projects, providing that significant computational resources are available in the form of compute clusters. To make the tool practical for vertebrate genomics communities without these resources, one future aim of the project is to provide a cloud service, where users could buy compute time to generate their alignments.

3.5 Methods

3.5.1 Alignment Assembly Hub Pipeline

From a set of input genome sequences (and any available annotations), users can create the comparative assembly hub using the two following commands:

1. *runProgressiveCactus.sh <seqFile> <workDir> <outputHalFile>*

2. *hal2assemblyHub.py <halFile> <outDir> –bedDirs <annotationDirs> –lod*

Command (1) generates the multiple sequence alignment, which is stored in the HAL format to the specified output file $outputHalFile$. $seqFile$ contains Newick-formatted phylogenetic tree of the input genomes (optional) and paths to the sequence FASTA files. $workDir$ specifies the working directory. More details can be found in Progressive Cactus Manual (`https://github.com/glennhickey/progressiveCactus`).

Command (2) produces necessary data and files for creating the comparative assembly hubs through the UCSC genome browser. $halFile$ is the HAL-formatted MSA file, which is the output ($outputHalFile$) from command (1). $outDir$ is the output directory where all the generated files are written into. Among the output files is a file named "hub.txt", which the users will upload to the UCSC genome browser (similarly to how a track hub is created [Raney et al., 2013], see `http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html` for more details) and the comparative assembly hubs will be created. $annotationDirs$ is a comma separated list of directories containning the annotation files, one directory per annotation type (e.g genes, pathogenic regions, antiobiotic resistance regions). Option *–lod* is specified to compute the levels of

detail, which is recommended for large datasets. For parallelism and job management, *hal2assemblyHub.py* uses jobTree (`https://github.com/benedictpaten/jobTree`), which is installed as part of Progressive Cactus installation process. Users can specify different jobTree options to speed up the running time.

In this work, I generated three different *E. coli/Shigella* comparative assembly hubs. One with duplications allowed (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliWithDups/hub/hub.txt`), one with duplications disallowed (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliNoDups/hub/hub.txt`), and one that disallowed duplications and required all genomes to be present in every block (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliCore/hub/hub.txt`).

Each of the hubs was generated by the two following commands:

1. *runProgressiveCactus.sh –legacy –configFile config.xml –maxThreads 24 –ktType snapshot seqFile.txt outdir outdir/alignment.hal*

2. *hal2assemblyHub.py alignment.hal outHubDir –maxThreads 24 –lod –bedDirs Genes,RNA,GI,PI,PathogenicGenes,ARGB –rmskDir rmskTracks –gcContent –alignability –conservation conservationRegions.bed –conservationGenomeName reference –conservationTree tree.nw –tree tree.nw –rename shortnames.txt –hub ecoliCompHub –shortLabel EcoliCompHub –longLabel "Escherichia coli Comparative Assembly Hub"*

All related files can be found at `http://compbio.soe.ucsc.edu/`

`reconstruction/ecoliComparativeHubs`, under directories "ecoliWithDups", "ecoliNoDups" and "ecoliCore', respectively. For more details of the options, please see the *hal2assemblyHub* documentation at `https://github.com/glennhickey/hal`.

3.5.2 Genome Sequence and Annotation Data

Nucleotide sequences of 57 *E. coli* and 9 *Shigella* spp. complete genomes were downloaded from the NCBI ftp site (`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz`, January 2013). The sequences were repeat-masked using RepeatMasker [**?**] with the '*-xsmall*' option and otherwise default settings. The repeat-masked sequences were used as inputs to construct the MSA. Other outputs of RepeatMasker were converted into bigBed format to build the "Repetitive Elements" track for each genome (`http://genomewiki.ucsc.edu/index.php/RepeatMasker`). For the 9 genomes ATCC 873, DH1 161951, KO11FL 162099, KO11FL 52593, O104 H4 2009EL 2050, O104 H4 2009EL 2071, O104 H4 2011C 3493, UM146, BL21 Gold DE3 pLysS AG, I used the reverse complement of their assemblies as the majority portion of those assemblies aligned to the reverse strand of other (57) genomes.

Gene, protein and non-coding RNA annotations for each genome were also obtained from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.gff.tar.gz`, `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.faa.tar.gz` and `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.rnt.tar.gz`, respectively).

In a number of genomes, I observed and corrected obvious errors (Appendix Section B.2), such as genes with positions that were out of range of the sequence length

(Appendix Table B.2.2) and genes with multiple exons that overlapped with each other (self-overlapped, Appendix Table B.2.2).

3.5.3 Gene and Operon Analyses

Paralogous and orthologous annotated coding genes were identified by BLAT amino acid sequence pairwise alignments. For each genome, genes were grouped into a single gene family if they shared at least 90% amino acid identity over at least 90% of the length of the longest gene.

To identify orthologous gene families shared among the genomes, I used the divide and conquer approach. Briefly, I started by breaking the input set of genomes into pairs. For each pair of genomes, I computed their union list of gene families by grouping orthologous gene families together. The resulted union gene family lists of all pairs were recursively treated as a new set of genomes and the process of finding union lists was repeated until orthologous gene families of all genomes were grouped together and one union gene family list was obtained. Two gene families of two genomes was identified as orthologous if at least one gene of one family had a reciprocal match with at least one gene of the other family. A match was defined as having at least 90% amino acid identity and 90% coverage.

To assess the multiple sequence alignment, for each pair of genomes, I computed the number of orthologous gene families that were aligned in the MSA and reported the average statistics of all pairs. Two orthologous gene families were considered as aligned in the MSA if at least one gene of one family was aligned to one gene of the

other family by the MSA with a minimum coverage of 90% of the longer gene.

As another assessment of the MSA, I analyzed the gene order and orientation conservation of the well-annotated *E. coli* K12 MG1655's operons that were also present in other genomes. Operons (or more accurately, transcription units) of K12 MG1655 were downloaded from RegulonDB (see Section 3.5.2). As the orders and orientations of the genes were of interest, only operons with two or more genes (and no pseudogene) were included in the analysis. In addition, I filtered out annotations without strong evidences, which I defined as operons with no other evidence than one of the following: "Inferred by computational analysis" (ICA), "Inferred computationally without human oversight" (ICWHO), "Non-traceable author statement" (NTAS) and "Polar mutation" (PM). After filtering, there were 535 operons total. For each genome other than K12 MG1655 (called target genome, 65 genomes total), I calculated the number of K12 MG1655 operons that had all their constituent genes present (having an ortholog) in the target genome and the percentage of these operons with the gene order and orientation conserved by the MSA. I reported the average statistics in Table 3.3.3.1.

Chapter 4

Comprehensive assessment of T-cell receptor repertoires

4.1 Overview

Advancements in high-throughput sequencing have enabled deep and quantitative analyses of the adaptive immune system. The emerging field of immunosequencing, based on the ability to read millions of B and T-cell receptor sequences in parallel for each sample, comprises of a large collection of immunological and clinical applications. As a result, many large datasets have been generated and many more are underway. This rapid production of data has necessitated the developments of new computational tools and algorithms to process and analyze this data. While there have been extensive tool developments for read mapping, gene calling, sequencing error corrections and individual repertoire assessments, very limited (if any) efforts have been invested in large-scale comparative analysis software. Here, I report on an open-source software package for comprehensive assessments and comparative analyses of T-cell receptor (TCR) repertoires, called the "Adaptive IMmunoSequencing ToolKit" or *aimseqtk*. The *aimseqtk* package comprises of four main components: The first three components address the three common applications of TCR sequencing: clone tracking across multiple repertoires (multiple tissues/conditions/time points), repertoire signature profiling (including individual and comparative analyses of diversity, similarity, gene-segment usage, CDR3 length distribution, clone size distribution and recombination model) and public clones identification. The fourth and last component, publication mining, utilizes the UCSC genome browser publication mining pipeline to identify previously published clones that are homologous to user-specified clones. The previously published clones are compared

with the user-specified clones for exploration of context, literature validations and assessments, and/or potential correlations. All four components are incorporated into an easily executed single pipeline for systematically and comprehensively analyzing the data at hand.

I applied the *aimseqtk* package to study TCR repertoires of the autoimmune disease Ankylosing Spondylitis (AS). The results show that in comparison with healthy repertoires, AS T-cell repertoires have higher diversity and similar CDR3 length distributions and gene-segment usage. Given the limited set of samples, the results presented here are preliminary, however, they demonstrate that with a sufficiently large set of samples, similar analyses can be conducted using the *aimseqtk* package to find evidence for antigen selection (if existed) in AS TCRβ repertoires and to identify potential autoreactive clones that may be involved in the disease development.

4.2 Introduction

Two main players of the adaptive immune system are B-cells and T-cells. The scope of this chapter and of the *aimseqtk* package will only focus on T-cells and TCR sequencing.

4.2.1 T-cell Receptor

The main players of the cell-mediated immunity are T lymphocytes, or T-cells. T-cells recognize foreign antigens that are presented by major histocompatability complex (MHC) molecules, and trigger appropriate immune responses to protect the body.

Which specific antigen that each T-cell responds to depends on the structure of the T-cell receptor expressed on its surface. The T-cell receptor, or TCR, is a membrane-bound molecule that is responsible for recognizing and binding to MHC-antigen complexes [Krogsgaard and Davis, 2005]. Each TCR is composed of two chains, α and β, or γ and δ. About 95% of T-cells in the body are $\alpha\beta$ T-cells, and only 5% are $\gamma\delta$ T-cells [Kindt et al., 2007]. The domain structures of TCRs are very similar to those of the immunoglobulins, and thus they belong to the immunoglobulin superfamily. Each chain in a TCR has two domains, one variable (V) and one constant (C). The constant region is anchored in the cell membrane, while the variable region faces outward and binds to the MHC-antigen complex. C domains stay constant across different TCRs, while V domains of both chains exhibit sequence variation. There are three hypervariable regions found in the V domains, called the complementarity-determining regions, or CDRs. Among these CDRs (CDR1, CDR2, and CDR3), CDR3 displays the greatest variability. In the interaction between TCRs with antigens, CDR3 provides the primary contact with the antigenic peptide. CDR1 and CDR2 interact with conserved surface features of the MHC molecules. Variations in the CDRs contribute to the diversity of TCRs, and allow for a vast repertoire of antigens that the TCRs recognize [Kindt et al., 2007].

The genes that encode the T-cell receptor are expressed only in cells of the T-cell lineage. In germlines, TCR loci are organized into multigene families corresponding to the α, β, γ, and δ chains. Each family contains multiple segments of genes called V (Variable), D (Diversity, D segments are only in β- and δ-, and not in α- and γ- chain families), and J (Joining). Functional TCR genes are produced by rearrangements of V

and J segments in the α-chain and γ-chain families and V, D, J segments in the β-chain and γ-chain families. V segments of the α-chain are called V_α, and similarly for V_β, J_α, etc. In human, the numbers of functional segments are: 79 V_α, and 38 J_α; 21 V_β, 2 D_β, and 11 J_β; 7 V_γ, and 3 J_γ; 6 V_δ, 2 D_δ, and 2 J_δ[Kindt et al., 2007].

During the development of the T lymphocyte, the T-cell receptor chains undergo V(D)J recombination, also known as somatic recombination. This process generates a diverse repertoire of TCRs that are necessary for the recognition of diverse antigens. In the β chain of the TCR, the first recombination event is between one D and one J gene segment. This can involve either the joining the $D_\beta 1$ gene segment to any of the $J_\beta 1$ segments or the joining of the $D_\beta 2$ gene segment to any of the $J_\beta 2$ segments. This D-J recombination is followed by the joining of one V gene, forming a rearranged VDJ gene. Any DNA between the selected V, D, and J is deleted. The rearrangement of the α-chain loci is similar to the β-chain process, except there is no D segment involved. One V gene joins with one J gene, and forms a VJ segment (Figure 4.1). The segments of V and J that get selected during recombination are random. Each combination of V_α, J_α and V_β, D_β, and J_β results in a unique antigen receptor. With more than 10^6 of possible combinations, somatic recombination provides the body with a rich repertoire of TCRs to protect against the variety of antigens [Kindt et al., 2007].

During the somatic recombination process, additional diversity is introduced by random deletions and insertions of nucleotides at the recombined junctions. In particular, random numbers of nucleotides (≥ 0) are deleted from the 3' end of V genes, both 5' and 3' ends of D genes and 5' end of J genes when these genes get rearranged. Ran-

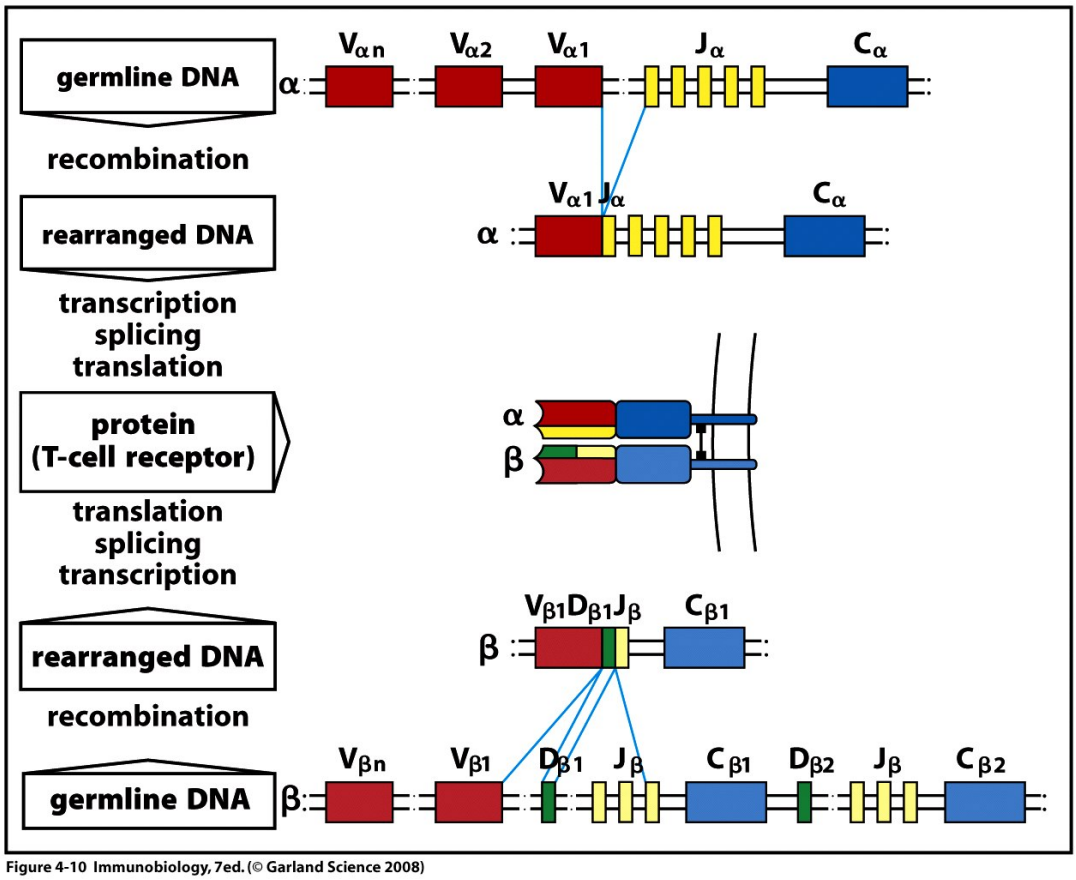

Figure 4.1: VDJ recombinations of the TCR. Figure is from Janeway *et al.*, 2004 [Janeway et al., 2004]

dom nucleotides are also inserted to the junctional regions [Kindt et al., 2007]. CDR1 and CDR2 are encoded within the V gene, while CDR3 encompasses the junction of V_α with J_α and V_β with D_β and D_β with J_β. The diversity is greater in CDR3, due to the junctional diversity and the addition and deletion of nucleotides at the junctions of the gene segments.

4.2.2 T-cell Development and Clonal Expansion

Progenitor T-cells arise from hematopoietic stem cells in the bone marrow. In human, progenitor T-cells begin to migrate to the thymus in the eighth or ninth week of gestation, and generate a large population of immature T-cells, called thymocytes. In the thymus, thymocytes proliferate and differentiate into functionally distinct subpopulations of mature T-cells. The maturation of T-cells involves rearrangements of the germline TCR genes and the expression of various membrane markers. This results in a diversity of T-cells, for which each produces a specific type of TCR. This diversity is shaped into an effective primary T-cell repertoire by two selection processes: one positive and one negative. The positive selection permits only T-cells whose TCRs can recognize self-MHC molecules, while the negative selection makes sure that T-cells that have too strong an affinity to self MHC or self MHC plus self peptides got eliminated. These processes ensure that only MHC-restricted and nonself-reactive T-cells mature [Kindt et al., 2007].

There are two subpopulations of T-cells: $CD4^+$ and $CD8^+$. $CD4^+$ T-cells recognize antigen presented by class II MHC molecules and generally function as helper T-cells. $CD8^+$ T-cells recognize antigen presented by class I MHC molecules and generally function as cytotoxic T-cells. $CD4^+$ and $CD8^+$ T-cells leave the thymus and enter the circulation as resting cells in the cell cycle G_0 stage. Naïve T-cells, T-cells that have not yet encountered antigen, continually recirculate between the blood and the lymph systems. Once an antigenic peptide is presented by the MHC molecules, the MHC-

peptide complex is scanned by the millions of circulating T-cells [Kindt et al., 2007].

If a naïve T-cell recognizes an antigen-MHC complex, it will be activated, initiating a primary response. The activated T-cell is induced to proliferate and differentiate. Each T-cell divides two to three times per day for four to five days, generating a clone of progeny cells, which differentiate into memory or effector T-cells. Effector T-cells are short-lived cells, and various effector cells have different functions to fight against the pathogen: help to B-cells, cytokine secretion, and cytotoxic killing activity. Effector cells are derived from both naïve and memory cells after antigen activation. The memory T-cell population is derived from both naïve cells and effector cells. Memory cells are long-lived cells that provide a heightened immune response to the subsequent attacks of the same antigen, generating a secondary response. Both primary and secondary responses enable a significant increase in the number of T-cells that recognize the attacking antigen, which is known as clonal expansion [Kindt et al., 2007].

4.2.3 Immunosequencing: Challenges and Applications

4.2.3.1 Challenges

Somatic recombination makes high-throughput sequencing (HTS) of the adaptive immune receptors (BCRs and TCRs) challenging. The rearranged receptor sequences are highly diverse and differ significantly from the (known) germline templates, making it difficult to design the primers for PCR amplification, currently a required step for targeted sequencing.

For cDNA, or reverse transcribed mRNA, in which the intron between the

recombined exon and the constant exon is spliced out, only two reverse primers specific to the constant region are required [Freeman et al., 2009]. Although amplification of receptor sequences is simplified with cDNA, quantitative interpretation of the data is difficult. The amount of TCR or BCR mRNAs being expressed varies across different T or B-cells as well as different developmental stages that they are in. It is therefore difficult to accurately quantify clone counts and analyze clonal expansion and contraction, which are important for studying adaptive immune functions [Robins, 2013]. For genomic DNA (gDNA), in which there is typically one copy of a productive TCR recombination per cell [Kindt et al., 2007], the aforementioned quantitative problem does not apply. However, the presence of the intron between the J and the constant region (Figure 4.1) distances the constant region from the junctional region and necessitates a large number of degenerate J-specific reverse primers, each with slightly different annealing efficiency.

4.2.3.2 Target Enrichment Solutions and Recent Developments

Recent developments have made HTS of the TCR repertoire possible. Various groups have been able to design a mixture of V and J primers that are inclusive of all V and J segments with high specificity [Robins et al., 2009, Wang et al., 2010]. Using the primer mixture, multiplex PCR is used to amplify the full set of potential recombinations.

The next challenge, expected from multiplex PCR, is amplification bias. When designed, the primers are computationally designed and selected for similar annealing

temperatures. This optimization, however, is not robust and significant amplification biases are still observed in practice [Robins et al., 2009]. To address this problem, two different solutions have been carried out and proved to sufficiently remove the majority of amplification bias. The first solution is to sequence a synthetic set of T-cells with known clonal frequencies, and to adjust the amount of each input primer (up or down) based on the differences between the expected and the observed frequencies [Carlson et al., 2013]. The process is then repeated until achieving minimal discordances, which are then removed computationally. The second approach is to use nested PCR in which a universal synthetic sequence is attached to the 5' end or 3' end of each V or J primer, respectively [Wang et al., 2010]. In the first few rounds of the PCR process, the V and J specific primers are used. However, in all subsequent rounds, the universal primer is used. The idea is that since a universal primer is used in most amplification rounds, the strategy avoids most of the amplification bias.

These approaches address the underlying challenges and allow for reading and accurately quantifying millions of BCRs and TCRs of each sample. The new field immunosequencing has emerged, which specializes in HTS of the adaptive immune repertoires (BCRs and TCRs) and expansive associated applications.

It is important to note that current technologies only allow for high-throughput investigation of either TCRβ chains or, less common, TCRα chains of the TCR molecules. Therefore, in the current context, a "clone" reflects only part of the TCR molecule. Tremendous efforts are invested in sequencing pairs of TCRα-TCRβ of the TCR molecules.

4.2.3.3 Immunological and Clinical Applications

The ability to sequence the adaptive immune receptor repertoires at high resolution has opened a door to countless immunological and clinical applications. The three main utilities most popular are repertoire profiling, clone tracking and public clone identification.

Repertoire profiling includes assessing various properties of the repertoires such as diversity, clonality, CDR3 length distribution, gene-segment usage, amino acid usage, nucleotide insertions and deletions. This has proved fruitful in many immunological basic research studies [Robins et al., 2009, Freeman et al., 2009, Klarenbeek et al., 2010, Wang et al., 2010, Robins et al., 2010]. For example, by using HTS, it has been found that the average diversity ($\geq$ one million clones) of the human TCRβ repertoire is discovered to be many times higher than previously estimated [Robins et al., 2009, Warren et al., 2011] and the number of clones shared among different individuals are 7000-fold higher than predicted [Robins et al., 2010]. In another example, the $CD4^+$ T-cell subset has been shown to have different repertoire signatures compared with $CD8^+$ T-cell subset, based on which a heuristic algorithm is developed to computationally estimate the ratio of $CD4^+/CD8^+$ T-cells in a mixture sample [Emerson et al., 2013]. Similarly, repertoire profiling can be applied to find signatures, if they exist, that are representative of specific diseases. A number of such studies are underway (see the Repertoire 10K (R10K) project `http://www.r10k.org/R10K/Projects/Projects.html`). Comparative analyses of repertoires of different conditions, such as different ages

[Rudd et al., 2011, Jiang et al., 2011], different tissues [Klarenbeek et al., 2012], or before and after treatments (such as transplantation or vaccination [Muraro et al., 2014]) help to study the dynamics of the adaptive immune system in response to or in correlation with the variables of interest.

Clone tracking is monitoring the presence or absence and the clonality (size/ frequency) of one or more specific clones across different time points or different conditions. It has been applied robustly in lymphoid malignancies to track the cancerous B or T-cell clones that present predominantly (in many cases $\geq 90\%$) in the blood, bone marrow, or lymph node of the patients. In fact, the first clinical application of TCR HTS, thanks to the technology's extreme sensitivity, is the tracking of MRD (minimal residue disease) before and after chemotherapy to detect the complete removal or reoccurrence of cancer clones [Wu et al., 2012]. Another common use of clone tracking is studying the dynamics of naïve, effector and memory T-cells, especially in the context of exposure to pathogens or vaccinations [Wang et al., 2010, Warren et al., 2011, Burrows et al., 2013]. In autoimmune diseases, the technique is used to look for autoreactive clones that are enriched in effected tissues/organs rather than in the other body regions [Maecker et al., 2012, Klarenbeek et al., 2012]. In immunotherapy, the set of infused clones can be monitored *in vivo* if HTS is applied to patient blood samples overtime [Grupp et al., 2013].

Public clones are present in repertoires of multiple individuals in the population. In contrast, private clones are present only in a single individual. Since the set of all possible TCR sequences is extremely large (e.g. $\sim 5 \times 10^{11}$ possible

TCRβ sequences), the theoretical probability that a clone is shared by more than one individual is extremely low [Robins et al., 2010]. In practice, unexpectedly high clonal overlaps have been observed in many healthy individuals. Possible explanations of this phenomenon have been attributed to MHC selection and antigen selection [Robins et al., 2010, Rudd et al., 2011, Koning et al., 2013]. As individuals are likely to be exposed to similar pathogens in the environment, similar TCRs are likely to be selected. The same concept is found in diseases involving sets of common antigens, presuming there exists a set of disease-specific clones, selected for targeting such antigens, commonly observed in patient repertoires but absent in others. Identifying such clones potentially accelerates the development of targeted therapeutic treatments, which may be less invasive and may be important medical advancements.

4.2.3.4 Software Developments

The broad scope of applications has yielded copious amounts of immunosequencing data. Followed this is the materialization of an active subfield which specializes in computational method and tool developments for handling this new of type data. Three main areas that have been concentrated on are: read mapping and gene calling, sequencing error correction, and individual repertoire assessment. Read mapping and gene calling involve efficiently aligning a large set of reads to the most likely V, D and J for each read, and characterizing other recombination information such as the number of deletion and insertion bases within a junction. Sequencing error correction typically involves collapsing identical reads into clones, clustering clones with high identity to

each other, and using read quality (and/or high-frequency clones of the same clusters) to either correct or filter out clones with low frequencies (low read counts). Individual repertoire assessment includes basic analyses of an individual repertoire such as calculating proportions of productive clones, diversity, clonality, gene-segment usage, amino acid usage and CDR3 length distribution.

Publicly available software for read mapping and gene calling of TCR data includes HighV-QUEST [Li et al., 2013] and Decombinator [Thomas et al., 2013], both of which employ the deterministic approach, and Murugan *et al.* [Murugan et al., 2012], which uses the probability approach utilizing expectation maximization. Decombinator also handles sequencing error correction, as does the recently published competitive software MiTCR [Bolotin et al., 2013]. For repertoire assessment, the only publicly available software, to my knowledge, is our local resource called the UCSC Immunobrowser [Kim et al., 2014]. As for private resources, the three main commercial companies that provide immunosequencing services (Adaptive Biotechnologies, iRepertoire and Sequenta) also provide similar computational analyses.

While the field has rapidly developed in the aforementioned areas, one emerging research area remains to be explored: comparative immunogenomics resources. While small-scale (ranging from 1 to < 50 samples) comparative analyses have been published in various studies, there has yet to exist a standard software to perform such analyses, especially at a large scale. Typically, each group needs to write its own scripts that are customized to the specific study of interest, or to purchase customized computational services. The importance of such comparative analyses in immunosequencing

applications calls for the development of such software.

4.2.4 Autoimmune Diseases

Autoimmune diseases arise when the host's immune system inappropriately reacts against self cells and organs. There are two main mechanisms that protect the body from self-reactivity. The first mechanism, termed central tolerance, eliminates immature B and T-cells when their affinity to self-antigens is higher than a specified threshold. The second mechanism, termed peripheral tolerance, inactivates self-reactive B and T lymphocytes that survive the initial screening process of central tolerance. Failure of these tolerance processes results in attacks against self components, and the possible onset of autoimmune disease [Goodnow et al., 2005, Hogquist et al., 2005].

The current understanding of autoimmunity consists of several themes. First, autoimmune disorders have a complex genetic basis that involves multiple causally related genes, each with a generally modest effect. Second, some genetic variants are clearly predisposed to multiple autoimmune diseases, implicating common pathways of pathogenesis. At the same time, the lack of such overlap for some diseases indicates that distinct mechanisms also exist [Gregersen and Olsson, 2009]. In addition, environmental component has been reported to be involved in the development of autoimmune diseases.

4.2.5 Ankylosing Spondylitis

Ankylosing Spondylitis (AS) is an autoimmune disease affecting approximately 350,000 persons in the US, 600,000 in Europe, and 0.1-1.0% of the world population. AS, a form of Spondyloarthritis, is a chronic, progressive, connective tissue disorder that is characterized by inflammation of the spinal and sacroiliac joints. AS can cause an eventual fusion, and may lead to a complete rigidity, of the spine. AS has a strong genetic predisposition: 95% affected individuals carry a specific allele of an MHC I gene, HLA-B27, which presents in only 9% of the general population [Märker-Hermann and Höhler, 1998, Khan, 2000]. Although the disease's major risk factor, HLA-B27, has been discovered for almost 40 years [Brewerton et al., 1973, Schlosstein et al., 1973], the etiology of AS is not yet clearly understood.

4.2.5.1 The Role of $CD8^+$ T-cells in AS Etiology

$CD8^+$ cytotoxic T-cells are thought to play an important role in AS pathogenesis. The rationale for their involvement comes from the disease's strong association with the MHC class I molecule HLA-B27, whose canonical function is peptide presentation to $CD8^+$ T-cells [Brewerton et al., 1973, Schlosstein et al., 1973]. This strong association has led to the long-standing arthritogenic hypothesis, which postulates that AS resulting from the ability of HLA-B27 to bind and present unique arthritogenic peptides to $CD8^+$ T-cells, with a response that cross-reacts with self-antigens and triggers the disease onset [Tam et al., 2010].

This arthritogenic hypothesis, which suggests a direct involvement of $CD8^+$ T-cells in AS, is supported by the increased number of circulating CD8+ T-cells in an AS patient's blood [Schirmer et al., 2002] and of CD8+ T-cell clonal expansions in their synovial fluid and peripheral blood [Dulphy et al., 1999, Duchmann et al., 2001, Mamedov et al., 2009]. Autoreactive T-cells, including B27-restricted $CD8^+$ T-cells with specificity for self-peptides and bacterially infected autologous cells, have been observed in AS and a closely related autoimmune disease, Reactive Arthritis (ReA) [Hermann et al., 1993, Duchmann et al., 1996, Appel et al., 2004, Fiorillo et al., 2000, Atagunduz et al., 2005]. Lastly, persistent [Mamedov et al., 2009] and shared clonal expansions [Dulphy et al., 1999, Mamedov et al., 2009] occur among AS patients.

This evidence suggests that $CD8^+$ T-cells participate in the development of AS. Understanding their etiology and functions will help to understand the disease mechanism. Identification of pathogenic auto-reactive $CD8^+$ T-cells may lead to successful selective suppression of these clones as in other diseases, through antibody therapy or immunization by TCR protein, peptide, or DNA [Mamedov et al., 2009].

4.2.5.2 AS TCR Repertoire Studies

At the time of the AS study reported in this chapter, the most (and only) mass sequencing published studies of AS T-cell repertoires were Mamedov *et al.* 2009 [Mamedov et al., 2009] and its follow-up study Britanova *et al.* 2012 [Britanova et al., 2012]. The first study used Sanger sequencing and generated TCR sequences for two patients, 2400 sequences per patient. The later study used 454 se-

quencing and generated between 11,000 and 19,500 TCR sequences for a single patient. Both studies assessed the changes of the AS T-cell repertoires over time. The studies were limited by the absence of healthy controls. Furthermore, current technologies using Illumia sequencing allow for yet higher throughput and provide a better resolution, which is at least two orders of magnitude greater than has been achieved in Britanova *et al.* 2012 [Britanova et al., 2012].

In collaboration with the Zuniga lab, the Pourmand lab and doctor Brent Culver, I applied Illumina sequencing to investigate the $CD8^+$ TCRβ repertoires of AS patients at high resolution, and compared them with $CD8^+$ TCRβ repertoires of healthy individuals. From the peripheral blood of 5 patients and 2 healthy controls, more than 28 million TCRβ sequences were generated. I comprehensively profiled and performed comparative analyses for these repertoires. The number of patient samples was too small to allow us to distinguish TCRβ sequences that were enriched in patients versus controls, but the work does demonstrate the use of the *aimseqtk* package in an actual research setting.

Of note, we are currently collaborating with Faham *et al.* (at Sequenta) to analyze a larger dataset of AS TCR repertoires (140 donors). However, the results of this collaboration will be reported elsewhere.

4.3 Results

In this section, I describe an open-source software package for comprehensive assessments and comparative analyses of TCR repertoires called "Adaptive IMmunoSequencing ToolKit" or *aimseqtk*. Given an input set of TCR sequencing data samples and accompanied meta information (e.g disease status), the *aimseqtk* package provides a thorough collection of analyses covering all major applications of the field: repertoire signature profiling and comparison, clone tracking, and public (or condition-associated) clone identification. In addition, the *aimseqtk* package offers a unique function that searches existing publications for homologous clones of any set of clones of interest. This function is especially applicable to condition-associated clones, aiming to assist the exploration for literature context and consistency of the identified association. Users choose which analyses to perform and the *aimseqtk* package returns the appropriate summary statistic tables and figures, and statistical test results when relevant. In the following subsections, I use the TCR sequencing data of 5 AS patients and 2 healthy donors as demonstration to describe in details the various functions of the software.

4.3.1 Sample Information

A summary of the sample information and sequencing results is listed in Table 4.1. Except for one patient (AS5), all donors carried HLA-B27. $CD8^+$ T-cells were purified from the peripheral blood of each sample (see Methods), then the DNA was extracted and TCRβ genes were amplified and sequenced by Adaptive Biotechnologies

Sample	Sex	Age	HLA-B27 Status	ERAP1 rs30187	ERAP1 rs10050860	BASDAI	Disease Duration	Treatment	Sequences	Clones
AS1	M	25	+	−+	−+	2.6	5	Enbrel	132,970	5,673,480
AS2	M	17	+	−+	++	3	5	NSAIDS	181,611	4,310,157
AS3	M	55	+	−−	++	4	5	Humira	150,496	5,834,002
AS4	M	24	+	−+	++	2.3	2	NSAIDS	12,407	261,775
AS5	M	20	−	++	++	4.6	13	NSAIDS	20,112	2,861,525
H1	M	58	+	−−	−−	N/A	N/A	N/A	56,018	6,019,470
H2	F	59	+	−+	++	N/A	N/A	N/A	51,657	3,842,153

Table 4.1: Sample summary. Rows: samples. Columns: 'Sample': sample ID, 'Gender': 'M' for male and 'F' for female, 'Age': number of years old, 'HLA-B27 Status': '+': sample is from $B27^+$ donor, '−': sample is from $B27^-$ donor, 'ERAP1 rs30187': '++', '−+' and '−−': sample is from donor carrying two risked alleles, one risked allele, and no risked allele at ERAP1 AS-associated SNP rs30187 [Evans et al., 2011a], respectively, 'ERAP rs10050860': similarly to'ERAP1 rs30187' but for ERAP1 AS-associated SNP rs10050860 [Evans et al., 2011a] (see Appendix Section C.2 for ERAP1 allele calling methods), 'BASDI': Bath Ankylosing Spondylitis Disease Activity Index, 'Disease Duration': number of years since the onset of the disease, 'Sequences': number of total productive sequences, 'Clones': number of total productive clones. 'N/A': not applicable. '.': missing information.

Corporation.

4.3.2 Preprocessing: Input Formats, Data Filtering and Downsampling

The standard practice for handling TCR sequencing data is to run the raw data produced by the sequencing machines through the error correction and mapping software. This software takes care of mapping reads, merging reads into clones, correcting potential sequencing errors and amplification bias and inferring junctional information. This step is essential and the available software has proven to be sufficient and robust

(see Section 4.2.3.4). The outputs of this step are typically files in a delimited format that is specific to the software used. The *aimseqtk* package is designed to take in these files as inputs and currently supports all the major formats: Adaptive Biotechnologies, Sequenta, iRepertoire, miTCR and its own format.

The example dataset used in this chapter was mapped and error-corrected using Adaptive Biotechnologies' ImmunoSeq standard procedure [Robins et al., 2010]. The gene names and CDR3 regions were defined following the IMGT nomenclature [Lefranc et al., 2009]. A distinct TCRβ amino acid sequence is referred to as a clone, and the number of copies of that particular sequence is its size or abundance.

The first stage of the *aimseqtk* package is preprocessing input data. Productive and non-productive clones are separated. Depending on their study design, users have the option to filter out clones based on the minimum and/or maximum clone frequencies and/or clone size (number of reads). For comparative analyses, ideally the experiment should be designed and carefully executed such that each sample has the same amount of starting cells and that the sequencing coverage is consistent across all samples. In practice however, due to various factors, these conditions are often not met, which results in sample size differences. To avoid biases introduced by these differences, larger samples can be reduced to be comparable with the smaller ones via the down-sampling function.

Normally, down-sampling involves reducing all the samples to the size of the smallest one. However, if there are more reads than needed to exhaustively sequence a sample, this approach does not account for the effect of sequencing saturation, or over-

sequencing. For example, in Figure 4.2, sample B has a total of nine million reads, but approximately only one million reads are needed to detect all sequences present in the sample. Sample A is smaller than B, with a total of five million reads, and has yet to reach its saturation. Comparing these two samples at five million reads, which is the size of the smaller sample A, sample A appears to be more diverse than sample B, with many more clones. That comparison, however, is inaccurate as it compares a well saturated sample B with an unsaturated sample A. The more appropriate comparison is at one million reads, before either sample passes its saturation point. In that comparison, sample B is less diverse than A.

Therefore, it is important to examine the data before determining the standard size to which the samples are normalized to. The standard size should be chosen based on the sample saturation points instead of the smallest sample size. To assist this, the rarefaction analysis function, which randomly resamples each repertoire over an increasing set of sizes (x-axis) and counts the number of unique clones each sampling contains (y-axis), produces an overview look of the data (Figure 4.2).

Figure 4.3 shows the rarefaction analyses of the AS and healthy repertoires. Based on this I pick two sampling sizes for this dataset: ten thousand sequences, the point before any of the samples saturated, and one million sequences, the point before all but the two smallest samples (AS4 and AS5) saturated. The *aimseqtk* package randomly selects ten thousand or one million sequences from each sample, performs analyses, and repeats the process 100 times. The average statistics of these samplings are reported in following sections. Of note, samplings of larger sizes than one million are also analyzed

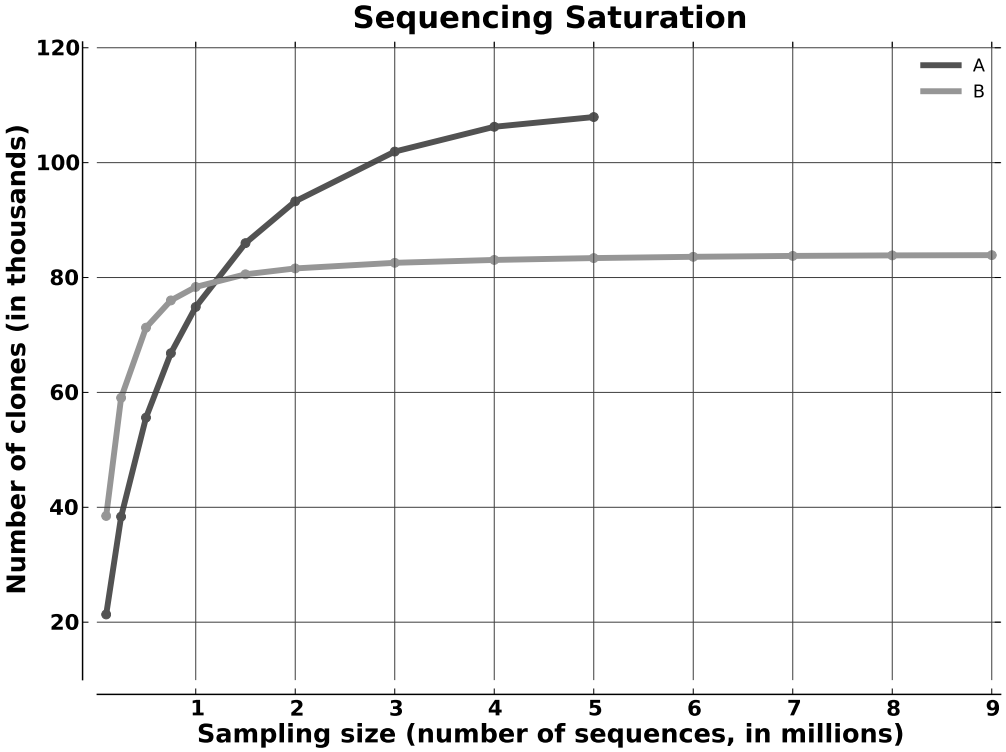

Figure 4.2: An illustration of sequencing saturation. Samples A and B are healthy samples obtained from Adaptive Biotechnologies

and found to yield similar results to those that obtained with one million sequences. Users may adjust the sampling sizes, the default setting is no sampling.

4.3.3 Repertoire Properties Profiling and Comparisons

The *aimseqtk* package provides comprehensive assessments of individual repertoires as well as comparative analyses of different repertoire groups. The list of incorporated assessments includes diversity, similarity, clonality, CDR3 length distribution, gene-segment usage and junctional insertions and deletions, each of which is described

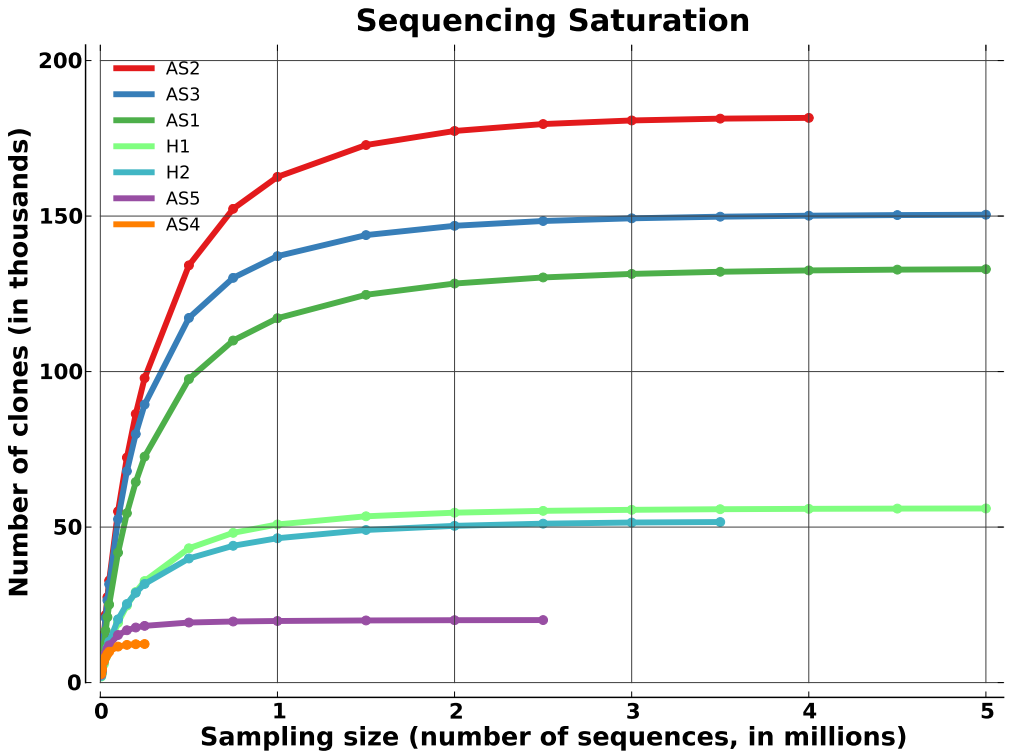

Figure 4.3: Rarefaction analyses of all samples.

in further detail below.

4.3.3.1 Diversity

Repertoire diversity (number of distinct clones and distribution of clone sizes) is important to the immune system's ability to protect the body against the vast number of antigens present in surrounding environments. For each repertoire, the *aimseqtk* package computes various diversity indices (Table 4.2) and performs Wilcoxon signed-rank (or Wilcoxon rank-sum or Mann-Whitney U) tests to compare groups of matched

(or unmatched) samples (Figure 4.4).

Table 4.2 and Appendix Table C.1 show that consistently across all the indices, the AS repertoires are more diverse than the controls. The number of clones and the Fisher Alpha index measure the species richness (number of distinct clones) while the Simpson and the Shannon indices measure species richness integrated together with species abundance (clone size). Out of one million sequences, all the AS repertoires have more than 110,000 clones, which is more than double the numbers of clones in the controls, 46,408 and 50,854. Similarly, Fisher Alpha, Simpson, and Shannon indices all indicate higher diversity in the AS samples than in the controls. The order of the samples from largest to smallest diversity is: AS2, AS3, AS1, AS4, AS5, followed by the controls H1 and H2, which closely resemble each other's diversity.

Since the dataset here is small, it is straightforward to investigate each sample individually. For larger datasets with many more samples, diversity differences among groups may be better summarized by the box plots as in Figure 4.4, one box plot per group showing the diversity distribution of samples in the group. Figure 4.4 shows the apparent higher number of distinct clones in the AS group compared with the Healthy group. Similar plots may be generated for other indices (one plot each) and statistical significances are reported (e.g Table 4.3).

4.3.3.2 Similarity

When comparing different groups of TCR repertoires, two common questions are whether within-group repertoires are more similar than between-group repertoires or

Sample	Unique Clones	Simpson	Shannon	Fisher Alpha
AS1	117,158 ± 100	0.999 ± 0.000	9.806 ± 0.002	34,431 ± 41
AS2	162,600 ± 102	1.000 ± 0.000	11.101 ± 0.001	55,066 ± 50
AS3	137,117 ± 87	0.999 ± 0.000	10.609 ± 0.002	43,002 ± 39
H1	50,849 ± 63	0.988 ± 0.000	7.190 ± 0.003	11,318 ± 17
H2	46,414 ± 55	0.985 ± 0.000	7.223 ± 0.003	10,072 ± 15

Table 4.2: Sample diversity indices of one million sequence samplings show that the AS repertoires are more diverse than the healthy repertoires. The rows represent the samples. The columns include: 'Sample' is the sample ID, 'Unique Clones' is the number of unique clones, 'Simpson', 'Shannon', 'Fisher Alpha' are the Simpson, the Shannon and the Fisher Alpha indices. Each cell contains the average and standard deviation of 100 samplings.

Diversity Index	Group 1	Group 2	p value	Mean 1 ± Std 1	Mean 2 ± Std 2
Fisher Alpha	Healthy	AS	0.017	14900.499 ± 980.063	62423.224 ± 10815.586
Unique Clones	Healthy	AS	0.009	62866.000 ± 3172.000	176098.667 ± 20483.268
Shannon	Healthy	AS	0.005	7.427 ± 0.010	10.715 ± 0.498
Simpson	Healthy	AS	0.004	0.988 ± 0.002	0.999 ± 4.08e-04

Table 4.3: Diversity index group comparisons show that the AS samples are more diverse than the healthy samples.

whether repertoire similarity correlates with an attribute of interest. There exist different measurements of similarity, among which I selected three popular ones to incorporate into the *aimseqtk* package: number of common clones, chao index [Chao et al., 2006] and horn index [Horn, 1966]. The software computes the similarity indices of choice for all pairs of input samples and returns a matrix table as well as a heatmap plot (Appendix Figure C.6) for each index.

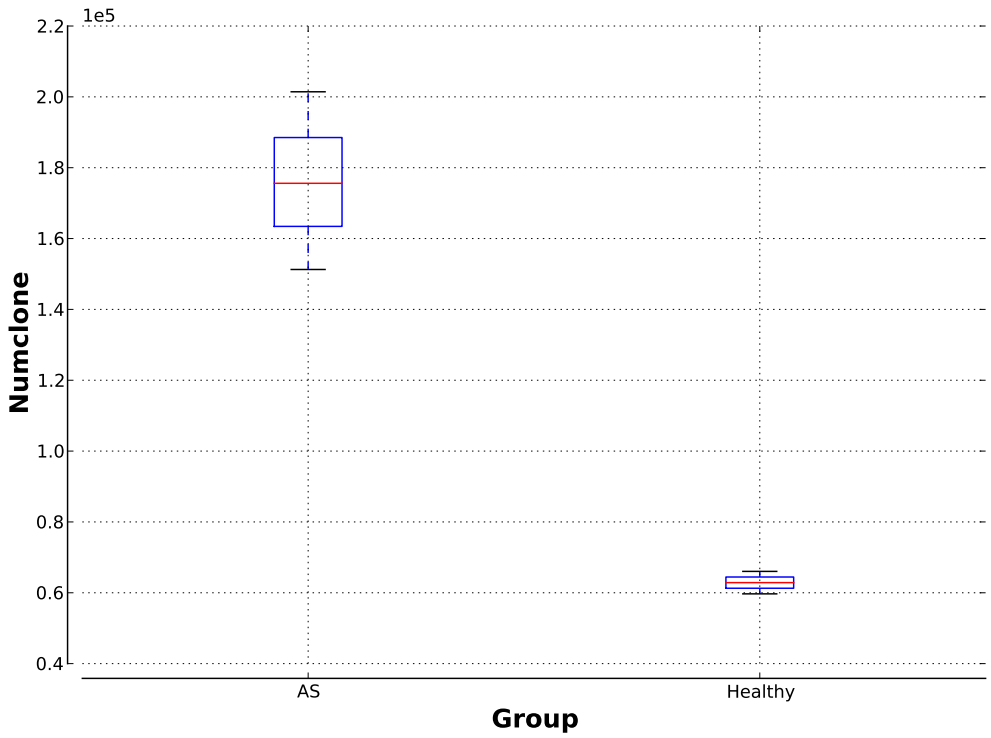

Figure 4.4: AS samples have more number of distinct clones (one million sequence samplings). 'Numclone' is the number of distinct clones each sample has. The red horizontal line marks the group median while the box's bottom and top boundaries mark the group 25th and 75th quartiles.

For any two groups A and B, statistical significances are reported if differences in similarity exist (Figure 4.5 and Table 4.4). Pairs of samples are separated into three categories: within-group pairs for group A, within-group pairs for group B, and between-group pairs. If pairs of samples within group A are observed to have higher similarity than do pairs belonging to the other two categories, the observation can be an indication/evidence of some common factors that are the results of, or have resulted in, the condition of group A, such as similar immune exposures, similar abnormalities

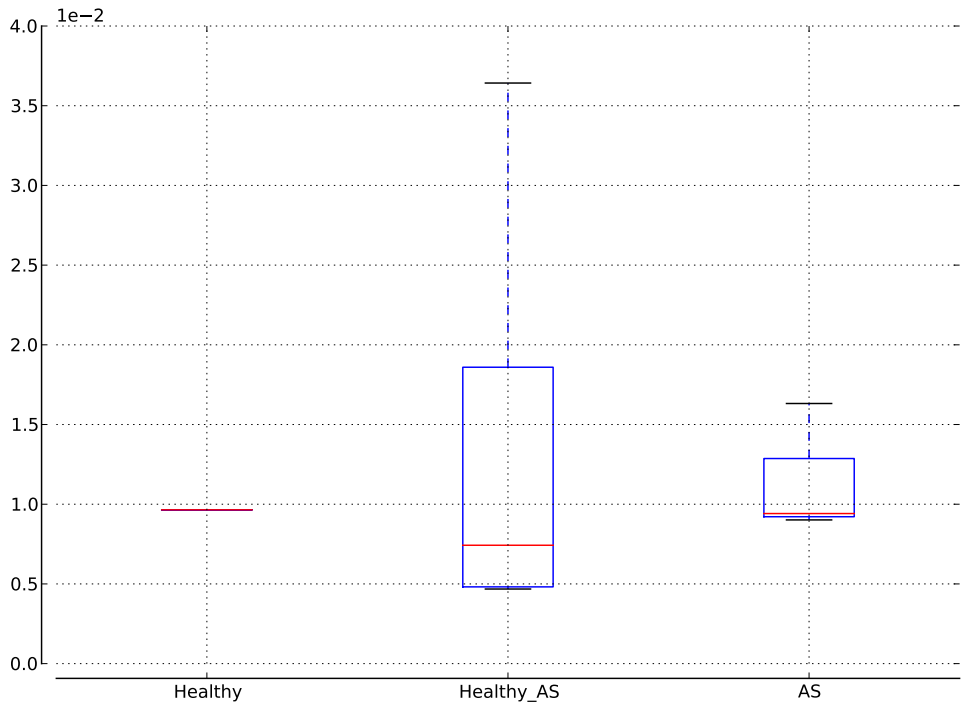

Figure 4.5: The Chao similarity index summary shows no significant differences among different groups. The x-axis shows the sample pair categories. The y-axis shows the Chao similarity index. The 'Healthy_AS' category contains all pairs of one Healthy sample and one AS sample. The 'AS' category contains all AS pairs. 'Category 1' and 'Category 2' are the two categories being compared.

in the recombination process or similar abnormalities in the tolerance system.

High clonal overlap of the repertoires can be an indication of similar immune experiences, such as common infections or immune experiences that are involved with a specific disease. To assess whether or not the AS repertoires have high overlaps with each other, I compare the overlaps of the AS repertoires with the overlaps of the control repertoires. An overlap is defined as the number of clones that the repertoires shared.

Category 1	Category 2	p value	Mean 1 ± Std 1	Mean 2 ± Std 2
Healthy	Healthy_AS	-0.292	0.010 +/- 0.000	0.014 +/- 0.012
Healthy_AS	AS	0.270	0.014 +/- 0.012	0.012 +/- 0.003
Healthy	AS	-0.408	0.010 +/- 0.000	0.012 +/- 0.003

Table 4.4: The Chao similarity index comparisons show no significant differences among different groups. Each row represents a pairwise comparison the group categories. The 'Healthy' category contains all sample pairs within the Healthy group. The 'Healthy_AS' category contains all pairs of one Healthy sample and one AS sample. The 'AS' category contains all AS pairs. 'Category 1' and 'Category 2' are the two categories being compared. 'p-value' shows the significance of the comparison. 'Mean 1' and 'Mean 2' are the average Chao similarity index of categories 1 and 2, respectively. 'Std 1' and 'Std 2' are the standard deviations of each category.

A clone is defined to be shared by two or more samples if it is observed in all of those samples.

To avoid biases introduced by differences in sample diversity (higher diversity, or more clones, increases the chance of clone sharing), I normalize the samples so that all samples have an equal number of clones (see Methods). Figure 4.6 shows the numbers of shared clones of all pairs of samples. Pairs of patients do not share more clones than pairs of controls or pairs of a patient and a control. However, $B27^+$ samples have more overlaps with each other than with the $B27^-$ sample.

4.3.3.3 Clonality

Clonality is important for understanding repertoire clonal dominance. The *aimseqtk* package computes two types of clone size distribution: the proportion of total clones as a function of clone size and the proportion of total sequences as a function of

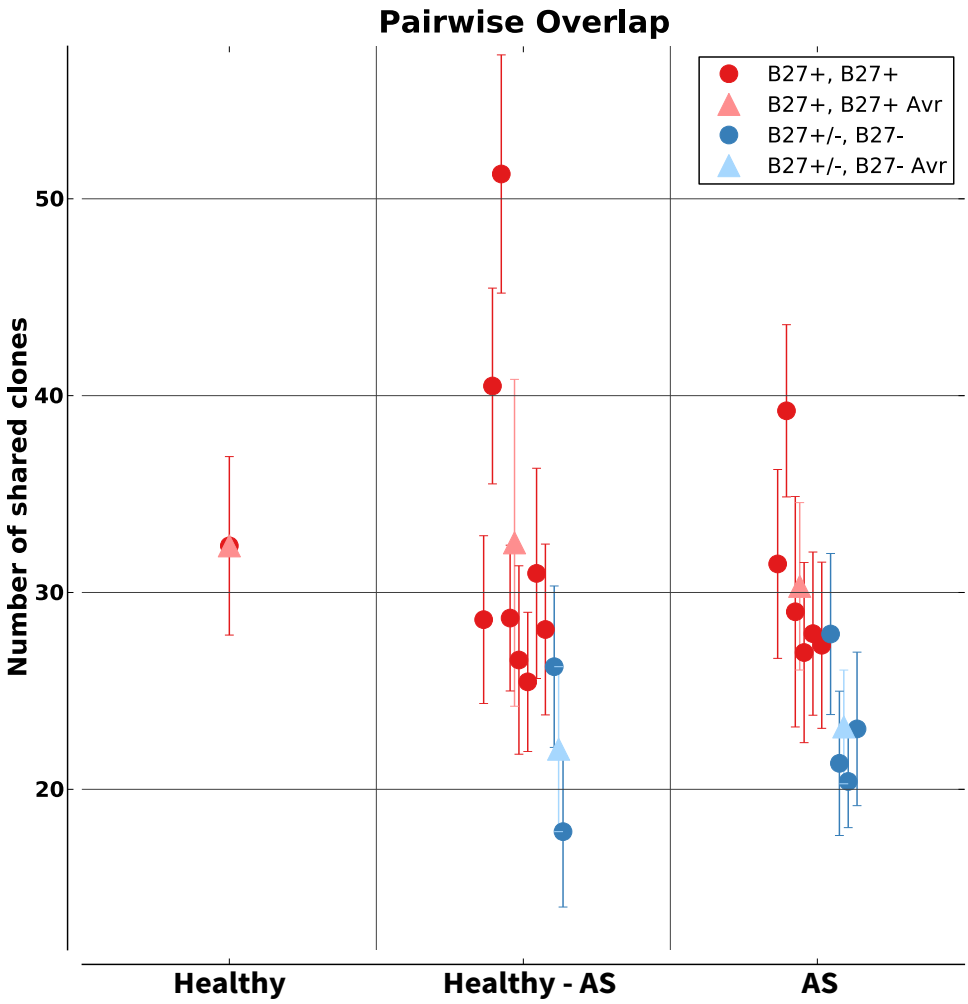

Figure 4.6: Sample pairwise overlaps, 12,000 clone samplings, show no higher overlaps among B27$^+$ AS repertoires in comparison with B27$^+$ healthy repertoires but higher overlaps among B27$^+$ repertoires in comparison with the B27$^-$ repertoire. 'B27+, B27+': both samples are B27+ (red), 'B27+/-, B27-': at least one sample is B27- (blue). Each circle shows the sampling average (of 100 samplings total) and the corresponding vertical bar shows the sampling standard deviation. 'B27+, B27+ Avr': average of all 'B27+, B27+' pairs within the category, 'B27+/-, B27- Avr': Average of all 'B27+/-, B27-' pairs within the category. The vertical bar of the triangles show the standard deviations of all pairwise overlap within the category. 'Healthy': pairs of two healthy control samples. 'Healthy - AS': pairs of one healthy control and one AS patient samples. 'AS - AS': pairs of two AS patient samples.

clone size (Figure 4.7 a, b). Additionally, an abundance summary of the largest clones of each sample helps to quickly assess how dominant the clonal expansions are in the sample (Figure 4.7 c). The size, or abundance, of a clone is defined as the percentage of total sequences belonging to that clone.

The healthy H1 and H2 clone size distributions are highly concentrated, with many highly dominant clonal expansions, while the B27$^+$ AS repertoires AS1, AS2, AS3 and AS4 are spread out and did not have highly dominant expansions (Figure 4.7 and Appendix Figures C.1, C.2, C.3). Figure 4.7A and Appendix Figure C.1A show the distributions of clones across different sizes of the samples. These distributions follow the expected negative exponential distribution [Sepúlveda et al., 2010]. The majority of clones ($> 70\%$) of each sample have very low frequencies (< 10 sequences or 0.001% for one million sequence samplings).

Relative to the AS samples, H1 and H2 have high proportions of large clones. In particular, out of one million sequences, H1 and H2 each had 0.024% clones with frequencies of $\geq 1\%$, 20 times larger than did AS1 (0.0009%), AS2 (0%), AS3 (0.0015%). In H1 and H2, the larger clones, even though exponentially less in number than the smaller ones, account for an equal if not a higher proportion of total sequences (Figure 4.7B). The 0.024% clones with frequencies $\geq 1\%$ account for 30% of the total sequences. In contrast, in the AS samples AS1 (1.68%), AS2 (0%), AS3 (4.66%) and AS4 (11.31%), large clones contribute significantly less to each repertoire. The B27$^-$ AS sample AS5, unlike the other ones, is neither spread out nor does it have many highly expanded clones. Instead, it has a hybrid state with two highly expanded clones followed by a

relatively even distribution. Multiple clonal expansions in H1 and H2 explain the lower diversity of these repertoires, while the absence of extremely dominant clones in the AS samples reflects the evenness of these repertoires and explained the higher diversity.

4.3.3.4 CDR3 Length Distribution

CDR3 length distribution is a standard analysis reported in most TCR studies, with preferential length usage often mentioned in traditional (spectra-type) autoimmune TCR comparative works [Miles et al., 2011]. The *aimseqtk* package computes and compares CDR3 length distributions of total clones and of total sequences of the samples (Figure 4.8 and Appendix Figure C.5, respectively). For each length, statistical significance (after Bonferroni correction) is reported if there exists preferential usage in a specific group over another. The median lengths of different groups are also compared and significant shifts in the distribution are reported.

The length distributions of clones are highly similar for all samples, there is no preferential length usage in the AS repertoires (Figure 4.8). The distributions follow the Gaussian distribution and are similar to CDR3 length distributions of healthy T-cell repertoires previously reported [Wang et al., 2010, Robins et al., 2009, Warren et al., 2011]. The length distributions of total sequences, however, are different in different samples and reflect the sample clonal expansions (Appendix Figure C.5).

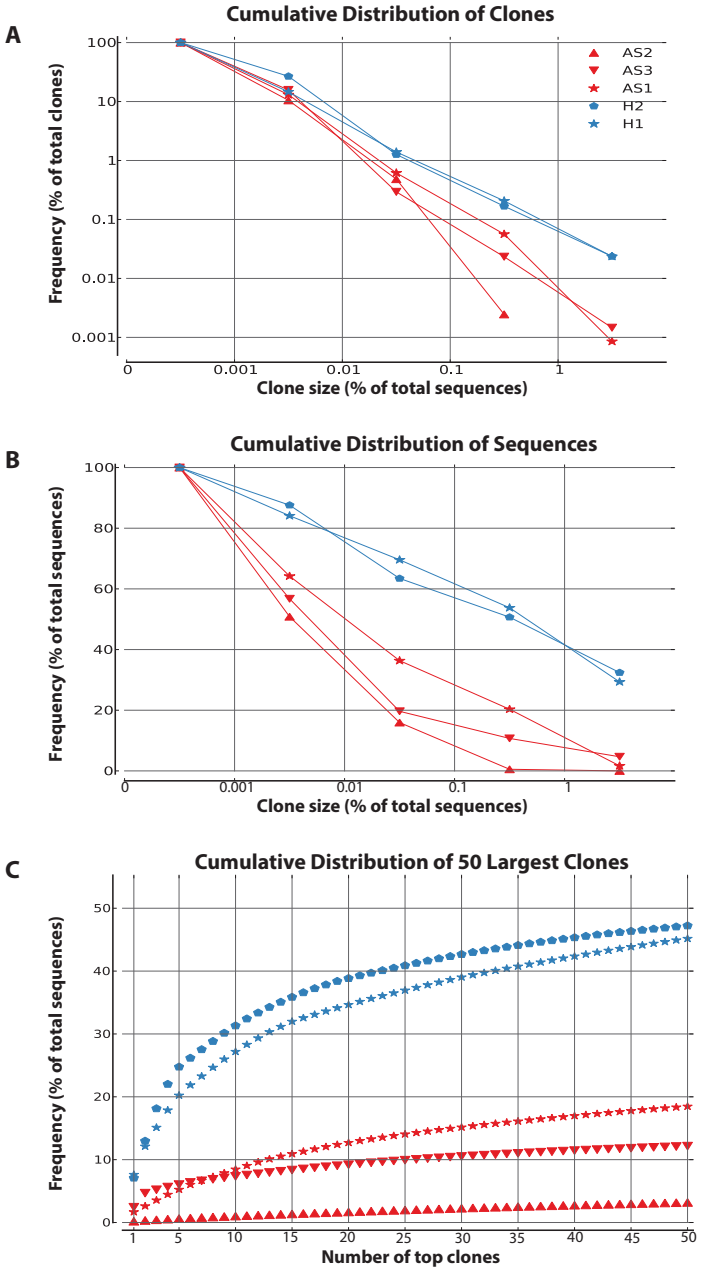

Figure 4.7: Cumulative clone size distributions of one million sequence samplings show no highly dominant clonal expansions in B27$^+$ AS repertoires. (A) Cumulative distribution of clones. (B) Cumulative distribution of sequences. (C) Cumulative distribution of sample 50 largest clones. AS samples are in red and healthy samples are in blue. In (A) and (B), each data point represents the proportion of total clones (A) or sequences (B) with frequencies $\geq$ 0%, 0.001%, 0.01%, 0.1% or 1%.

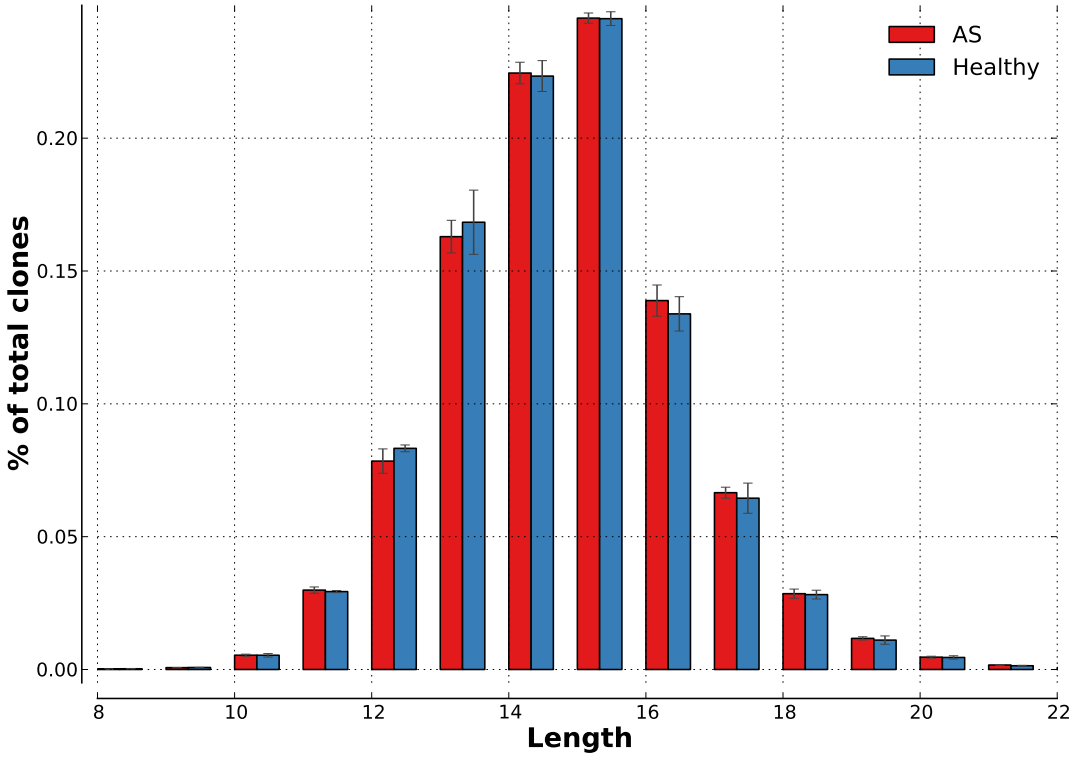

Figure 4.8: CDR3 length distributions of distinct clones.

4.3.3.5 Recombination Model: Gene-segment Usage and Junctional Insertions and Deletions

Let R_θ be the recombination event that results in a TCR nucleotide sequence θ, R_θ involves: the selected gene segments (V_θ, D_θ and J_θ), the number of deleted nucleotides of each gene segment ($delV_\theta$, $del5D_\theta$, $del3D_\theta$, $delJ_\theta$), and the inserted nucleotides at the VD ($insVD_\theta$) and DJ ($insDJ_\theta$) junctions. The generative probability

P_G of a recombination event R_θ is [Murugan et al., 2012]:

$$\begin{aligned} P_G(R_\theta) =& P(V_\theta) * P(D_\theta, J_\theta) * \\ & P(delV_\theta|V_\theta) * P(del5D_\theta, del3D_\theta|D_\theta) * P(delJ_\theta|J) * \\ & P(insVD_\theta) * P(insDJ_\theta) \end{aligned}$$

Joint or conditional probabilities are used respectively when the involved variables are correlated or dependent [Murugan et al., 2012]. $P(V_\theta)$ is the probability that V_θ gets selected for recombination among all possible V genes. $P(D_\theta, J_\theta)$ is the joint probability that D_θ and J_θ get selected among all possible D and J pairs. $P(delV_\theta|V_\theta)$ is the conditional probability that $delV_\theta$ nucleotides get deleted given that V is V_θ. $P(del5D_\theta, del3D_\theta|D_\theta)$ and $P(delJ_\theta|J)$ are similar, but for D and J. $P(insVD_\theta)$ ($P(insDJ_\theta)$) is the probability that the nucleotide sequence $insVD_\theta$ ($insDJ_\theta$) get inserted at the V_D (D_J) junction:

$$P(insVD_\theta) = P(L(insVD)) * \prod_{i=1}^{L(insVD)} P_{VD}(x_i|x_{i-1})$$

$$P(insDJ_\theta) = P(L(insDJ)) * \prod_{i=1}^{L(insDJ)} P_{DJ}(y_i|y_{i+1})$$

P_{VD} and P_{DJ} are conditional probabilities of inserting a specific nucleotide (x_i or y_i) given the immediately 5' or 3' nucleotide (x_{i-1} or y_{i+1}). $L(insVD)$ and $L(insDJ)$ are lengths of inserted sequences at VD and DJ junctions.

The various distributions account for different aspects of the recombination process and together, influence the shape of the repertoire. The *aimseqtk* package computes each of these usage distributions for each sample and checks for differences (if

they exist) in each distribution among different groups. The software provides two types of usage profiles: the sequence profile, in which the usage (frequency) of each attribute (e.g each gene segment or each recombination) is based on the number of sequences carrying that attribute (e.g gene or recombination), and the clone profile, in which the usage is computed using the number of clones. Sampling effect aside, the clone profile is expected to reflect the antigen-naïve state of the repertoire while the sequence profile reflects its immune encounters and expansions of clones. In the following examples, the clone profile is used.

For the current dataset, the overall variable region gene-segment usage of the samples is consistent with previously reported healthy repertoires [Robins et al., 2010] (Appendix Tables C.2, C.3, C.4). The repertoires use both Ds, all Js, all possible D-J combinations, 86-88% of total Vs, 83-93% of all possible V-J and 80-91% V-D-J recombinations. Figures 4.9 and 4.10 show the usage of J and V genes in the two groups AS and Healthy. There is no significant differential usage in V, D, J, D-J or V-J detected between AS and healthy repertoires. Principal component analyses are also performed for each attribute, an example is shown in Figure 4.11.

Figure 4.12 shows the distribution of the number of nucleotides inserted at the DJ junction represented by box plots. In comparison with the healthy repertoires, the AS repertoires show slightly higher proportions of longer inserted lengths ($>= 4$ nucleotides) and lower proportions of shorter lengths (< 4).

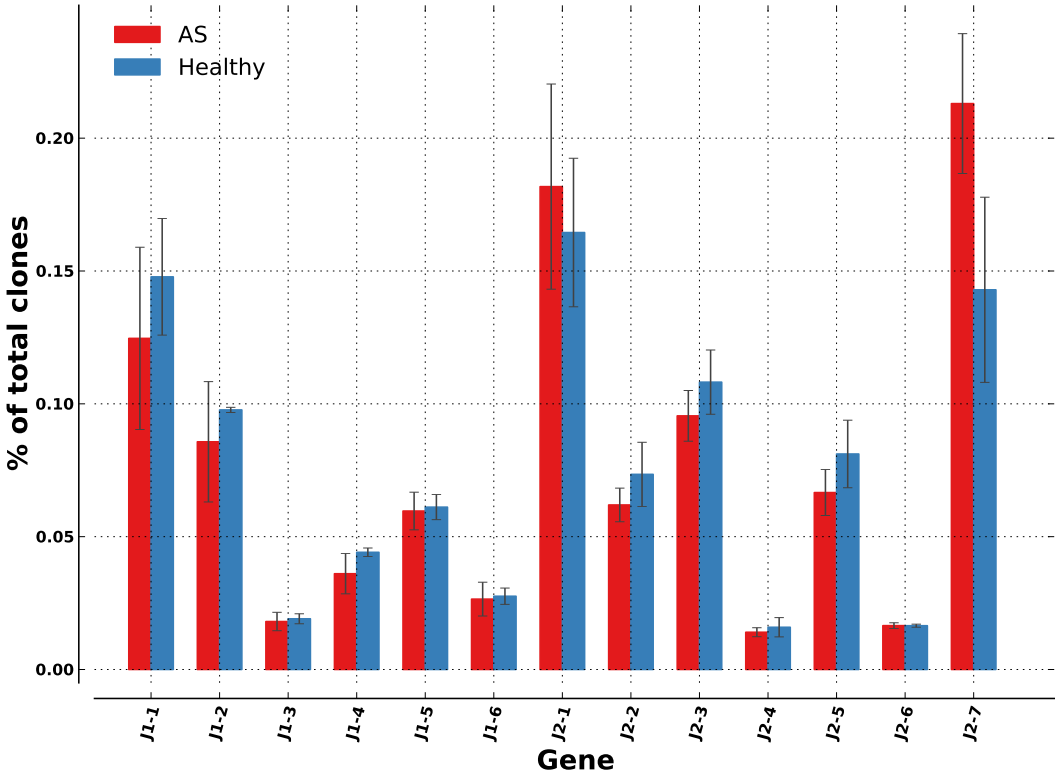

Figure 4.9: No significant preferential TRBJ usage is detected between AS and Healthy repertoires.

4.3.4 Clone Tracking

The *aimseqtk* package tracks abundances of clones across different samples and groups (e.g different time points, tissues, or conditions). Users can specify a list of specific clones to track, examples are cancerous clones, autologous clones, clones injected into hosts, or clones known to be associated with certain bacteria or disease of interest. In addition, the software identifies highly expanded clones in each sample (the default

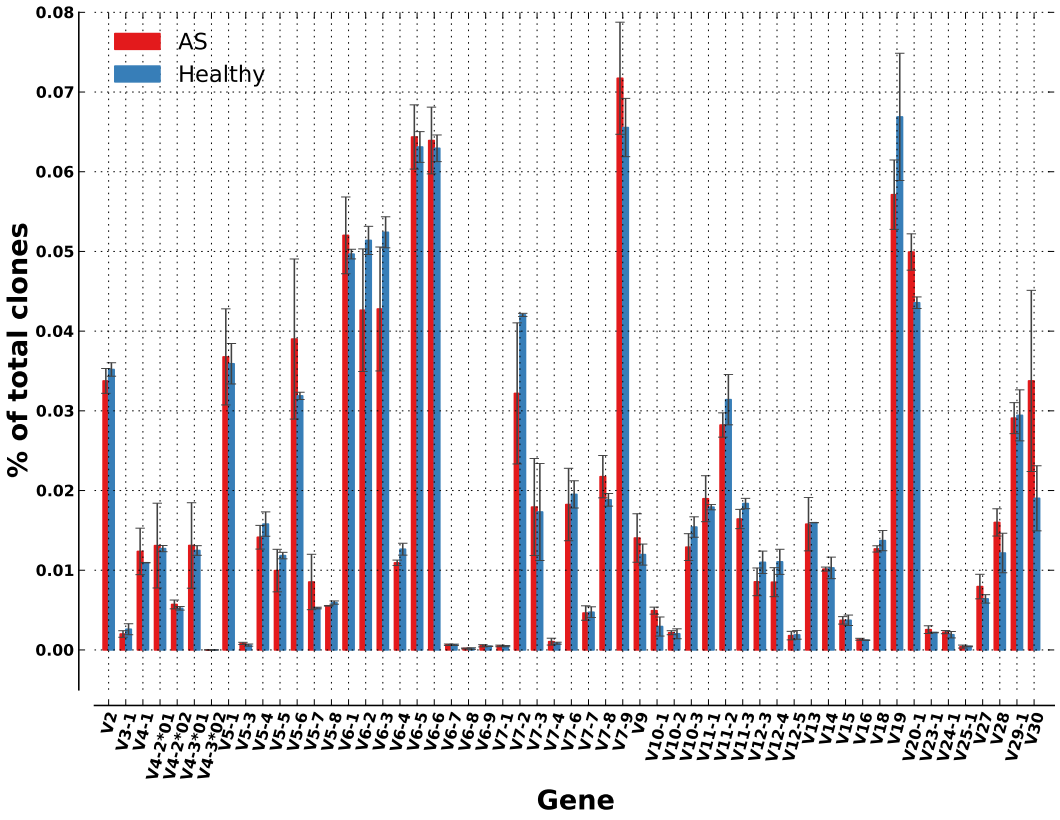

Figure 4.10: No significant preferential TRBV usage is detected between AS and Healthy repertoires.

minimum frequency is 1%) and tracks their abundances in other samples (Table 4.5, Figure 4.13).

Table 4.5 shows a summary of the abundances of each sample's most expanded clones across all samples of both groups AS and Healthy. In this limited set of samples, there is no common clonal expansion that is shared by two or more samples. Figure 4.13 shows an example box plot that tracks the abundance of one of sample H2's highly

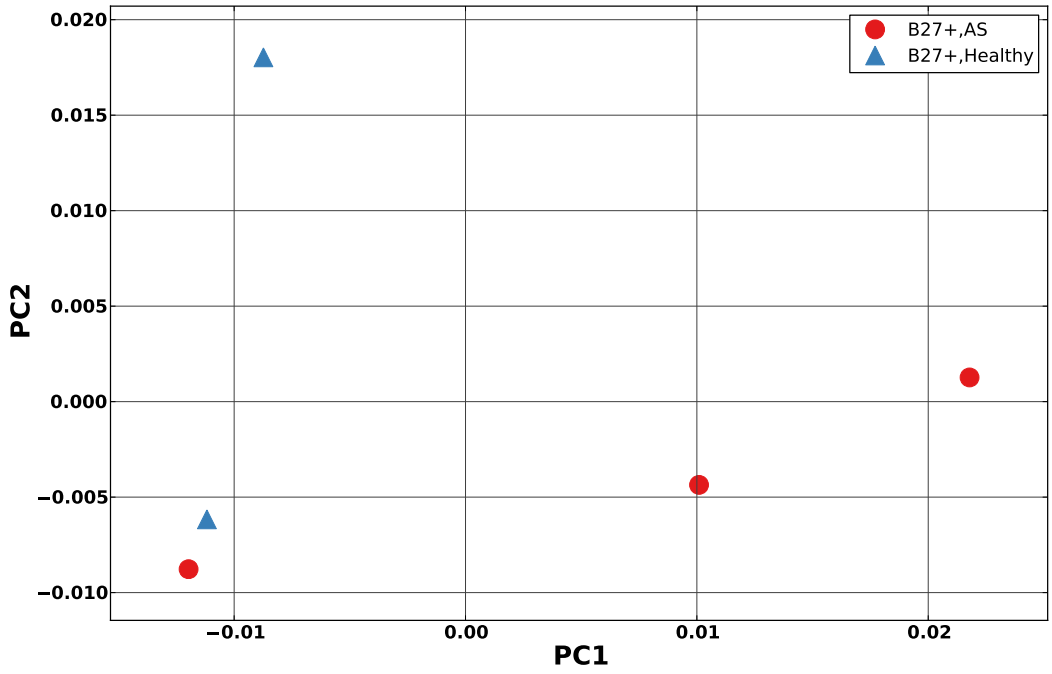

Figure 4.11: Principal component analysis of VJ usage shows no discrimination between AS and Healthy repertoires.

expanded clone TRBV11-1_CASSLFYSPYNEQFF_TRBJ2-1 (2.77%), which is either absent, or present with a very low abundance, in other samples. With large dataset, this type of plot is useful to visualize differential abundance of a specific clone between different groups, such as abundance of a cancerous clone before (typically >= 90%) and after treatment (very small or if successful, absent).

4.3.5 Public Clones

The *aimseqtk* package helps to identify clones that are dominantly present in a particular group of samples (in-group) compared with other groups (out-groups).

V	Sequence	J	AS	Healthy
TRBV2	CASNTRLPNTEAFF	TRBJ1-1	Absent	1.3170
TRBV4-1	CASSQEGSSYNEQFF	TRBJ2-1	Absent	1.3836
TRBV4-3	CASSQDEGTGANVLTF	TRBJ2-6	1.6722,0.0008	Absent
TRBV4-3	CASSQDSGSGANVLTF	TRBJ2-6	2.429	0.0006
TRBV6-4	CASSDTLAADSNEQFF	TRBJ2-1	Absent	7.6397
TRBV6-5	CASKGTGDTDTQYF	TRBJ2-3	5.1687e-05	4.5018
TRBV6-5	CASRQGRGAFF	TRBJ1-1	Absent	3.884
TRBV7-2	CASSLTLGSEQFF	TRBJ2-1	Absent	1.4139
TRBV7-8	CASSLWSDYPYEQYF	TRBJ2-7	0.0002	5.1564,0.0003
TRBV7-9	CASTLSGMNTEAFF	TRBJ1-1	Absent	7.1387,0.0002
TRBV9	CASSPSPKLAAHEQYF	TRBJ2-7	0.0003	1.1028
TRBV10-3	CAIRPGLAGIQETQYF	TRBJ2-5	Absent	1.4114
TRBV10-3	CATIPQGQNEQFF	TRBJ2-1	Absent	2.3883
TRBV11-1	CASSLFYSPYNEQFF	TRBJ2-1	8.8578e-05	2.7687,0.0002
TRBV14	CASSHLYTEAFF	TRBJ1-1	Absent	5.9167
TRBV15	CATSRERTGGGEKLFF	TRBJ1-4	2.2664	0.0007
TRBV19	CASSISVSQPQHF	TRBJ1-5	Absent	1.3011
TRBV19	CASSITSGAYNEQFF	TRBJ2-1	Absent	0.0054
TRBV20-1	CSASRQGGGEQFF	TRBJ2-1	Absent	3.003
TRBV29-1	CSARILDHEQFF	TRBJ2-1	0.0001	2.7582
TRBV29-1	CSLEWGNNEQFF	TRBJ2-1	Absent	1.6535
TRBV29-1	CSVEDNRGPYEQYF	TRBJ2-7	Absent	1.1608
TRBV30	CAWGGDDSYEQYF	TRBJ2-7	3.4457e-05	1.0409

Table 4.5: Tracking abundances of each sample's most expanded clones (frequency >= 1%) shows that there is no common clonal expansions shared by the samples of both groups. Columns "AS" and "Healthy" show the (comma separated) abundances of each clone (in percentage) in samples of group AS and group Healthy, respectively. Zero frequencies are not reported unless the clone is absent in all samples of a specific group, then it is marked as 'Absent' in that group.

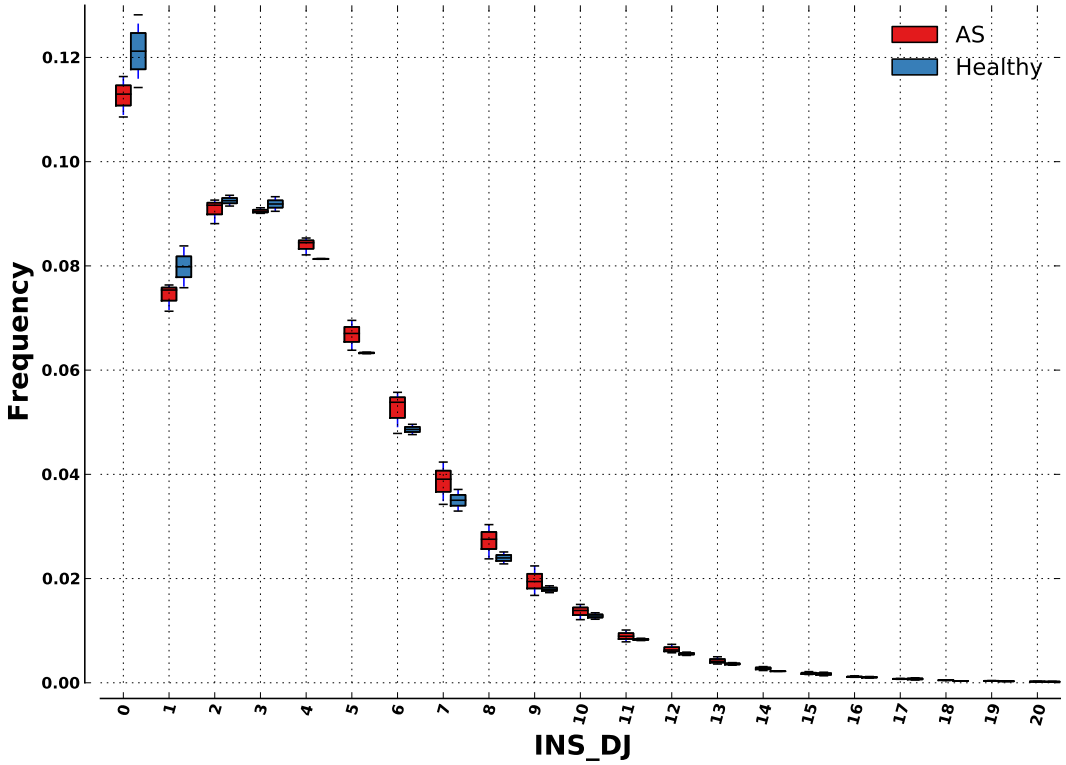

Figure 4.12: Length distributions of DJ inserted nucleotide sequences show slightly higher proportions of longer lengths in AS repertoires.

Users can adjust the minimum proportion of the in-group samples and the maximum proportion of the out-group samples that contain each clone. Fisher's Exact test and multiple testing correction are used to filter for clones that are significantly associated with the in-group. There were 34 clones that were present in at least 75% of the AS (4 samples) and absent in all the healthy samples. None of these clones was significantly associated with AS. This was expected because the number of studied samples was

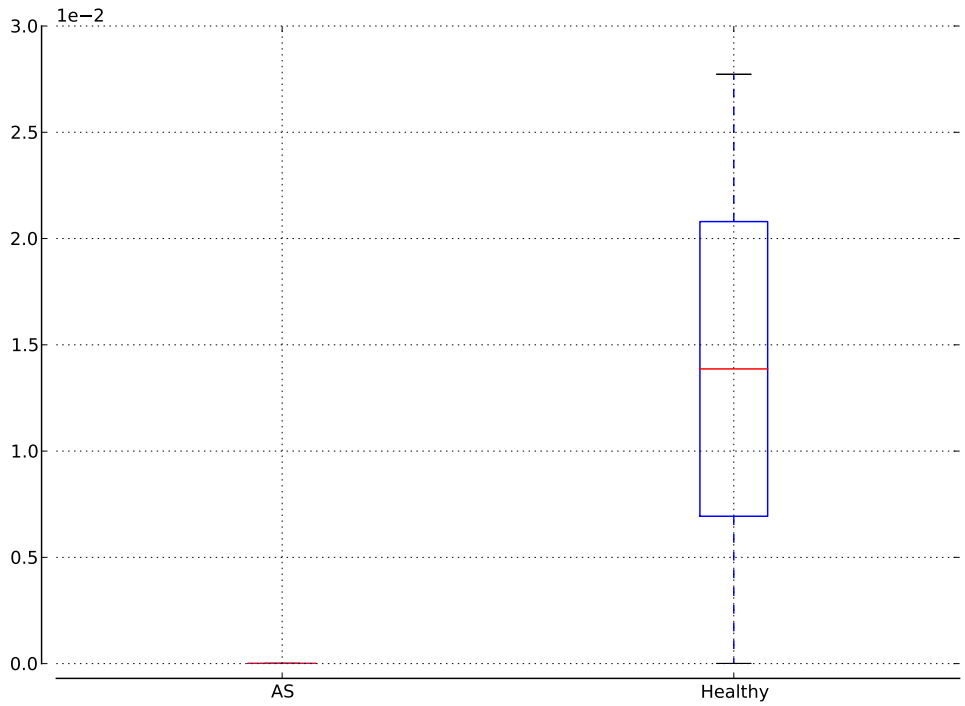

Figure 4.13: Tracking abundances of one of the expanded clones, clone TRBV11-1_CASSLFYSPYNEQFF_TRBJ2-1, of sample H2. The clone is absent in all AS samples.

small.

4.3.6 Publication Mining

The UCSC Browser publication mining tool [Haeussler et al., 2011] is incorporated into the *aimseqtk* package to assist researchers in investigating the published literature for information on the set of clones that they are interested in. This tool collects sequences from databases like IMGT and the full text of published research articles and runs BLAST [McGinnis and Madden, 2004] to compare these to the CDR3

sequences of the input samples. Homologous sequences together with information of the studies in which they are included are reported. The *aimseqtk* package includes various filtering criteria, such as sequence identity and keywords, to help narrowing down to relevant studies, of which users can conduct further exploration for literature context and/or literature validation. In the following subsections, I demonstrate the tool's utilization in inquiring information for 1/ expanded clones of each sample and 2/ AS-public clones.

4.3.6.1 Expanded Clones of AS Patients are Reported in Previous Studies on Autoimmune Diseases

BLAST is used to search the 10 most expanded clones of each sample (70 clones total) in databases and published articles. A larger proportion of the AS clones (12%) have autoimmune-related ('relevant') matches than of the healthy clones (5%) (Table 4.6). Of the 12% (6) patient clones with relevant matches, four are from sample AS5, one from AS1 and one from AS4 (Table 4.7). The matches in the literature come from patients with Rheumatoid Arthritis (RA), Reactive Arthritis (ReA) and Multiple Sclerosis (MS). Repeating the search with the 10 most expanded clones of 10 healthy samples from published high-throughput sequencing studies (100 clones total) (Appendix Section C.1) confirms this observation. 6% of these clones have relevant matches, which is consistent with the 5% and is half of the 12%. This suggests the hypothesis that patient expanded clones are more similar to clones of related autoimmune diseases than are expanded clones from healthy individuals. However, a larger patient

cohort and set of matched normals would be required to check this.

4.3.6.2 Clones Shared Among AS Patients are Reported in Previous Studies on Autoimmune Diseases

BLAST search of clones that are shared by four or more patients and and that are absent in the controls results in relevant matches for 13 of those clones (Table 4.8). Out of the 13 clones, there are four that have high similarity to sequences previously observed in AS or Spondyloarthropathy (SpA). The other clones with matches include two in ReA, two in RA, six in MS and one autoreactive clone specific to melanoma tumor antigen glycoprotein 100. These autoimmune diseases, especially SpA, which is a group of joint diseases including AS (and RA), and ReA, which is HLA-B27 associated, are related to AS. The three clones with matches in SpA (**CASS**M**GQGYEQYF**, **CASS**I**GQGAYEQYF** and **CASS**L**GQGAYEQYF**) are similar to each other. In Section 4.3.5, I pointed out that these clones were not significantly associated with AS, which was expected given the small sample size. I show here, as an example that if they were, their sequence similarity with each other and with previously reported SpA clones would provide further support of their involvements in SpA disease mechanisms.

4.3.7 Clustering

Clustering analyses help to assess sequence homology of clones within a sample and across different samples. Using a greedy algorithm, the software clusters together clones of all samples that have identical V and J genes and the same CDR3 length, and

Group	Total	Matches	% Matches/Total
AS	50	6	12
Healthy	20	1	5
Healthy, published	100	6	6

Table 4.6: Literature search summary of 10 most expanded clones from each sample: larger proportion of the AS expanded clones have high sequence similarity with clones previously reported in related autoimmune diseases than of the healthy expanded clones. Columns: 'Group': sample group, 'Total': total number of clones included in the analysis, 'Matches': number of clones with matches of $\geq 85\%$ identity from previous autoimmunity studies, '% Matches/Total': the percentage of the total clones with matches. 'Healthy': healthy samples H1 and H2,'AS': AS samples AS1-AS5, 'Healthy, published': healthy samples from previously published TCRβ high-throughput sequencing studies.

shared at least 85% (approximately ≤ 2 amino acids difference, this cutoff is adjustable) CDR3 sequence identity (see Methods). For the AS dataset, a total of 605,271 clones results in 481,148 clusters, 77.8% (374,388) of which are singletons (clones that do not cluster with any other clone).

A clone is labeled 'expanded' if its frequency is 0.1% or greater (this cutoff is adjustable as well). There are 382 expanded clones total. I classify clusters with multiple expanded clones as *confident* candidates for further analyses to identify potential disease-associated clones. I require a *confident* cluster to have at least three expanded clones. Out of more than 100 thousand non-singleton clusters, one is identified as *confident*.

Further analyses of this *confident* cluster shows a strong indication of antigen selection. The cluster includes 18 clones carrying the motif TRBV4-3 - **CASSQD*G*GANVLTF** - TRBJ2-6 (Table 4.9). The 18 clones consist of 7 abun-

Clones			Samples							Matches		
V	**CDR3**	**J**	**AS1**	**AS2**	**AS3**	**AS4**	**AS5**	**H1**	**H2**	**CDR3**	**Alignment**	**Disease**
7-6	CASSVTGANNEQFF	2-1			41		395928	48		CASSFTGAGNEQFF	CASS TGA NEQFF	RA
19	CASSIFGEQFF	2-1,2-7		5	1427		45504			CASSIYGEQFF	CASSI+GEQFF	RA
7-8	CASSKGTANYGYTF	1-2					15962			CASSAGTENYGYTF	CASS GT NYGYTF	RA
										CASSLGTENYGYTF	CASS GT NYGYTF	RA
9	CASSVGGRATGELFF	2-2			108		7266			CASSVGGRRSTGELFF	CASSVGGR +TGELFF	MS
5-1	CASSPGLSNTEAFF	1-1	44286	17						CASSPGQVNTEAFF	CASSPG NTEAFF	MS
7-9	CASHSVDRHQETQYF	2-5				4408				CAS-SVDRFQETQYF	CAS SVDR QETQYF	ReA

Table 4.7: A summary of AS patients' most expanded clones with high sequence similarity to clones reported by previous disease studies. 'Clones': TRBV, CDR3 and TRBJ information of the clone. 'Samples': number of sequences of a specific clone each sample has, an empty cell indicates the absence of the clone. 'Matches': CDR3 sequences reported by previous autoimmune disease studies: 'CDR3': the CDR3 amino acid sequence, 'Alignment': alignment of the matched CDR3 sequence with the clone CDR3 sequence (the residue is shown if it is conserved, '+' indicates a positive score for the match, '-' indicates a gap, and a space indicates a mismatch), 'Disease': the disease of the host of the match. 'ReA': Reactive Arthritis (IMGT, unpublished, Genbank accession number AJ296351, AJ296374), 'RA': Rheumatoid Arthritis [Striebich et al., 1998, Li et al., 1994], 'MS': Multiple Sclerosis [Biegler et al., 2006, Babbe et al., 2000]

Clones			Samples							Matches		
V	**CDR3**	**J**	**AS1**	**AS2**	**AS3**	**AS4**	**AS5**	**H1**	**H2**	**CDR3**	**Alignment**	**Disease**
5-1	CASSLGGGYEQYF	2-7	19	7	62		24			CASSLGRGYEQYF	CASSLG GYEQYF	MS
5-1	CASSLAGGPYNEQFF	2-1	53	31	33	4				CASRLAGGPFNEQFF	CAS LAGGP+NEQFF	MS
5-6	**CASSLGQGAYEQYF**	**2-7**	**10**	**30**	**39**		**108**			**CASSV-QGAYEQYF**	**CASS+ QGAYEQYF**	**SpA**
5-6	CASSLGGSSYEQYF	2-7	9	10	245	15				CASSLGSSSYEQYF	CASSLG SSYEQYF	ReA
6-3	CASSSYNEQFF	2-1	28	28	17	8				CASSSVNEQFF	CASSS NEQFF	AS
6-5,6-6	CASSYSGGNTEAFF	1-1	26	17	46	2				CASSYSRKNTEAFF	CASSYS NTEAFF	MS
										CASSLRGGNTEAFF	CASS GGNTEAFF	MS
6-5,6-6	CASSYGDSSYEQYF	2-7	3	42	50	16				CASSLGSSSYEQYF	CASS G SSYEQYF	ReA
7-9	CASSLGGDEQFF	2-1	101	12	158		9			CASSLRGDEQFF	CASSL GDEQFF	RA
										CASSLRGDEQFF	CASSL GDEQFF	RA
12-4	CASSLGGNEQFF	2-1	17	71	82		17			CASSPGGNEQFF	CASS GGNEQFF	$HER2_{369}$
13	CASSLGRGTYEQYF	2-7	2	10	37		31			CASSLGRG-YEQYF	CASSLGRG YEQYF	MS
19	**CASSMGQGYEQYF**	**2-7**	**12**	**6**	**23**	**7**				**CASSFGQGYEQYF**	**CASS GQGYEQYF**	**SpA**
19	**CASSIGQGAYEQYF**	**2-7**	**34**	**26**	**17**		**34**			**CASSV-QGAYEQYF**	**CASS+ QGAYEQYF**	**SpA**
										CASSIGQENYEQYF	CASSIGQ YEQYF	RA
										CASSIGTGAHEQYF	CASSIG GA+EQYF	MS
20-1	CSAR-GTASYEQYF	2-7	9	11	14		243			CSARAGGASYEQYF	CSAR G ASYEQYF	MS

Table 4.8: A summary of clones that are shared by at least four patients and absent or present with low frequencies ($\leq$ 10 sequences) in the controls and have high sequence similarity ($\geq$85% identity, approximately $\leq$ 2 amino acids difference) with previously reported CDR3 sequences of autoimmune diseases. 'Clones': TRBV, CDR3 and TRBJ information of the clone. 'Samples': number of sequences of a specific clone each sample has, an empty cell indicates the absence of the clone. 'Matches': CDR3 sequences reported by previous autoimmune disease studies: 'CDR3': the CDR3 amino acid sequence, 'Alignment': alignment of the matched CDR3 sequence with the clone CDR3 sequence (the residue is shown if it is conserved, '+' indicates a positive score for the match, '-' indicates a gap, and a space indicates a mismatch), 'Disease': the disease of the host of the match. 'AS': Ankylosing Spondylitis [Duchmann et al., 2001], 'SpA': Spondyloarthropathy (IMGT, unpublished, Genbank accession numbers AY145777, AY145767), 'ReA': Reactive Arthritis (IMGT, unpublished, Genbank accession number AJ296361), 'RA': Rheumatoid Arthritis [Striebich et al., 1998, Li et al., 1994], 'MS': Multiple Sclerosis [Biegler et al., 2006, Babbe et al., 2000, Ristori et al., 2000], '$HER2_{369}$': allorestricted TCR with specificity for the HER2/neu-derived peptide 369 [Liang et al., 2010].

dant clones (with frequencies ranging from 0.01% to 2.42%) followed by 11 smaller ones. All 7 abundant clones are from two patient samples, AS1 and AS3, including the most expanded clone of AS1 (**CASSQD**E**G**T**GANVLTF**, 1.66%, 94,393 sequences) and the most expanded clone of AS3 (**CASSQD**S**G**S**GANVLTF**, 2.42%, 141,012 sequences). There are 3 clones in this confident cluster that are also present in the healthy samples, however only at very low frequencies (0.0006% - 0.002%). Multiple clonal expansions from two different patients sharing the same motif (**CASSQD***__G__***GANVLTF**) suggests that there may have been independent selection for this motif in the two patient repertoires and it may be AS-associated. If this result is confirmed in a sufficiently large dataset, then the two most expanded clones of AS1 and AS3, **CASSQD**E**G**T**GANVLTF** and **CASSQD**S**G**S**GANVLTF**, are potential candidates for further study.

4.4 Discussion

In this chapter, I described an open-source software package named *aimseqtk* for performing comprehensive assessments and comparative analyses of TCR repertoires. The *aimseqtk* package contains a comprehensive list of analyses, including normalization (downsampling), repertoire signature profiling and comparisons, clone tracking, public clones identification and publication mining. In the following, I will discuss the results obtained from applying the *aimseqtk* package to study the TCR data of AS patients and healthy individuals.

V	CDR3	J	Sample	Count	Frequency (%)
4-3	CASSQD**S**G**S**GANVLTF	2-6	**AS3**	**141012**	**2.417**
			H1	36	0.0006
4-3	CASSQD**E**G**T**GANVLTF	2-6	**AS1**	**94393**	**1.664**
			AS2	36	0.0008
4-3	CASSQD**G**G**S**GANVLTF	2-6	**AS1**	**9677**	**0.171**
4-3	CASSQD**Q**G**T**GANVLTF	2-6	**AS1**	**5199**	**0.092**
4-3	CASSQD**G**G**A**GANVLTF	2-6	**AS1**	**2599**	**0.046**
4-3	CASSQD**E**G**S**GANVLTF	2-6	**AS1**	**1336**	**0.024**
4-3	CASSQD**P**G**S**GANVLTF	2-6	**AS1**	**1101**	**0.019**
			AS3	108	0.002
4-3	CASSQD**A**G**A**GANVLTF	2-6	AS1	397	0.007
			AS2	162	0.004
4-3	CASSQD**R**G**S**GANVLTF	2-6	AS2	180	0.004
			H2	90	0.002
			AS1	54	0.001
4-3	CASSQD**R**G**T**GANVLTF	2-6	AS1	162	0.003
4-3	CASSQD**L**G**A**GANVLTF	2-6	AS3	90	0.0015
4-3	CASSQD**M**G**A**GANVLTF	2-6	AS1	54	0.001
4-3	CASSQD**L**G**T**GANVLTF	2-6	AS1	54	0.001
4-3	CASSQD**I**G**G**GANVLTF	2-6	H2	36	0.001
4-3	CASSQD**G**G**R**GANVLTF	2-6	AS2	36	0.0008
4-3	CASSQD**R**G**N**GANVLTF	2-6	AS1	36	0.0006
4-3	CASSQD**N**G**S**GANVLTF	2-6	AS1	36	0.0006
4-3	CASSQD**T**G**Y**GANVLTF	2-6	AS3	36	0.0006

Table 4.9: Cluster of homologous clones from multiple patients carrying the motif TRBV4-3 - CASSQD*G*GANVLTF - TRBJ2-6 ('*' can be replaced by different amino acids). Rows: clones in descending order of frequencies. The top two clones are the most expanded clones of two AS samples AS3 and AS1. Columns: 'V': TRBV gene, 'CDR3': the CDR3 amino acid sequence, 'J': TRBJ gene, 'Sample': sample(s) carrying the corresponding clone, 'Count': number of sequences of the corresponding clone each sample has, 'Frequency': frequency of the corresponding clone in each sample. Bolded are expanded clones with frequencies $\geq 0.01\%$.

AS predominantly affects the joints. Previous studies of AS and related autoimmune diseases suggest that the patient synovial fluid yields stronger disease-implicating signals than does peripheral blood [Atagunduz et al., 2005, Striebich et al., 1998]. However, detection of disease-associated signals in blood is desirable since obtaining synovial fluid is an invasive procedure. Thus, future diagnostic and prognostic applications of sequence analysis of disease-associated repertoires in patient blood are more likely to become widely used. There have been reports of clonal expansions and HLA-B27-restricted auto-reactive $CD8^+$ T-cells in blood samples from AS patients [Duchmann et al., 2001, Atagunduz et al., 2005]. However, these studies were limited by low-throughput methods with low sensitivity. With immunosqeuencing providing a sensitivity better than 0.001%, I reinvestigated the AS peripheral blood $CD8^+$ TCR repertoires to search for AS-associated signatures.

For the first time, high resolution snapshots of AS TCR repertoires are profiled and compared. The preliminary results show that except for having a higher diversity, AS peripheral blood $CD8^+$ TCR repertoires are overall similar to healthy ones. In particular, AS repertoires have similar CDR3 length distribution and similar gene-segment usage. They do not share more clones with each other than they do with healthy repertoires. No clone is detected to be significantly associated with AS. The lack of differentiation between AS and healthy repertoires may be attributed to the small sample size and consequently, the low statistical power. It is therefore critical to repeat the analyses with a larger dataset.

The higher diversity in AS is consistent with the absence of highly dominant

clonal expansions in the AS repertoires. In contrast, the repertoires of the two healthy samples had multiple highly dominant clonal expansions. Neither of the healthy donors had a known infection at the time of the blood draw. A previous blood draw of healthy donor H1 two years prior to the current blood draw had similar clonal expansions. 10 of 12 high frequency ($\geq$ 1%) clones of the current blood draw were present in the previous blood sample: 4 with frequencies $\geq$ 1% and 6 with frequencies $\geq$ 0.1%). This consistency of expanded clones confirmed that the highly dominant clonal expansions observed in the healthy samples were not the result of acute infection or other short-term antigen exposure. My analysis (Figure C.4) of the healthy repertoires from other high-throughput TCR sequencing studies showed that multiple highly dominant clonal expansions are not uncommon in healthy individuals. These clonal expansions in healthy individuals might reflect previous common infections. However, since only TCRβ chains are investigated (TCRα chains are missing), it is possible that these expansions reflect the host's preferential usage of the particular TCRβ chains that got expanded, regardless of the specificities of the T cells. It is an open, interesting scientific question awaiting to be answered once pair sequencing of TCR chains becomes successful.

Only one of the AS patients, AS5, had highly dominant clonal expansions. Patient AS5's repertoire had two extremely dominant expansions. These two expansions had much higher abundances (20.75% and 13.85%) than did the largest clones ($\leq$ 7.62%) in the healthy patient repertoires. These aberrant expansions might be related to the fact that at the time of the blood draw AS5 was experiencing an active flare-up and had a large swelling of the knee. The extreme frequencies suggest the possibility that

the two dominant clones in AS5 are disease-related.

Evidence of antigen selection was observed in the clonal expansions of the AS repertoires. I identified a cluster of homologous expanded clones from two different patients, including each patient's most expanded clone, sharing the motif TRBV4-3 - CASSQD*G*GANVLTF - TRBJ2-6. Being shared by multiple clonal expansions suggests a selective advantage of the motif. One possible scenario is that TCRs of the two most expanded clones (CASSQDSGSGANVLTF and CASSQDEGTGANVLTF) have the best fit for a particular peptide-MHC complex while less expanded clones have less-fit TCRs. The peptide that is involved may either belong to a common antigen in the environment or may be disease-associated. More samples are needed to clarify this.

The results of the AS-healthy TCR repertoire comparative analyses are preliminary due to the small dataset. However, they help demonstrating the variety of applications of the *aimseqtk* package. With a sufficiently large dataset, similar analyses can be repeated to search for evidence of antigen selection and/or identify potential disease-associated clones if exist.

The *aimseqtk* package has been tested with a larger dataset of ∼250 samples (from 140 donors), each sample having an average size of two million sequences and 200,000 clones. The most computationally expensive analyses are the pair-wise analyses for all pairs of samples, which in this example, involve approximately 12.5 billion (250 x 250 x 200,000) comparisons. To scale with such intensive analyses, the *aimseqtk* package optionally parallelizes all analyses, as well as takes advantage of the natural decomposition of the sequences by their gene-segment information. In light of the

Repertoire 10K project, which aims to sequence the TCR repertoires of 100 different diseases, 100 patients per disease, an abundance of TCR data will soon be available. In the near future, I plan to further scale the *aimseqtk* package to sufficiently handle analyses of thousands of TCR repertoires. In addition, for normalization purposes (down-sampling), currently the software provides rarefaction analyses to guide the users to pick an appropriate normalized size. I plan to incorporate mathematical models to improve this normalization process. Finally, I plan to integrate the *aimseqtk* package with the UCSC Immunobrowser to make the software more accessible to users without them having to download and run the software via command-lines.

4.5 Material and Methods

4.5.1 Implementation

The code is written in Python and available at `https://github.com/ngannguyen/aimseqtk.git`. The pipeline includes analyses on repertoire properties, such as diversity, clone size distribution, CDR3 length distribution, gene-segment usage, junctional insertion, junctional deletion and amino acid usage. Comparative analyses include monitoring changes of an individual repertoire over different time points or different conditions, comparing multiple repertoires for differences in any of the repertoire properties, and analyses of shared or persistent clones. In addition, the pipeline has the sampling option that performs the sampling process as described in the previous section. Multiple analyses can be run in parallel for efficiency.

4.5.2 Statistics

Diversity indices are calculated using the R *vegan* [Oksanen et al., 2012] package. Differences between groups are analyzed using the Wilcoxon signed-rank test for groups of matched samples and the Wilcoxon rank-sum (or Mann-Whitney U) for groups of unmatched samples. Alternatively, users can choose to use the Student's t test. All tests are computed using the python *scipy* package [Jones et al., 01].

4.5.3 Clustering

Clones are clustered using a greedy algorithm as described in [Edgar, 2010]. The default identity threshold is 85% and is adjustable. Gaps are not allowed in the alignments (only sequences of equal length are clustered together).

4.5.4 Publication mining

UCSC obtained permission from the publishers Elsevier (`http://www.elsevier.com/`) and the American Association of Immunology (`http://www.aai.org/`) to download more than 2 million research articles. We added more than 250,000 articles from the open-access archive PubmedCentral (`http://www.ncbi.nlm.nih.gov/pmc/`). The UCSC Browser publication mining pipeline is an automated pipeline that parses these articles in various formats, like PDF and XML, to raw text and then extracts the DNA or protein sequences in them [Haeussler et al., 2011]. We added sequences from databases such as IMGT [Lefranc, 2003] and Genbank [Benson et al., 2012]. DNA and protein sequences were then compared with BLAST

[McGinnis and Madden, 2004] against our sample CDR3 sequences.

In this study, I required a match to have 85% or higher identity (approximately $\leq$ 2 amino acids different) with the query. To focus on relevant matches to AS, I searched for matches from studies that had titles or abstracts containing one or more of the following keywords: *arthritis*, *ankylosing*, *spondy*, *autoreactive*, *autoantigen*, *reactive arthritis*, *rheumatoid arthritis*, *multiple sclerosis*, *self*, *cross-reactive*, *mimicry*, *synovial*, *crohn*, *psoriasis*, *inflammatory bowel disease*, *ibd*, *ulcerative colitis*, *uveitis*. I classified these matches as matches of previous autoimmunity studies. To reduce noise, I excluded high-throughput studies ($\geq$ 1000 sequences) in the search.

4.5.5 Sample collection, preparation and sequencing

4.5.5.1 Human subjects

Patients with definite AS were defined by the modified New York criteria for diagnosis of ankylosing spondylitis [van der Linden et al., 1984]. The two healthy control samples free of inflammatory disease were included in the study. Human subject characteristics are provided in Table 4.1. Subjects were recruited according to the Palo Alto Medical Foundation Institutional Review Board guidelines (Protocol #09-49) with an informed consent agreement.

4.5.5.2 $CD8^+$ T-cell isolation

PBMCs were isolated from freshly collected blood samples by Ficoll density centrifugation using Ficoll-Paque^{TM} PREMIUM from GE Healthcare Bio-Sciences AB

(Uppsala, Sweden). Cells were washed in calcium-free, magnesium-free DPBS (Mediatech, Inc., Manassas, VA) containing 2% fetal calf serum (Thermo Scientific Hyclone, Santa Clara, CA). $CD8^+$ T-cells were isolated with anti-human CD8 IMagTM (Becton Dickinson, San Jose, CA) according to the manufacturers instructions. Purity of $CD8^+$ T-cells was determined by flow cytometric analysis using monoclonal antibodies specific for CD8 (clone 32-M4, labeled with PE) and CD3 (clone HIT3a, labeled with FITC) obtained from Santa Cruz Biotechnology (Santa Cruz, CA). Flow cytometric analysis was performed on a BD LSR II flow cytometer (Becton Dickinson, San Jose, CA). Flow cytometry data analysis was performed with FlowJo data cytometric analysis tools (Tree Star, Inc., Ashland, OR).

4.5.5.3 DNA Isolation from $CD8^+$ T-cell

DNA was extracted from 3-4 x 106 cells using an Invitrogen Purelink gDNA mini kit (Lot: 1089136, Catalogue #: K1820-01) following manufacturers specific protocols. Samples were eluted in 100 μl of supplied elution buffer for a total yield of 0.5-3 μg of purified DNA. Extracted DNA samples were frozen and shipped to Adaptive Biotechnologies Corporation (Seattle, WA) for sequencing.

4.5.6 Sample samplings

4.5.6.1 Standardizing sample sizes

Except for the publication mining analyses, all analyses were done using the normalized samples. I picked two sizes for normalization: ten thousand sequences and

one million sequences. The analyses using samples normalized to one million sequences consisted of donors AS1, AS2, AS3, H1, and H2. The analyses using samples normalized to ten thousand sequences consisted of all donors (AS1-AS5, H1 and H2). The *aimseqtk* package randomly selected ten thousand or one million sequences from each sample, performed the analyses and computed the statistics, and this process was repeated 100 times. The average statistics of these samplings are reported in this study.

4.5.6.2 Samplings of clones

To normalize a sample to a subset of clones, we randomly selected that many clones from the sample using each clone's frequency as its probability to get selected. Similarly as in the above, we computed the statistics for the samplings, repeated the process 100 times, and reported the average and standard deviation of the results.

Chapter 5

Conclusion

In this dissertation, I presented three extensive efforts that aimed to enhance and facilitate genomic research: the construction of a pan-genome reference, the comparative assembly hub pipeline and the adaptive immunosequencing toolkit pipeline. In each respective effort, I illustrated the resulting algorithms and software by building and assessing the pan-genome reference for the human major histocompatibility complex region, constructing the *E. coli* comparative assembly hubs and updating the *E. coli* pan-genome, core genome and phylogenies, and comprehensively profiling and comparing $CD8^+$ TCR repertoires of AS patients to ones of healthy individuals.

It is inspiring and fascinating to reflect on the constant evolving of the world in general and on human capabilities in particular. To send a letter from one town to another, we have transitioned from depending on a messenger traveling on foot for many months to the convenience and instant delivery of electronic mails. To make a phone call, we have moved from depending on switchboard operators to simply getting out

the mobile phone from our pockets. To comprehend a human's genome, we have progressed from the sequencing efforts of an international collaboration between multiple research institutions, fifteen years and three billion dollars to those sequencing efforts of one individual, a few days and one thousand dollars. It is therefore not surprising that one day, going to space will just be another field trip, and most likely before then, our young scientists will be able to investigate the genome and biology of the organisms in their backyard. In fact, in recent years, the media has featured fascinating headlines on high-school students devising potential cures for cancer [Hartman, 2012] and making a genetic discovery of their own rare disease [Honeyman et al., 2014, Naggiar, 2014]. With sequencing costs continuing to decrease, we are moving forward into the direction of individualizing genomic research. I envision the day when an average family is able to reconstruct their pan-genome reference or ancestral reference, generate the browser collection of their genomes, exploring and performing comparative genomic and immunogenomic analyses on their own data as well as against publicly available data to understand their own digital footprint, health and diseases. Until that happens, I hope that the tools presented here help to enable researchers to analyze their data and to test their hypotheses and to facilitate research, from well-funded, human-oriented, medical-driven projects to basic research of less well investigated species, such as the common ant.

As brilliantly illustrated by Matt Might (Figure 5.1), I hope that the works in this dissertation have contributed a "dent" to the continuous quest of expanding human knowledge.

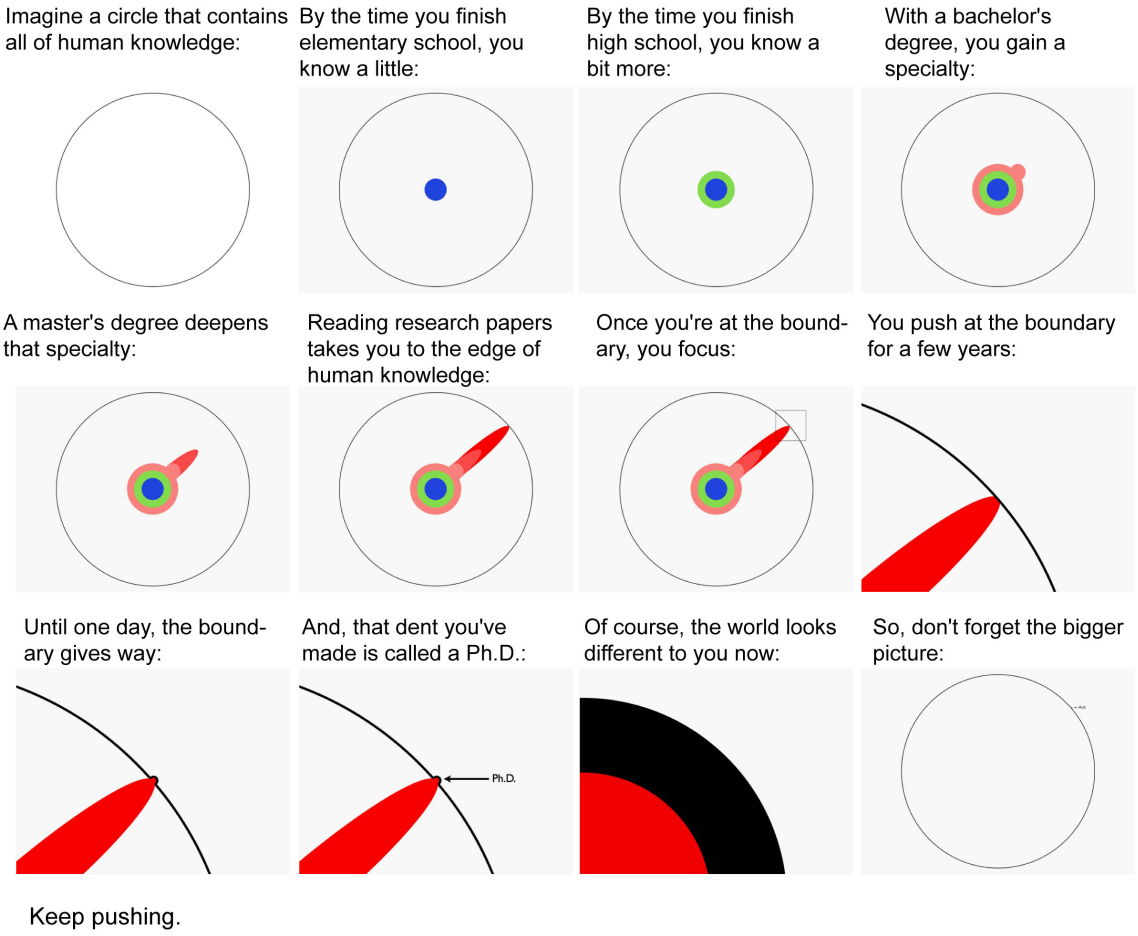

Figure 5.1: Matt Might's illustration of what a Ph.D. is (`http://matt.might.net/articles/phd-school-in-pictures/`).

Bibliography

[1000-Genomes-Project-Consortium, 2010] 1000-Genomes-Project-Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–73.

[10k-Community-of Scientists, 2009] 10k-Community-of Scientists, G., 2009. Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*, **100**(6):659–74.

[Abeel et al., 2012] Abeel, T. T., Van Parys, T. T., Saeys, Y. Y., Galagan, J. J., and Van de Peer, Y. Y., 2012. GenomeView: a next-generation genome browser. *Nucleic Acids Research*, **40**(2):e12–e12.

[Adams et al., 2000] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.*, 2000. The genome sequence of drosophila melanogaster. *Science*, **287**(5461):2185–95.

[Alföldi et al., 2011] Alföldi, J., Palma, F. D., Grabherr, M., Williams, C., Kong, L., Mauceli, E., Russell, P., Lowe, C. B., Glor, R. E., Jaffe, J. D., *et al.*, 2011. The genome

of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**(7366):587–91.

[Aparicio et al., 2002] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.*, 2002. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, **297**(5585):1301–10.

[Appel et al., 2004] Appel, H., Kuon, W., Kuhne, M., Wu, P., Kuhlmann, S., Kollnberger, S., Thiel, A., Bowness, P., Sieper, J., et al., *et al.*, 2004. Use of hla-b27 tetramers to identify low-frequency antigen-specific t cells in chlamydia-triggered reactive arthritis. *Arthritis Res. Ther*, **6**(6):521–534.

[Aradidopsis-Genome-Initiative, 2000] Aradidopsis-Genome-Initiative, 2000. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, **408**(6814):796–815.

[Atagunduz et al., 2005] Atagunduz, P., Appel, H., Kuon, W., Wu, P., Thiel, A., Kloetzel, P.-M., and Sieper, J., 2005. HLA-B27-restricted CD8+ T cell response to cartilage-derived self peptides in ankylosing spondylitis. *Arthritis & Rheumatism*, **52**(3):892–901.

[Babbe et al., 2000] Babbe, H. H., Roers, A. A., Waisman, A. A., Lassmann, H. H., Goebels, N. N., Hohlfeld, R. R., Friese, M. M., Schröder, R. R., Deckert, M. M., Schmidt, S. S., *et al.*, 2000. Clonal expansions of CD8(+) T cells dominate the T cell

infiltrate in active multiple sclerosis lesions as shown by micromanipulation and single cell polymerase chain reaction. *Journal of Experimental Medicine*, **192**(3):393–404.

[Bartlett and Stirling, 2003] Bartlett, J. M. S. and Stirling, D., 2003. A short history of the polymerase chain reaction. *Methods Mol Biol*, **226**:3–6.

[Belov et al., 2006] Belov, K., Deakin, J. E., Papenfuss, A. T., Baker, M. L., Melman, S. D., Siddle, H. V., Gouin, N., Goode, D. L., Sargeant, T. J., Robinson, M. D., *et al.*, 2006. Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biol*, **4**(3):e46.

[Benson et al., 2012] Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., and Sayers, E. W., 2012. Genbank. *Nucleic Acids Res*, **40**(1):D48–53.

[Bentley et al., 2008] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.*, 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218):53–9.

[Berard et al., 2009] Berard, S., Chateau, A., Chauve, C., Paul, C., and Tannier, E., 2009. Computation of perfect dcj rearrangement scenarios with linear and circular chromosomes. *Journal of Computational Biology*, **16**(10):1287–1309.

[Bertrand et al., 2009] Bertrand, D., Blanchette, M., and El-Mabrouk, N., 2009. Genetic map refinement using a comparative genomic approach. *J Comput Biol*, **16**(10):1475–86.

[Biegler et al., 2006] Biegler, B. W., Yan, S. X., Ortega, S. B., Tennakoon, D. K., Racke, M. K., and Karandikar, N. J., 2006. Glatiramer acetate (GA) therapy induces a focused, oligoclonal CD8+ T-cell repertoire in multiple sclerosis. *Journal of Neuroimmunology*, **180**(1-2):13–13.

[Blattner et al., 1997] Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.*, 1997. The complete genome sequence of escherichia coli k-12. *Science*, **277**(5331):1453–62.

[Bolotin et al., 2013] Bolotin, D. A., Shugay, M., Mamedov, I. Z., Putintseva, E. V., Turchaninova, M. A., Zvyagin, I. V., Britanova, O. V., and Chudakov, D. M., 2013. MiTCR: software for T-cell receptor sequencing data analysis. *Nature methods*, **10**(9):813–814.

[Brewerton et al., 1973] Brewerton, D., Hart, F., Nicholls, A., Caffrey, M., James, D., and Sturrock, R., 1973. Ankylosing spondylitis and HL-A 27. *The Lancet*, **301**(7809):904–907.

[Britanova et al., 2012] Britanova, O. V., Bochkova, A. G., Staroverov, D. B., Fedorenko, D. A., Bolotin, D. A., Mamedov, I. Z., Turchaninova, M. A., Putintseva, E. V., Kotlobay, A. A., Lukyanov, S., *et al.*, 2012. First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. *Bone Marrow Transplantation*, **47**(11):1479–1481.

[Burrows et al., 2013] Burrows, J. M., Rist, M. J., Miles, J. J., and Burrows, S. R., 2013. High frequency of herpesvirus-specific clonotypes in the human T cell repertoire can remain stable over decades with minimal turnover. *Journal of Virology*, **87**(1):697–700.

[Cantor and Cantor, 1969] Cantor, T. H. and Cantor, C. R., 1969. Evolution of protein molecules. *Mammalian protein metabolism*, **1**:22–123.

[Carlson et al., 2013] Carlson, C. S., Emerson, R. O., Sherwood, A. M., Desmarais, C., Chung, M.-W., Parsons, J. M., Steen, M. S., LaMadrid-Herrmannsfeldt, M. A., Williamson, D. W., Livingston, R. J., *et al.*, 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Communications*, **4**:2680–2680.

[Chao et al., 2006] Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J., 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, **62**(2):361–371.

[Chattopadhyay et al., 2009] Chattopadhyay, S., Weissman, S. J., Minin, V. N., Russo, T. A., Dykhuizen, D. E., and Sokurenko, E. V., 2009. High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *PNAS*, **106**(30):12412–12417.

[Chaudhuri et al., 2010] Chaudhuri, R. R., Sebaihia, M., Hobman, J. L., Webber, M. A., Leyton, D. L., Goldberg, M. D., Cunningham, A. F., Scott-Tucker, A., Ferguson, P. R., Thomas, C. M., *et al.*, 2010. Complete genome sequence and comparative

metabolic profiling of the prototypical enteroaggregative Escherichia coli strain 042. *PLoS ONE*, **5**(1):e8801–e8801.

[Chimpanzee-Sequencing-Analysis-Consortium, 2005] Chimpanzee-Sequencing-Analysis-Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**(7055):69–87.

[Church et al., 2011] Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., *et al.*, 2011. Modernizing reference genome assemblies. *PLoS Biol*, **9**(7):e1001091.

[Coffey et al., 2011] Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., Joyce, C. J., Leproust, E. M., Harrow, J., Hunt, S., *et al.*, 2011. The gencode exome: sequencing the complete human exome. *Eur J Hum Genet*, **19**(7):827–31.

[Consortium et al., 1998] Consortium, S. et al., 1998. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, **282**(5396):2012–8.

[Cooper et al., 2005] Cooper, G. M. G., Stone, E. A. E., Asimenos, G. G., Green, E. D. E., Batzoglou, S. S., and Sidow, A. A., 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genes & Development*, **15**(7):901–913.

[Darling et al., 2004] Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, **14**(7):1394–1403.

[Darling et al., 2010] Darling, A. E., Mau, B., and Perna, N. T., 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, **5**(6):e11147–e11147.

[Didelot et al., 2012] Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., and Crook, D. W., 2012. Transforming clinical microbiology with bacterial genome sequencing. *Nature reviews. Genetics*, **13**(9):601–612.

[Duchmann et al., 2001] Duchmann, R., Lambert, C., May, E., hler, T. H., and rker Hermann, E. M., 2001. CD4+ and CD8+ clonal T cell expansions indicate a role of antigens in ankylosing spondylitis; a study in HLA-B27+ monozygotic twins. *Clinical and Experimental Immunology*, **123**(2):315–322.

[Duchmann et al., 1996] Duchmann, R. R., May, E. E., Ackermann, B. B., Goergen, B. B., zum Büschenfelde, K. H. K. M., and Märker-Hermann, E. E., 1996. HLA-B27-restricted cytotoxic T lymphocyte responses to arthritogenic enterobacteria or self-antigens are dominated by closely related TCRBV gene segments. A study in patients with reactive arthritis. *Scandinavian Journal of Immunology*, **43**(1):101–108.

[Dulphy et al., 1999] Dulphy, N. N., Peyrat, M. A. M., Tieng, V. V., Douay, C. C., Rabian, C. C., Tamouza, R. R., Laoussadi, S. S., Berenbaum, F. F., Chabot, A. A., Bonneville, M. M., *et al.*, 1999. Common intra-articular T cell expansions in patients with reactive arthritis: identical beta-chain junctional sequences and cytotoxicity toward HLA-B27. *Journal of Immunology*, **162**(7):3830–3839.

[Earl et al., 2011] Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., *et al.*, 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res*, **21**(12):2224–41.

[Edgar, 2010] Edgar, R. C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19):2460–2461.

[Emerson et al., 2013] Emerson, R., Sherwood, A., Desmarais, C., Malhotra, S., Phippard, D., and Robins, H., 2013. Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. *Journal of Immunological Methods*, **391**(1-2):14–21.

[ENCODE-Project-Consortium et al., 2011] ENCODE-Project-Consortium, Myers, R. M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R. C., Bernstein, B. E., Gingeras, T. R., Kent, W. J., Birney, E., *et al.*, 2011. A user's guide to the encyclopedia of dna elements (encode). *PLoS Biol*, **9**(4):e1001046.

[Erdos and Rényi, 1960] Erdos, P. and Rényi, A., 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**:17–61.

[Evans et al., 2011a] Evans, D., Spencer, C., Pointon, J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., Stone, M., *et al.*, 2011a. Interaction between erap1 and hla-b27 in ankylosing spondylitis implicates peptide handling in the mechanism for hla-b27 in disease susceptibility. *Nature genetics*, **43**(8):761–767.

[Evans et al., 2011b] Evans, D. M., Spencer, C. C. A., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Opperman, U., Dilthey, A., Pirinen, M., Stone, M. A., *et al.*, 2011b. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics*, **43**(8):761–767.

[Fagin et al., 2002] Fagin, R., Kumar, R., and Sivakumar, D., 2002. Comparing Top k Lists. *SIAM J. DISCRETE MATH*, **17**(1):134–160.

[Fellay et al., 2007] Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., *et al.*, 2007. A whole-genome association study of major determinants for host control of hiv-1. *Science*, **317**(5840):944–7.

[Fernando et al., 2008] Fernando, M. M. A., Stevens, C. R., Walsh, E. C., Jager, P. L. D., Goyette, P., Plenge, R. M., Vyse, T. J., and Rioux, J. D., 2008. Defining the role of the mhc in autoimmunity: a review and pooled analysis. *PLoS Genet*, **4**(4):e1000024.

[Fiorillo et al., 2000] Fiorillo, M. T. M., Maragno, M. M., Butler, R. R., Dupuis, M. L. M., and Sorrentino, R. R., 2000. CD8(+) T-cell autoreactivity to an HLA-B27-restricted self-epitope correlates with ankylosing spondylitis. *Journal of Clinical Investigation*, **106**(1):47–53.

[Flicek et al., 2011] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen,

Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., *et al.*, 2011. Ensembl 2011. *Nucleic Acids Res*, **39**(Database issue):D800–6.

[Freeman et al., 2009] Freeman, J., Warren, R., Webb, J., Nelson, B., and Holt, R., 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research*, **19**(10):1817.

[Fujita et al., 2011] Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., *et al.*, 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, **39**(Database issue):D876–82.

[Fukiya et al., 2004] Fukiya, S., Mizoguchi, H., Tobe, T., and Mori, H., 2004. Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. *Journal of Bacteriology*, **186**(12):3911–3921.

[Ghai et al., 2004] Ghai, R., Hain, T., and Chakraborty, T., 2004. GenomeViz: visualizing microbial genomes. *BMC Bioinformatics*, **5**:198–198.

[Gnerre et al., 2011] Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., *et al.*, 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA*, **108**(4):1513–8.

[Goffeau et al., 1996] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B.,

Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.*, 1996. Life with 6000 genes. *Science*, **274**(5287):546, 563–7.

[Goodnow et al., 2005] Goodnow, C., Sprent, J., de St Groth, B., and Vinuesa, C., 2005. Cellular and genetic mechanisms of self tolerance and autoimmunity. *Nature*, **435**(7042):590–597.

[Grant et al., 2012] Grant, J. R., Arantes, A. S., and Stothard, P., 2012. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics*, **13**(1):202–202.

[Gregersen and Olsson, 2009] Gregersen, P. and Olsson, L., 2009. Recent advances in the genetics of autoimmune disease. *Annual review of immunology*, **27**:363–391.

[Griffiths et al., 1999] Griffiths, A. J. F., Miller, J. H., and Suzuki, D. T., 1999. An introduction to genetic analysis. .

[Grupp et al., 2013] Grupp, S. A., Kalos, M., Barrett, D., Aplenc, R., Porter, D. L., Rheingold, S. R., Teachey, D. T., Chew, A., Hauck, B., Wright, J. F., *et al.*, 2013. Chimeric Antigen Receptor–Modified T Cells for Acute Lymphoid Leukemia. *New England Journal of Medicine*, **368**(16):1509–1518.

[H et al., 2007] H, W., PF, H., TM, W., and DW, U., 2007. Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. *Genome Biology*, **8**(12):R267–R267.

[Haas et al., 2004] Haas, B. J., Delcher, A. L., Wortman, J. R., and Salzberg, S. L.,

2004. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**(18):3643–3646.

[Haeussler et al., 2011] Haeussler, M., Gerner, M., and Bergman, C. M., 2011. Annotating genes and genomes with DNA sequences extracted from biomedical articles. *Bioinformatics*, **27**(7):980–986.

[Hartman, 2012] Hartman, S., 2012. Teen Makes Genetic Discovery of Her Own Rare Cancer. *CBSNews.com*, `http://www.cbsnews.com/news/calif-hs-student-devises-possible-cancer-cure/`.

[Hayashi et al., 2006] Hayashi, K. K., Morooka, N. N., Yamamoto, Y. Y., Fujita, K. K., Isono, K. K., Choi, S. S., Ohtsubo, E. E., Baba, T. T., Wanner, B. L. B., Mori, H. H., *et al.*, 2006. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Molecular Systems Biology*, **2**:2006–0007.

[Hellsten et al., 2010] Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N. H., Shu, S., Taher, L., *et al.*, 2010. The genome of the western clawed frog xenopus tropicalis. *Science*, **328**(5978):633–6.

[Herbig et al., 2012] Herbig, A., Jäger, G., Battke, F., and Nieselt, K., 2012. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, **28**(12):i7–15.

[Hermann et al., 1993] Hermann, E. E., Yu, D. T. D., zum Büschenfelde, K. H. K. M., and Fleischer, B. B., 1993. HLA-B27-restricted CD8 T cells derived from synovial

fluids of patients with reactive arthritis and ankylosing spondylitis. *The Lancet*, **342**(8872):646–650.

[Hey and Harris, 1999] Hey, J. and Harris, E., 1999. Population bottlenecks and patterns of human polymorphism. *Mol Biol Evol*, **16**(10):1423–6.

[Hickey et al., 2013] Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D., 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics (Oxford, England)*, **29**(10):1341–1342.

[Hill and Harnish, 1981] Hill, C. W. and Harnish, B. W., 1981. Inversions between ribosomal RNA genes of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **78**(11):7069–7072.

[Hogquist et al., 2005] Hogquist, K., Baldwin, T., and Jameson, S., 2005. Central tolerance: learning self-control in the thymus. *Nature Reviews Immunology*, **5**(10):772–782.

[Holmes and Durbin, 1998] Holmes, I. and Durbin, R., 1998. Dynamic programming alignment accuracy. *J Comput Biol*, **5**(3):493–504.

[Honeyman et al., 2014] Honeyman, J. N., Simon, E. P., Robine, N., Chiaroni-Clarke, R., Darcy, D. G., Lim, I. I. P., Gleason, C. E., Murphy, J. M., Rosenberg, B. R., Teegan, L., *et al.*, 2014. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science*, **343**(6174):1010–1014.

[Horn, 1966] Horn, H. S., 1966. Measurement of'overlap'in comparative ecological studies. *Ann. Rev. Ecol. Syst.*, **5**:25–37.

[Horton et al., 2008] Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J., Forbes, S., Gilbert, J. G. R., Halls, K., Harrow, J. L., *et al.*, 2008. Variation analysis and gene annotation of eight mhc haplotypes: the mhc haplotype project. *Immunogenetics*, **60**(1):1–18.

[i5K Consortium, 2013] i5K Consortium, 2013. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *Journal of Heredity*, **104**(5):595–600.

[International-Chicken-Genome-Sequencing-Consortium, 2004] International-Chicken-Genome-Sequencing-Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**(7018):695–716.

[International-Human-Genome-Sequencing-Consortium, 2004] International-Human-Genome-Sequencing-Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011):931–45.

[Janeway et al., 2004] Janeway, C., Travers, P., Walport, M., and Schlomchik, M., 2004. Immunobiology: Garland Science. *New York*, .

[Jen et al., 2009] Jen, D., Engels, R., and Stolte, C., 2009. Medea: Comparative genomic visualization with adobe flash, `http://www.broadinstitute.org/annotation/medea`.

[Jiang et al., 2011] Jiang, N., Weinstein, J. A., Penland, L., White, R. A., Fisher, D. S.,

and Quake, S. R., 2011. 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.

[Jones et al., 01] Jones, E., Oliphant, T., Peterson, P., et al., 2001–. SciPy: Open source scientific tools for Python.

[Kaas et al., 2012] Kaas, R. S., Friis, C., Ussery, D. W., and Aarestrup, F. M., 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics*, **13**:577.

[Karolchik et al., 2014] Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., *et al.*, 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, **42**(1):D764–D770.

[Karolchik et al., 2004] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J., 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, **32**(Database issue):D493–D496.

[Karp, 1972] Karp, R., 1972. Reducibility among combinatorial problems. *Plenum*, (Complexity of Computer Computations):85–103.

[Kendall, 1938] Kendall, M., 1938. A new measure of rank correlation. *Biometrika*, **30**(1/2):81–93.

[Kent, 2002] Kent, W. J., 2002. Blat–the blast-like alignment tool. *Genome Res*, **12**(4):656–64.

[Kent et al., 2003] Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA*, **100**(20):11484–9.

[Kerkhoven et al., 2004] Kerkhoven, R., van Enckevort, F. H. J., Boekhorst, J., Molenaar, D., and Siezen, R. J., 2004. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics*, **20**(11):1812–1814.

[Khan, 2000] Khan, M., 2000. HLA-B27 polymorphism and association with disease. *The Journal of rheumatology*, **27**(5):1110.

[Kim et al., 2014] Kim, H. J., Nguyen, N., Haeussler, M., Haussler, D., and Pourmand, N., 2014. The UCSC Immunobrowser: Interactive Analysis of T-cell Receptor Sequencing Experiments, `http://immuno.soe.ucsc.edu`.

[Kindt et al., 2007] Kindt, T., Goldsby, R., Osborne, B., and Kuby, J., 2007. *Kuby immunology*. WH Freeman.

[Kirkpatrick, 2010] Kirkpatrick, M., 2010. How and why chromosome inversions evolve. *PLoS Biol*, **8**(9).

[Klarenbeek et al., 2012] Klarenbeek, P. L., de Hair, M. J. H., Doorenspleet, M. E., van Schaik, B. D. C., Esveldt, R. E. E., van de Sande, M. G. H., Cantaert, T., Gerlag, D. M., Baeten, D., van Kampen, A. H. C., *et al.*, 2012. Inflamed target tissue provides

a specific niche for highly expanded T-cell clones in early human autoimmune disease. *Annals of the Rheumatic Diseases*, .

[Klarenbeek et al., 2010] Klarenbeek, P. L., Tak, P. P., van Schaik, B. D. C., Zwinderman, A. H., Jakobs, M. E., Zhang, Z., van Kampen, A. H. C., van Lier, R. A. W., Baas, F., and de Vries, N., *et al.*, 2010. Human T-cell memory consists mainly of unexpanded clones. *Immunology Letters*, **133**(1):42–48.

[Koning et al., 2013] Koning, D., Costa, A. I., Hoof, I., Miles, J. J., Nanlohy, N. M., Ladell, K., Matthews, K. K., Venturi, V., Schellens, I. M. M., Borghans, J. A. M., *et al.*, 2013. CD8+ TCR Repertoire Formation Is Guided Primarily by the Peptide Component of the Antigenic Complex. *Journal of Immunology*, **190**(3):931–939.

[Krogsgaard and Davis, 2005] Krogsgaard, M. and Davis, M., 2005. How T cells' see'antigen. *Nature Immunology*, **6**(3):239–245.

[Krzywinski et al., 2009] Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A., 2009. Circos: an information aesthetic for comparative genomics. *Genome research*, **19**(9):1639–1645.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

[Lefranc et al., 2009] Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., *et al.*,

2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Research*, **37**(Database):D1006–D1012.

[Lefranc, 2003] Lefranc, M.-P. M., 2003. IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis, http://imgt.cines.fr. *Leukemia*, **17**(1):260–266.

[Leimbach et al., 2013] Leimbach, A., Hacker, J., and Dobrindt, U., 2013. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. *Current topics in microbiology and immunology*, **358**:3–32.

[Levy et al., 2007] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., *et al.*, 2007. The diploid genome sequence of an individual human. *PLoS Biol*, **5**(10):e254.

[Li and Durbin, 2009] Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14):1754–60.

[Li and Durbin, 2011] Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357):493–6.

[Li et al., 2010] Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., *et al.*, 2010. Building the sequence map of the human pan-genome. *Nature biotechnology*, **28**(1):57–63.

[Li et al., 2013] Li, S., Lefranc, M.-P., Miles, J. J., Alamyar, E., Giudicelli, V., Duroux, P., Freeman, J. D., Corbin, V. D. A., Scheerlinck, J.-P., Frohman, M. A., *et al.*, 2013.

IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications*, **4**:–.

[Li et al., 1994] Li, Y. Y., Sun, G. R. G., Tumang, J. R. J., Crow, M. K. M., and Friedman, S. M. S., 1994. CDR3 sequence motifs shared by oligoclonal rheumatoid arthritis synovial T cells. Evidence for an antigen-driven response. *Journal of Clinical Investigation*, **94**(6):2525–2531.

[Liang et al., 2010] Liang, X., Weigand, L. U., Schuster, I. G., Eppinger, E., van der Griendt, J. C., Schub, A., Leisegang, M., Sommermeyer, D., Anderl, F., Han, Y., *et al.*, 2010. A single TCR alpha-chain with dominant peptide recognition in the allorestricted HER2/neu-specific T cell repertoire. *Journal of Immunology*, **184**(3):1617–1629.

[Lindblad-Toh et al., 2005] Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., *et al.*, 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**(7069):803–19.

[Lukjancenko et al., 2010] Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W., 2010. Comparison of 61 sequenced Escherichia coli genomes. *Microbial ecology*, **60**(4):708–720.

[Maecker et al., 2012] Maecker, H. T. H., Lindstrom, T. M. T., Robinson, W. H. W., Utz, P. J. P., Hale, M. M., Boyd, S. S., Scott, S. D. S., Shen-Orr, S. S. S., and

Fathman, C. G. C., 2012. New tools for classification and monitoring of autoimmune diseases. *Nature Reviews: Rheumatology*, **8**(10):562–562.

[Mamedov et al., 2009] Mamedov, I. Z., Britanova, O. V., Chkalina, A. V., Staroverov, D. B., Amosova, A. L., Mishin, A. S., Kurnikova, M. A., Zvyagin, I. V., Mutovina, Z. Y., Gordeev, A. V., *et al.*, 2009. Individual characterization of stably expanded T cell clones in ankylosing spondylitis patients. *Autoimmunity*, **42**(6):525–536.

[Märker-Hermann and Höhler, 1998] Märker-Hermann, E. and Höhler, T., 1998. Pathogenesis of human leukocyte antigen B27-positive arthritis: Information from Clinical Materials. *Rheumatic Disease Clinics of North America*, **24**(4):865–881.

[Marth et al., 2004] Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**(1):351–72.

[Mayor et al., 2000] Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I., 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**(11):1046–1047.

[McGinnis and Madden, 2004] McGinnis, S. and Madden, T. L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, .

[Medini et al., 2005] Medini, D., Donati, C., Tettelin, H., and Masignani, V., 2005. The microbial pan-genome. *Current opinion in Genetics and Development*, .

[Medvedev and Brudno, 2009] Medvedev, P. and Brudno, M., 2009. Maximum likelihood genome assembly. *J Comput Biol*, **16**(8):1101–16.

[Mellmann et al., 2011] Mellmann, A. A., Harmsen, D. D., Cummings, C. A. C., Zentz, E. B. E., Leopold, S. R. S., Rico, A. A., Prior, K. K., Szczepanowski, R. R., Ji, Y. Y., Zhang, W. W., *et al.*, 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE*, **6**(7):e22751–e22751.

[Miles et al., 2011] Miles, J. J., Douek, D. C., and Price, D. A., 2011. Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunology and Cell Biology*, **89**(3):375–387.

[Miller et al., 2007] Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., *et al.*, 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, **17**(12):1797–808.

[Mouse-Genome-Sequencing-Consortium et al., 2002] Mouse-Genome-Sequencing-Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., *et al.*,

2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–62.

[Muraro et al., 2014] Muraro, P. A., Robins, H., Malhotra, S., Howell, M., Phippard, D., Desmarais, C., de Paula Alves Sousa, A., Griffith, L. M., Lim, N., Nash, R. A., *et al.*, 2014. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *Journal of Clinical Investigation*, **124**(3):1168–1172.

[Murugan et al., 2012] Murugan, A. A., Mora, T. T., Walczak, A. M. A., and Callan, C. G. C., 2012. 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.

[Naggiar, 2014] Naggiar, S., 2014. Teen Makes Genetic Discovery of Her Own Rare Cancer. *NBCNews.com*, `http://www.nbcnews.com/health/cancer/teen-makes-genetic-discovery-her-own-rare-cancer-n75991`.

[Newman, 2008] Newman, A., 2008. Max-cut. *Encyclopedia of Algorithms*, **1**:489–492.

[Nguyen et al., 2014a] Nguyen, N., Hickey, G., Raney, B. J., Armstrong, J., Clawson, H., Zweig, A., Karolchik, D., Kent, J., Haussler, D., and Paten, B., *et al.*, 2014a. Comparative Assembly Hubs: Web Accessible Browsers for Comparative Genomics. *Preprint*, **NA**.

[Nguyen et al., 2014b] Nguyen, N., Hickey, G., Zerbino, D. R., Raney, B. J., Earl, D., Armstrong, J., Haussler, D., and Paten, B., 2014b. *Building a Pangenome Reference*

for a Population. In Sharan, R., editor, *Proceedings of RECOMB 2014*, pages 207–221.

[Nguyen et al., 2014c] Nguyen, N., Zerbino, D., Earl, D., Raney, B., Hickey, G., Diekhans, M., Haussler, D., and Paten, B., 2014c. Towards the Universal Human Reference Genome: Building a Comprehensive Consensus Sequence for the Major Histocompatibility Complex. *Preprint*, **NA**.

[Nielsen et al., 2010] Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., and Wang, T., 2010. Visualizing genomes: techniques and challenges. *Nature methods*, **7**(3 Suppl):S5–S15.

[Ogura et al., 2009] Ogura, Y. Y., Ooka, T. T., Iguchi, A. A., Toh, H. H., Asadulghani, M. M., Oshima, K. K., Kodama, T. T., Abe, H. H., Nakayama, K. K., Kurokawa, K. K., *et al.*, 2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. *Audio, Transactions of the IRE Professional Group on*, **106**(42):17939–17944.

[Ohta et al., 1991] Ohta, K. K., Beall, D. S. D., Mejia, J. P. J., Shanmugam, K. T. K., and Ingram, L. O. L., 1991. Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. *Applied and Environmental Microbiology*, **57**(4):893–900.

[Ohtsubo, 2008] Ohtsubo, 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics*, **9**(1):376–376.

[Oksanen et al., 2012] Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H., *et al.*, 2012. *vegan: Community Ecology Package*. R package version 2.0-4.

[Osoegawa et al., 2001] Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., and de Jong, P. J., 2001. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res*, **11**(3):483–96.

[Overmars et al., 2013] Overmars, L., Kerkhoven, R., Siezen, R. J., and Francke, C., 2013. MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC genomics*, **14**:209.

[Park, 2009] Park, P. J., 2009. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**(10):669–80.

[Paten et al., 2011a] Paten, B., Diekhans, M., Earl, D., John, J. S., Ma, J., Suh, B., and Haussler, D., 2011a. Cactus graphs for genome comparisons. *J Comput Biol*, **18**(3):469–81.

[Paten et al., 2011b] Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D., 2011b. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res*, **21**(9):1512–28.

[Paten et al., 2008] Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E., 2008.

Enredo and pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*, **18**(11):1814–1828.

[Paten et al., 2014] Paten, B., Novak, A., and Haussler, D., 2014. Mapping to a Reference Genome Structure. *Preprint*, **NA**.

[Patterson et al., 2006] Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, **441**(7097):1103–8.

[Perna, 2011] Perna, N. T., 2011. Genomics of escherichia and shigella. In *Genomics of Foodborne Bacterial Pathogens.*, pages 119–139. Springer.

[Pruitt et al., 2012] Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R., 2012. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res*, **40**(1):D130–5.

[Raney et al., 2013] Raney, B., Dreszer, T., Barber, G., Clawson, H., Fujita, P., Wang, T., Nguyen, N., Paten, B., Zweig, A., Karolchik, D., *et al.*, 2013. Track Data Hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, :btt637.

[Rasko et al., 2011] Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E. E., Sebra, R., Chin, C.-S., Iliopoulos, D., *et al.*, 2011. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England journal of medicine*, **365**(8):709–717.

[Raymond et al., 2005] Raymond, C. K., Kas, A., Paddock, M., Qiu, R., Zhou, Y., Subramanian, S., Chang, J., Palmieri, A., Haugen, E., Kaul, R., *et al.*, 2005. Ancient haplotypes of the hla class ii region. *Genome Res*, **15**(9):1250–7.

[Ristori et al., 2000] Ristori, G., Giubilei, F., Giunti, D., Perna, A., Gasperini, C., Buttinelli, C., Salvetti, M., and Uccelli, A., 2000. Myelin basic protein intramolecular spreading without disease progression in a patient with multiple sclerosis. *Journal of Neuroimmunology*, **110**(1-2):240–243.

[Robins, 2013] Robins, H., 2013. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, **25**(5):646–652.

[Robins et al., 2012] Robins, H. H., Desmarais, C. C., Matthis, J. J., Livingston, R. R., Andriesen, J. J., Reijonen, H. H., Carlson, C. C., Nepom, G. G., Yee, C. C., and Cerosaletti, K. K., *et al.*, 2012. Ultra-sensitive detection of rare T cell clones. *Journal of Immunological Methods*, **375**(1-2):6–6.

[Robins et al., 2009] Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., and Carlson, C. S., 2009. Comprehensive assessment of T-cell receptor -chain diversity in T cells. *Blood*, **114**(19):4099–4107.

[Robins et al., 2010] Robins, H. S., Srivastava, S. K., Campregher, P. V., Turtle, C. J., Andriesen, J., Riddell, S. R., Carlson, C. S., and Warren, E. H., 2010. Overlap and

Effective Size of the Human CD8+ T Cell Receptor Repertoire. *Science Translational Medicine*, **2**(47):47ra64–47ra64.

[Rocha, 2008] Rocha, E. P. C., 2008. The Organization of the Bacterial Genome. *Annual Review of Genetics*, **42**(1):211–233.

[Rohde et al., 2011] Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J., *et al.*, 2011. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *The New England journal of medicine*, **365**(8):718–724.

[Rosenbloom et al., 2010] Rosenbloom, K. R., Dreszer, T. R., Pheasant, M., Barber, G. P., Meyer, L. R., Pohl, A., Raney, B. J., Wang, T., Hinrichs, A. S., Zweig, A. S., *et al.*, 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic acids research*, **38**(Database issue):D620–D625.

[Rudd et al., 2011] Rudd, B. D., Venturi, V., Davenport, M. P., and Nikolich-Zugich, J., 2011. Evolution of the Antigen-Specific CD8+ TCR Repertoire across the Life Span: Evidence for Clonal Homogenization of the Old TCR Repertoire. *The Journal of Immunology*, **186**(4):2056–2064.

[S et al., 2004] S, K., A, P., AL, D., M, S., M, S., C, A., and SL, S., 2004. Versatile and open software for comparing large genomes. *Genome Biology*, **5**(2):R12–R12.

[Sahl et al., 2011] Sahl, J. W. J., Steinsland, H. H., Redman, J. C. J., Angiuoli, S. V. S., Nataro, J. P. J., Sommerfelt, H. H., and Rasko, D. A. D., 2011. A comparative

genomic analysis of diverse clonal types of enterotoxigenic Escherichia coli reveals pathovar-specific conservation. *Audio, Transactions of the IRE Professional Group on*, **79**(2):950–960.

[Salzberg et al., 2011] Salzberg, S. L., Phillippy, A. M., Zimin, A. V., Puiu, D., Magoc, T., Koren, S., Treangen, T., Schatz, M. C., Delcher, A. L., Roberts, M., *et al.*, 2011. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, .

[Schirmer et al., 2002] Schirmer, M. M., Goldberger, C. C., Würzner, R. R., Duftner, C. C., Pfeiffer, K.-P. K., Clausen, J. J., Neumayr, G. G., and Falkenbach, A. A., 2002. Circulating cytotoxic CD8(+) CD28(-) T cells in ankylosing spondylitis. *Arthritis Research*, **4**(1):71–76.

[Schlosstein et al., 1973] Schlosstein, L., Terasaki, P., Bluestone, R., and Pearson, C., 1973. High association of an HL-A antigen, W27, with ankylosing spondylitis. *The New England journal of medicine*, **288**(14):704.

[Sepúlveda et al., 2010] Sepúlveda, N., Paulino, C. D., and Carneiro, J., 2010. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *Journal of Immunological Methods*, **353**(1-2):14–14.

[Sherry et al., 2001] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K., 2001. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, **29**(1):308–11.

[Siepel et al., 2006] Siepel, A., Pollard, K., and Haussler, D., 2006. New methods for detecting lineage-specific selection. *In Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, :190–205.

[Smit et al., 2010] Smit, A. F. A., Hubley, R., and Green, P., 2010. Repeatmasker open-3.0. .

[Stamatakis, 2006] Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21):2688–2690.

[Stewart et al., 2004] Stewart, C. A., Horton, R., Allcock, R. J. N., Ashurst, J. L., Atrazhev, A. M., Coggill, P., Dunham, I., Forbes, S., Halls, K., Howson, J. M. M., *et al.*, 2004. Complete mhc haplotype sequencing for common disease gene mapping. *Genome Res*, **14**(6):1176–87.

[Striebich et al., 1998] Striebich, C. C., Falta, M. T., Wang, Y., Bill, J., and Kotzin, B. L., 1998. Selective accumulation of related CD4+ T cell clones in the synovial fluid of patients with rheumatoid arthritis. *Journal of Immunology*, **161**(8):4428–4436.

[Tam et al., 2010] Tam, L.-S., Gu, J., and Yu, D., 2010. Pathogenesis of ankylosing spondylitis. *Nature Publishing Group*, **6**(7):399–405.

[Tannier et al., 2009] Tannier, E., Zheng, C., and Sankoff, D., 2009. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, **10**:120.

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

[Thorvaldsdóttir et al., 2013] Thorvaldsdóttir, H. H., Robinson, J. T. J., and Mesirov, J. P. J., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**(2):178–192.

[Touchon et al., 2009] Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., *et al.*, 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genetics*, **5**(1):e1000344.

[Traherne, 2008] Traherne, J. A., 2008. Human mhc architecture and evolution: implications for disease association studies. *Int J Immunogenet*, **35**(3):179–92.

[Traherne et al., 2006] Traherne, J. A., Horton, R., Roberts, A. N., Miretti, M. M., Hurles, M. E., Stewart, C. A., Ashurst, J. L., Atrazhev, A. M., Coggill, P., Palmer, S., *et al.*, 2006. Genetic analysis of completely sequenced disease-associated mhc haplotypes identifies shuffling of segments in recent human history. *PLoS Genet*, **2**(1):e9.

[Trapnell et al., 2009] Trapnell, C., Pachter, L., and Salzberg, S. L., 2009. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9):1105–11.

[Turner et al., 2012] Turner, P. C. P., Yomano, L. P. L., Jarboe, L. R. L., York, S. W. S., Baggett, C. L. C., Moritz, B. E. B., Zentz, E. B. E., Shanmugam, K. T. K., and Ingram, L. O. L., 2012. Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis pdc* and *adhB* genes. *Journal of Industrial Microbiology & Biotechnology*, **39**(4):629–639.

[van der Linden et al., 1984] van der Linden, S. S., Valkenburg, H. A. H., and Cats, A. A., 1984. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis & Rheumatism*, **27**(4):361–368.

[Vieira et al., 2011] Vieira, G., Sabarly, V., Bourguignon, P.-Y., Durot, M., Le Fevre, F., Mornico, D., Vallenet, D., Bouvet, O., Denamur, E., Schachter, V., *et al.*, 2011. Core and panmetabolism in Escherichia coli. *Journal of Bacteriology*, **193**(6):1461–1472.

[Wang et al., 2010] Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., Davis, R. W., *et al.*, 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences*, **107**(4):1518–1523.

[Wang et al., 2008] Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., *et al.*, 2008. The diploid genome sequence of an asian individual. *Nature*, **456**(7218):60–5.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M., 2009. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1):57–63.

[Warren et al., 2011] Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R., and Holt, R. A., 2011. 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.

[Waterhouse et al., 2009] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J., 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9):1189–1191.

[Wellcome-Trust-Case-Control-Consortium, 2007] Wellcome-Trust-Case-Control-Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145):661–78.

[Wheeler et al., 2008] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., *et al.*, 2008. The complete genome of an individual by massively parallel dna sequencing. *Nature*, **452**(7189):872–6.

[Wu et al., 2012] Wu, D. D., Sherwood, A. A., Fromm, J. R. J., Winter, S. S. S., Dunsmore, K. P. K., Loh, M. L. M., Greisman, H. A. H., Sabath, D. E. D., Wood, B. L. B., and Robins, H. H., *et al.*, 2012. High-throughput sequencing detects minimal

residual disease in acute T lymphoblastic leukemia. *Science Translational Medicine*, **4**(134):134ra63–134ra63.

[Xu, 2009] Xu, A. W., 2009. A fast and exact algorithm for the median of three problem: a graph decomposition approach. *J Comput Biol*, **16**(10):1369–81.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, **18**(5):821–9.

[Zhang et al., 2006] Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., and Scherer, S. W., 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res*, **115**(3-4):205–14.

[Zhu et al., 2007] Zhu, J., Sanborn, J. Z., Diekhans, M., Lowe, C. B., Pringle, T. H., and Haussler, D., 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Computational Biology*, **3**(12):e247–e247.

Appendix A

Supplement for: "Building a Pan-genome Reference for a Population"

A.1 NP-hardness of the Pan-genome Reference Problem

The pan-genome reference problem is NP-hard and can be projected onto the problem of finding maximum weight subgraphs of a bidirected graph that do not contain characteristic classes of simple cycle. See Appendix Section A.1 for a full proof of the problem's NP-hardness.

A M, N *bidirected simple cycle*, henceforth abbreviated to a M, N-cycle, is a simple cycle in a bidirected graph containing M vertices such that $M \geq N$, $M - N$ of the vertices have both their sides incident with an edge in the cycle (called *balanced vertices*) and the other N vertices have only one side incident with edges in the cycle (called *unbalanced vertices*). A M, N-cycle is odd if N is odd, else it is called even. A

bidirected graph is *strongly acyclic* if it contains no $M, 0$-cycles or odd M, N-cycles. Let $\hat{\mathbf{G}}$ be the set of all strongly acyclic subgraphs of $\hat{G}$ of maximum weight. The following lemma shows the relationship between maximum weight strongly acyclic subgraphs and maximum weight pan-genome references.

Lemma 1 *There exists a surjection $f : \mathbf{F} \twoheadrightarrow \hat{\mathbf{G}}$, such that for all F in $\mathbf{F}$, $f(F) = (V, \hat{E}_F)$.*

Proof 1 *Let $F \in \mathbf{F}$, the threads in F orient all the vertices, partitioning the sides into two sets according to if they appear in a pan-genome reference thread or not. By definition, the consistent edges and this bipartition of the sides form a bipartite graph. If there exists an odd M, N-cycle in $f(R)$, then it defines an odd cycle in this bipartite graph (a contradiction), hence $f(R)$ contains no odd M, N-cycles.*

A pan-genome reference induces a partial $<_F$ order on the vertices. If there exists a $M, 0$-cycle $\{\{X_1, -X_2\}, \{X_2, -X_3\}, \ldots, \{X_n, -X_1\}\} \in f(R)$, as these edges are consistent with F, this implies that both $\{X_1, -X_1\} <_F \{X_n, -X_n\}$ and $\{X_n, -X_n\} <_F \{X_1, -X_1\}$, but a partial order is asymmetric (a contradiction), therefore $f(R)$ contains no $M, 0 - cycles$.

As $f(F)$ is strongly acyclic, if it is not in $\hat{\mathbf{G}}$ then it must be possible to add an edge to $f(F)$ without creating a $M, 0$-cycle or odd M, N-cycle. Assume therefore that $f(F)$ is a subgraph of some $\hat{G}' \in \hat{\mathbf{G}}$. Let $\{X, Y\}$ be an edge in $\hat{G}'$ but not in $f(F)$. By definition, $\{X, Y\}$ has non-zero weight. Between $\{X, -X\}$ and $\{Y, -Y\}$ of the three other possible edges, $\{\{X, -Y\}, \{-X, Y\}, \{-X, -Y\}\}$, one must be in $\hat{E}_F$, else F is not

a maximum weight solution to the pan-genome reference problem, because in this case there must exist two threads in F, one that contains X or $-X$ and one that contains Y or $-Y$, and these two threads can be concatenated together to create a new pan-genome reference additionally consistent with one of the four possible edges between $\{X, -X\}$ and $\{Y, -Y\}$. If $\{X, -Y\} \in \hat{E}_F$ then $\hat{G}'$ contains a $2,1$-cycle $\{\{X, -Y\}, \{Y, X\}\}$, if $\{-X, -Y\}$ then $\hat{G}'$ contains a $2,0$-cycle $\{\{-X, -Y\}, \{Y, X\}\}$ and if $\{-X, Y\}$ then $\hat{G}'$ contains a $2,1$-cycle $\{\{-X, Y\}, \{Y, X\}\}$. A contradiction is derived in all cases, therefore $f(R) \in \hat{\mathbf{G}}$.

It remains to prove that for every member of $\hat{G}'$ in $\hat{\mathbf{G}}$ there exists F such that $f(F) = \hat{G}'$. For $\hat{G}' = (\hat{V}', \hat{E}') \in \hat{\mathbf{G}}$ a side bicolouring *is a labelling function* colour*, such that each vertex and edge's sides are coloured such that one is* black *and the other is* red*, i.e. it creates a bipartition of the sides of the graph.*

To construct such a colouring for $\hat{G}'$ use a depth first search. In each connected component of $\hat{G}'$ pick an unlabeled vertex and colour one of it sides red and the other black. The depth first search then recurses from this vertex such that for each edge of the form $\{X, Y\}$ if X is coloured red and Y is unlabeled then Y is coloured black and $-Y$ is coloured red and vice versa if X is coloured black. If during this recursion an edge is encountered such that both sides are already labeled then the depth first search has traversed a M, N-cycle. Further, if the sides of this edge are labeled with the same colour then the depth first search has failed to produce a side bicolouring. Suppose such a cycle is encountered in $\hat{G}'$, either there are two excess black sides or two excess red sides, as only the last edge encountered does not have sides of distinct colours.

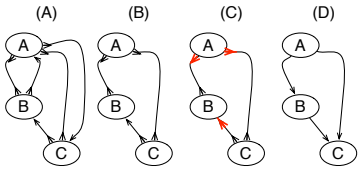

Figure A.1: (A) A bidirected graph with three vertices A, B and C. (B) A subgraph of (A) containing no $M, 0$-cycles or odd M, N-cycles. (C) A side bicolouring of (B). (D) A digraph for (C).

Each balanced vertex contributes a black and a red side while each unbalanced vertex contributes either two black sides or two red sides, therefore $N \geq 1$. *Furthermore, as there are only two excess vertices of one colour* N *must be odd, implying* $\hat{G}'$ *is not strongly acyclic, therefore there exists a side bicolouring of* $\hat{G}'$. *Given a side bicolouring of* $\hat{G}'$ *let* $\hat{G}'' = (\hat{V}'', \hat{E}'')$ *be a digraph, such that* $\hat{V}'' = \{X : \{X, -X\} \in \hat{V}' \wedge \text{colour}(X) = \textit{red}\}$ *and* $\hat{E}'' = \{(X, Y) : \{X, -Y\} \in \hat{E}' \wedge \text{colour}(X) = \textit{red} \wedge \text{colour}(-Y) = \textit{black}\}$, *where* (a, b) *is a directed edge from* a *to* b. *The graph* $\hat{G}''$ *is isomorphic to* $\hat{G}'$, *except that the arbitrary orientations of the sides within the vertices have been reassigned so that there is only one type of edge in the graph (Fig. A.1). A directed cycle in* $\hat{G}''$ *would be a* $M, 0$*-cycle, but as* $\hat{G}''$ *is strongly acyclic it must contain no directed cycles, therefore* $\hat{G}''$ *is a DAG. Any topological sort* $F = \{X_1, X_2, \ldots, X_n\}$ *of the vertices of* $\hat{G}''$ *is a pan-genome reference for which* $f(F) = \hat{G}'$. □

Theorem 1 *The pan-genome reference problem is NP-hard.*

Proof 2 *The problem of finding a maximum weight strongly acyclic subgraph of a bidirected graph is polynomial-time reducible to the pan-genome reference problem, because, by the previous lemma, the consistent subgraph of any solution to the pan-genome ref-*

erence problem is a maximum weight strongly acyclic subgraph. It remains to prove that the problem of finding a maximum weight strongly acyclic subgraph of a bidirected graph is NP-hard. This is proved here by reduction of the minimum feedback arc set problem [Karp, 1972], which is to find the smallest set of edges in a directed graph that when removed result in a graph containing no directed cycles. Using the demonstration in the previous lemma, a digraph can be equivalently represented as a side bicoloured bidirected graph. An unbalanced vertex in an M, N-cycle is red if the endpoints of the edges incident with it in the cycle are colored red, else it is black. Suppose there exists an M, N-cycle in a side bicoloured bidirected graph with i balanced vertices, j unbalanced red vertices and k unbalanced black vertices. As in a side bicoloured bidirected graph each edge has one red endpoint and one black endpoint the total number of red and black endpoints is equal, therefore $i+2j = i+2k$, thus $k = j$ and therefore it is not possible to construct an odd M, N-cycle in a side bicoloured bidirected graph. As a directed cycle in a digraph corresponds to an $M, 0$-cycle in the equivalent side bicoloured bidirected graph, the minimum feedback arc set problem is thus polynomial-time reducible to the problem of finding a maximum weight strongly acyclic subgraph of a side bicolored bidirected graph (i.e. eliminating $M, 0$-cycles). □

An alternative, similarly simple proof of NP-hardness uses the elimination of odd M, N-cycles rather than the $M, 0$-cycles, reducing the maximum bipartite subgraph problem[Newman, 2008].

A.2 C. Ref. Sample Composition

Sample	Repeat	GRCh37	C. Ref.	Total
venter	2,520,085 (52.40 %)	4,719,450 (98.13 %)	4,782,784 (99.45 %)	4,809,387
cox	2,520,556 (52.56 %)	4,687,818 (97.76 %)	4,790,727 (99.90 %)	4,795,371
qbl	2,252,520 (52.44 %)	4,200,985 (97.80 %)	4,294,505 (99.98 %)	4,295,325
dbb	2,192,556 (52.15 %)	4,089,233 (97.26 %)	4,202,262 (99.95 %)	4,204,302
NA12878	1,952,189 (46.56 %)	3,997,156 (95.34 %)	4,127,361 (98.44 %)	4,192,579
ssto	2,229,126 (53.41 %)	3,953,873 (94.74 %)	4,152,690 (99.50 %)	4,173,551
mann	2,215,801 (54.03 %)	3,877,792 (94.56 %)	4,090,561 (99.75 %)	4,100,741
MCF	1,939,138 (51.10 %)	3,706,922 (97.68 %)	3,794,583 (99.99 %)	3,794,911
YH1	1,190,870 (36.07 %)	3,260,622 (98.77 %)	3,282,982 (99.45 %)	3,301,296
NA19240	1,068,403 (32.86 %)	3,209,465 (98.72 %)	3,213,608 (98.85 %)	3,251,154
NA19239	905,038 (29.48 %)	3,024,497 (98.52 %)	3,032,949 (98.80 %)	3,069,926
NA12892	955,690 (31.22 %)	3,033,815 (99.11 %)	3,039,870 (99.31 %)	3,061,138
NA18507	1,019,209 (33.49 %)	2,991,861 (98.30 %)	3,016,172 (99.09 %)	3,043,745
apd	1,222,215 (52.66 %)	2,293,261 (98.82 %)	2,320,668 (100.00 %)	2,320,747
NA19238	461,000 (20.42 %)	2,233,579 (98.92 %)	2,234,998 (98.98 %)	2,258,041
average	1,704,147 (43.98 %)	3,639,782 (97.64 %)	3,708,307 (99.48 %)	3,727,669
all	NA	4,970,600 (88.84 %)	5,285,388 (94.46 %)	5,595,284
panTro3	2,499,545 (52.01 %)	4,517,719 (94.01 %)	4,620,359 (96.14 %)	4,805,689

Table A.1: The number of bases from each sample classified as repetitive by repeat masker (Repeat column), aligned to GRCh37 (GRCh37 column), aligned to C. Ref. (C. Ref. column) and covered by the Cactus MSA (Total column). The 'average' category gives the average over all human samples. The 'all' category considers all columns and unaligned bases in all the human samples, e.g. as 1 base per homology set.

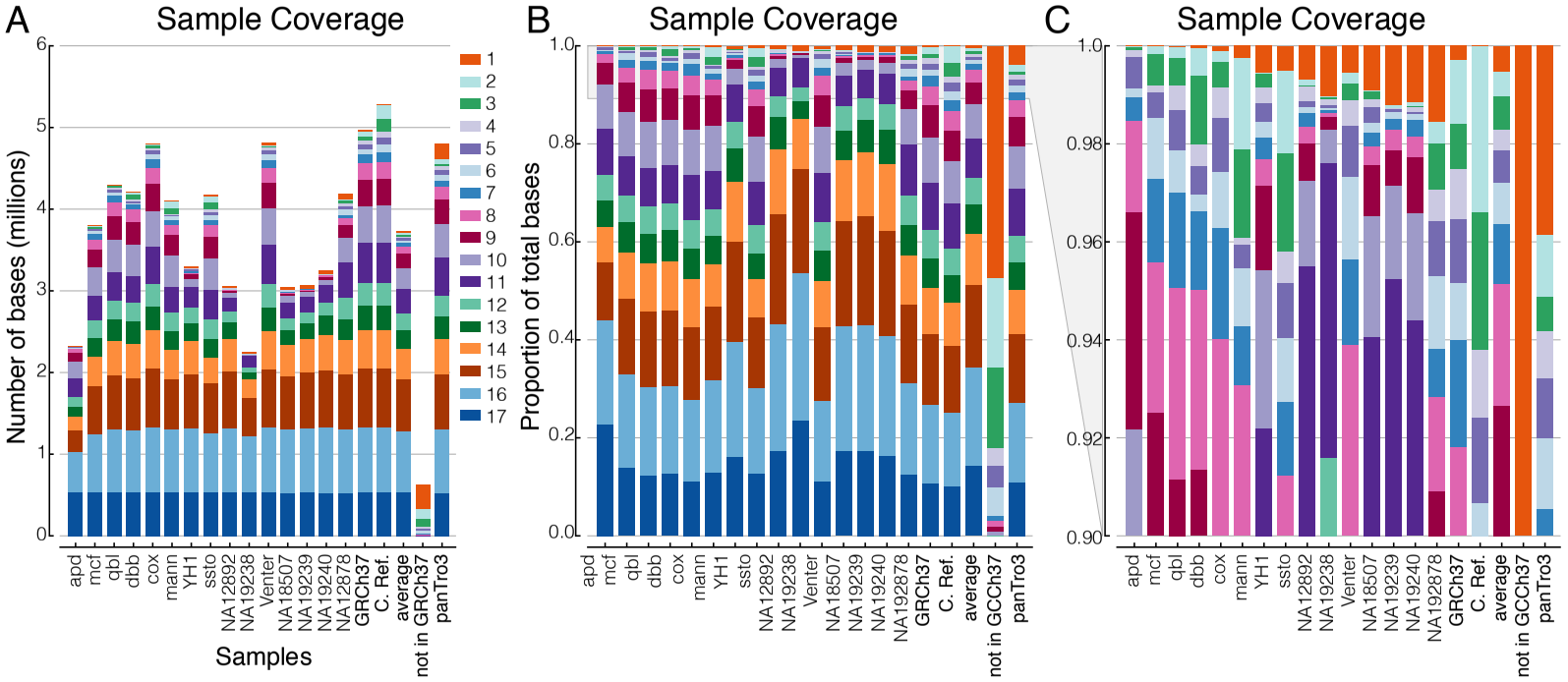

Figure A.2: The distribution of occurrence within the samples of homologous bases. (A) A stacked histogram showing the number of bases in each sample mapped to the Cactus MSA, where each colour is associated with an integer that gives the number of bases in the sample contained in columns of that cardinality. (B) As in (A), but showing the proportion of bases. (C) As in (B), but zoomed in to show the 10% of least shared bases. Samples are labelled by name, the 'not in GRCh37' category gives all bases in the samples not mapped to bases in GRCh37.

A.2.1 Manual Analysis of False Positive SNVs

dbSNP Validation of Single Nucleotide Variations

Sample	All				Recurrent			
	T#	TP	STP	SFN	T#	TP	STP	SFN
NA12878	12351	90	57	53	10700	97	63	55
NA12892	8949	77	63	57	6660	95	79	60
NA19238	5370	67	58	77	3475	95	83	78
NA19239	9116	72	59	64	6574	94	77	66
NA19240	10315	73	59	58	7346	94	77	61
apd	4726	99	NA	NA	4555	99	NA	NA
cox	15028	98	NA	NA	14148	98	NA	NA
dbb	13329	98	NA	NA	12840	99	NA	NA
mann	14144	97	NA	NA	13099	98	NA	NA
mcf	12012	98	NA	NA	11359	98	NA	NA
nigerian	7199	90	72	67	6089	98	78	70
qbl	14336	97	NA	NA	13127	97	NA	NA
ssto	14173	98	NA	NA	12938	98	NA	NA
venter	14322	95	67	34	12885	98	70	38
yanhuang	8394	78	63	67	6291	98	79	69
panTro3	65877	25	NA	NA	12965	96	NA	NA
aggregate	56080	76	NA	NA	34402	95	NA	NA
C. Ref.	10461	97	NA	NA	10461	97	NA	NA

Table A.2: All: all SNVs detected in each sample with respect to GRCh37. Recurrent: as All, but excluding SNVs not present in at least two samples, including chimp. T#: Total number of SNVs. TP: Percentage true positives, as validated by a matching SNV in dbSNP. STP: Percentage (sample) true positives, as validated by those reported for the sample in question. SFN: Percentage (sample) false negatives, as validated by those reported for the sample in question. An NA entry denotes that the data was not available. Aggregate row: gives the total SNVs in human samples (excluding chimp). C. Ref. row: gives SNVs in C. Ref. with respect to GRCh37

dbSNP Validation of Filtered Single Nucleotide Variations

Sample	All				Recurrent			
	T#	TP	STP	SFN	T#	TP	STP	SFN
NA12878	10252	92	60	59	8952	99	66	61
NA12892	6668	83	71	64	5303	96	83	67
NA19238	3730	78	70	80	2744	96	88	82
NA19239	6494	81	68	71	5165	95	81	72
NA19240	7731	79	65	65	5851	95	80	68
apd	3987	100	NA	NA	3832	100	NA	NA
cox	12639	99	NA	NA	11901	99	NA	NA
dbb	11235	99	NA	NA	10825	100	NA	NA
mann	11926	99	NA	NA	11037	99	NA	NA
mcf	10109	99	NA	NA	9552	99	NA	NA
nigerian	5630	94	77	72	4897	99	82	75
qbl	11946	98	NA	NA	10920	98	NA	NA
ssto	11939	99	NA	NA	10900	99	NA	NA
venter	12017	97	69	43	10878	99	72	46
yanhuang	6646	80	66	72	5051	98	81	74
panTro3	57001	25	NA	NA	10908	97	NA	NA
aggregate	43485	82	NA	NA	28344	97	NA	NA
C. Ref.	8843	98	NA	NA	8843	98	NA	NA

Table A.3: Experiment and format same as in Table A.2, but only for 'Filtered SNVs', as defined in Supplementary Figure 2.9

dbSNP Validation of Non-Repetitive Single Nucleotide Variations

Sample	All				Recurrent			
	T#	TP	STP	SFN	T#	TP	STP	SFN
NA12878	5684	97	69	49	5316	98	71	51
NA12892	5180	84	71	72	4126	97	83	74
NA19238	3677	71	64	82	2474	96	88	84
NA19239	5116	84	71	76	4253	95	81	77
NA19240	5256	86	73	73	4377	95	82	75
apd	1798	99	NA	NA	1739	100	NA	NA
cox	6183	98	NA	NA	5945	98	NA	NA
dbb	5239	99	NA	NA	5062	99	NA	NA
mann	5752	98	NA	NA	5417	98	NA	NA
mcf	5094	98	NA	NA	4817	98	NA	NA
nigerian	4429	92	77	78	3795	99	82	80
qbl	6168	98	NA	NA	5765	98	NA	NA
ssto	5938	98	NA	NA	5541	98	NA	NA
venter	5330	97	70	74	4903	99	72	76
yanhuang	4138	92	79	80	3682	98	83	81
panTro3	28036	25	NA	NA	5657	97	NA	NA
aggregate	22505	81	NA	NA	14735	96	NA	NA
C. Ref.	4359	99	NA	NA	4359	99	NA	NA

Table A.4: Experiment and format same as in Table A.2, but only for SNVs at bases not defined as repetitive in GRCh37.

dbSNP Validation of Filtered, Non-Repetitive Single Nucleotide Variations

Sample	All				Recurrent			
	T#	TP	STP	SFN	T#	TP	STP	SFN
NA12878	4658	98	71	57	4351	99	73	59
NA12892	3966	88	76	77	3292	98	85	79
NA19238	2638	80	74	85	1994	96	90	86
NA19239	3899	87	75	80	3347	96	83	81
NA19240	4128	87	75	79	3448	96	84	80
apd	1461	100	NA	NA	1412	100	NA	NA
cox	5022	99	NA	NA	4832	99	NA	NA
dbb	4223	100	NA	NA	4070	100	NA	NA
mann	4694	99	NA	NA	4400	99	NA	NA
mcf	4149	99	NA	NA	3911	99	NA	NA
nigerian	3448	96	82	82	3038	99	85	83
qbl	4943	99	NA	NA	4626	99	NA	NA
ssto	4855	99	NA	NA	4532	99	NA	NA
venter	4298	98	72	79	3965	100	74	80
yanhuang	3236	94	81	84	2920	98	84	85
panTro3	23916	24	NA	NA	4607	98	NA	NA
aggregate	17342	87	NA	NA	11862	97	NA	NA
C. Ref.	3564	99	NA	NA	3564	99	NA	NA

Table A.5: Experiment and format same as in Table A.2, but only for filtered SNVs at bases not defined as repetitive in GRCh37.

A.2.2 Indels

To allow for some alignment uncertainty I permit up to a 5 bp disagreement in the exact location of the indel, in Supplementary Figure *A*.3 I analyze the effect of allowing location disagreement, notably without it the true positive rate falls to 23%. Manual analysis of the MSA and our previous work demonstrating the accuracy of the Cactus MSA program [Paten et al., 2011b] lead us to conclude that this large discrepancy relates to alignment uncertainty [Holmes and Durbin, 1998] rather than a source of systematic error in our alignments.

The overall false negative rate for short indels was 35% in haploid samples and 81% in diploid samples, which is substantially higher than the SNV rate. This is probably partially explainable by the lower true positive rate, i.e. there is likely not an undercalling of indels in the MSA, rather a larger number of indels that are possibly false positives. Looking only at short indels not deemed repetitive, the overall true positive rate increases to 80%, but this accounts for only 31% of all indels. Supplementary Tables *A*.7, *A*.8, *A*.9 and *A*.10 analyse short insertions (sequences present in the sample, but not in the reference) and short deletions (sequences present in the reference, but not in the sample) separately. For these short indels I do not see substantial differences in the level of agreement with the dbSNP/1000 Genomes Project data, e.g. I observe a 64% overall true positive rate for insertions and a 66% overall true positive rate for deletions.

Chrom	Start	C.Ref. Allele	GRCh37 Allele	Annotation
chr6	30463203	T	G	C
chr6	30463204	T	G	C
chr6	32535163	C	G	C
chr6	32535165	G	C	C
chr6	32535237	A	G	C
chr6	32536032	A	T	C
chr6	32536033	C	A	C
chr6	32536034	C	T	C
chr6	32536962	C	T	B
chr6	32547908	G	A	ID
chr6	32548727	C	G	S
chr6	32550935	G	A	B
chr6	32551306	G	A	R
chr6	32551307	A	G	R
chr6	32551852	C	G	S
chr6	32553317	A	C	B
chr6	32570019	T	C	C
chr6	32633098	A	C	B
chr6	32633101	T	C	B
chr6	32633906	T	C	B
chr6	32689731	G	C	C
chr6	32689732	C	T	ID
chr6	32689733	A	T	ID

Table A.6: A manual analysis of C.Ref. non-repetitive and filtered SNVs with respect to GRCh37 that were not in dbSNP or 1000 Genome Project data. These SNVs were neither within the repetitive regions (non-repetitive) nor proximal to a breakpoint (filtered). 'Chrom', 'Start': location of each SNV relative to the positive strand of GRCh37. Annotation: 'C': SNVs were confirmed by an independent MULTIZ multiple sequence alignment (see Supplementary Section 2.5.6). 'B': a bug in dbSNP build 134 that had been fixed in build 135, SNVs were indeed in dbSNP. 'ID': disaggreement between Cactus MSA and other alignments, in which Cactus MSA called substitutions while other alignments called indels. 'R': SNVs were not confirmed by MULTIZ MSA but recurrent within the input samples. 'S': SNVs were not confirmed by MULTIZ MSA and not reccurent (single).

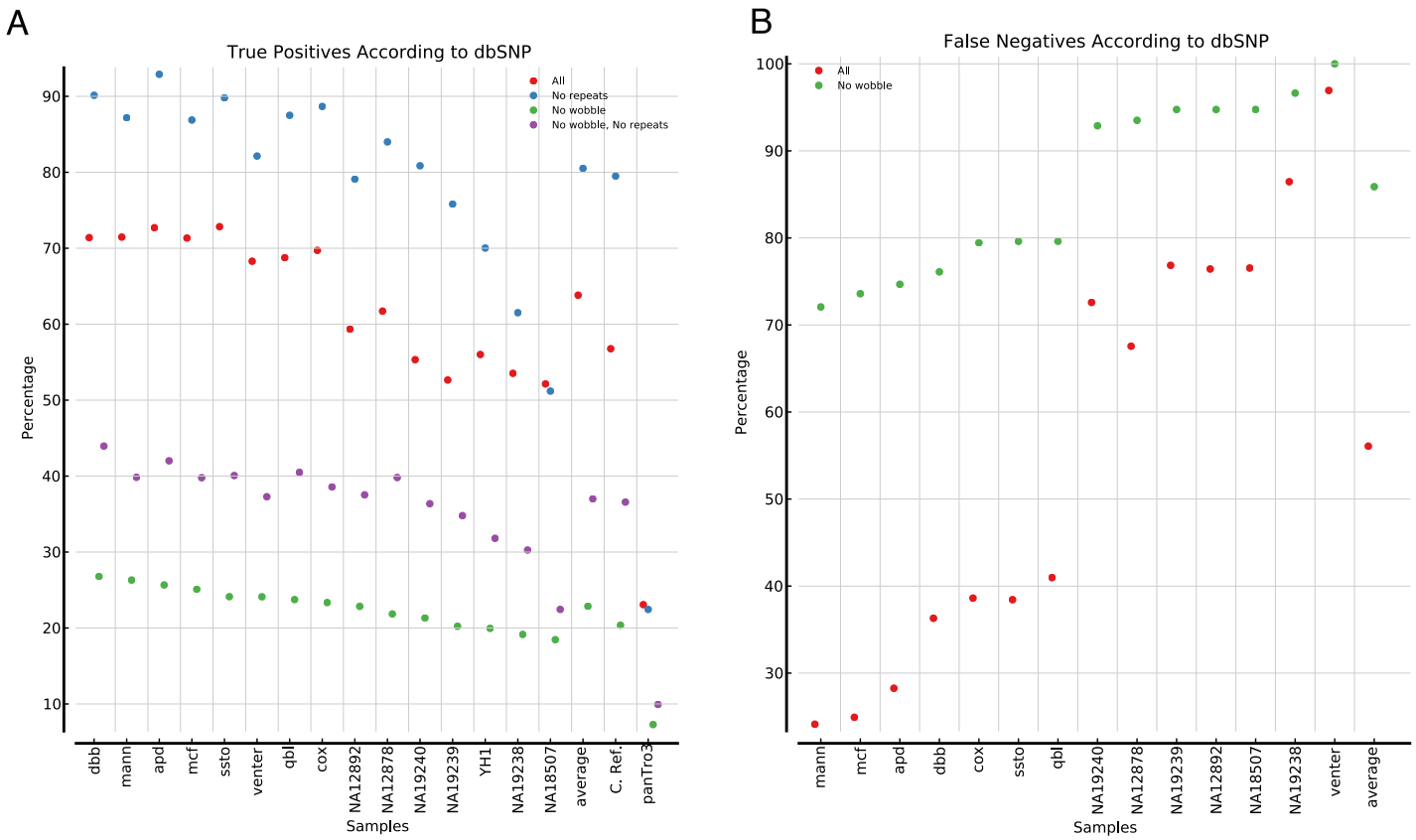

Figure A.3: A detailed comparison of the short indels predicted by the Cactus MSA to those in dbSNP. (A) The proportion of short indels predicted by the MSA with respect to GRCh37 already present in dbSNP. (B) The proportion of (previously reported) short indels for a given sample in dbSNP not predicted by the MSA. Key gives categories of indels predicted by the MSA; 'All': All short indels, allowing for a 5 base disagreement in the local of the indel, 'No repeats': As 'All', but ignoring indels that are labeled repetitive, 'No wobble': As 'All', but not allowing for any disagreement in the location of the indel, 'No wobble, no repeats': intersection of previous two categories.

A.3 Correct Contiguity

Formally, two positions x_i and x_j such that $i < j$ are *contiguous* in a sequence y if there exists two positions y_k and y_l such that (1) $k < l$, (2) $[x_i] = [y_k]$ and (3) $[x_j] = [y_l]$. For a reference (either a chosen input sample or C. Ref.), which is represented as a set of one or more contigs, x and y are *correctly contiguous* (similarly defined in [Earl et al., 2011]) if they are contiguous in the forward or reverse complement of a contig in the reference.

For each contig of length n there are $\binom{n}{2}$ possible pairs of positions, to avoid testing them all 100 million pairs were sampled from each sample with respect to each tested reference. For a contig $x = x_1, x_2 \ldots x_n$ a pair x_i, x_j is selected at random such that a distance $j - i$ is selected with probability proportional to $log_{10}(j - i)/log_{10}(n)$.

dbSNP Validation of Short Insertion Variations

Sample	T#	All			No wobble		
		TP	STP	SFN	TP	STP	SFN
NA12878	796	65	19	63	30	4	93
NA12892	480	60	9	87	30	1	98
NA19238	250	61	14	90	30	3	98
NA19239	526	57	15	79	29	3	96
NA19240	493	62	9	84	31	2	96
apd	353	68	NA	NA	27	NA	NA
cox	1020	66	NA	NA	26	NA	NA
dbb	878	68	NA	NA	30	NA	NA
mann	853	69	NA	NA	31	NA	NA
mcf	781	68	NA	NA	27	NA	NA
nigerian	562	57	12	79	25	2	96
qbl	1025	65	NA	NA	25	NA	NA
ssto	965	69	NA	NA	28	NA	NA
venter	1005	64	0	98	26	0	100
yanhuang	635	56	NA	NA	24	NA	NA
panTro3	3621	25	NA	NA	10	NA	NA
aggregate	10622	64	NA	NA	28	NA	NA
C. Ref.	2469	57	NA	NA	21	NA	NA

Table A.7: All: insertions detected in each sample with respect to GRCh37, allowing a match to an insertion within 5 bases of its location in dbSNP. No wobble: as All, matched precisely to insertions in dbSNP (location and length) T#: Total number of insertions. TP: Percentage true positives, as validated by a match in dbSNP. STP: Percentage (sample) true positives, as validated by those reported for the sample in question. SFN: Percentage (sample) false negatives, as validated by those reported for the sample in question. An NA entry denotes that the data was not available. Aggregate row: gives the total insertions in human samples (excluding chimp). C. Ref. row: gives insertions in C. Ref. with respect to GRCh37

dbSNP Validation of Short Deletion Variations

Sample	T#	All			No wobble		
		TP	STP	SFN	TP	STP	SFN
NA12878	902	59	17	79	15	4	95
NA12892	627	59	21	81	18	6	94
NA19238	387	49	19	88	12	6	96
NA19239	680	50	18	83	13	5	95
NA19240	736	51	18	77	15	5	93
apd	321	78	NA	NA	25	NA	NA
cox	1087	73	NA	NA	21	NA	NA
dbb	884	75	NA	NA	24	NA	NA
mann	903	74	NA	NA	22	NA	NA
mcf	845	74	NA	NA	23	NA	NA
nigerian	700	48	25	77	13	6	94
qbl	1001	73	NA	NA	22	NA	NA
ssto	979	76	NA	NA	21	NA	NA
venter	1048	72	0	99	22	0	100
yanhuang	638	56	NA	NA	16	NA	NA
panTro3	3677	21	NA	NA	4	NA	NA
aggregate	11738	66	NA	NA	19	NA	NA
C. Ref.	168	50	NA	NA	8	NA	NA

Table A.8: All: deletions detected in each sample with respect to GRCh37, allowing a match to an deletion within 5 bases of its location in dbSNP. No wobble: as All, matched precisely to insertions in dbSNP (location and length) T#: Total number of deletions. TP: Percentage true positives, as validated by a match in dbSNP. STP: Percentage (sample) true positives, as validated by those reported for the sample in question. SFN: Percentage (sample) false negatives, as validated by those reported for the sample in question. An NA entry denotes that the data was not available. Aggregate row: gives the total deletions in human samples (excluding chimp). C. Ref. row: gives deletions in C. Ref. with respect to GRCh37

dbSNP Validation of Short Non-Repetitive Insertion Variations

Sample	T#	All			No wobble		
		TP	STP	SFN	TP	STP	SFN
NA12878	298	85	29	51	48	6	90
NA12892	199	77	9	95	46	1	100
NA19238	122	75	24	91	48	6	98
NA19239	215	78	24	86	46	6	97
NA19240	185	86	17	88	49	4	97
apd	80	94	NA	NA	48	NA	NA
cox	272	87	NA	NA	44	NA	NA
dbb	228	89	NA	NA	48	NA	NA
mann	259	85	NA	NA	47	NA	NA
mcf	222	86	NA	NA	44	NA	NA
nigerian	248	62	18	86	32	4	97
qbl	277	86	NA	NA	45	NA	NA
ssto	263	89	NA	NA	48	NA	NA
venter	250	80	0	100	41	0	100
yanhuang	210	77	NA	NA	41	NA	NA
panTro3	1281	26	NA	NA	14	NA	NA
aggregate	3328	82	NA	NA	45	NA	NA
C. Ref.	649	79	NA	NA	37	NA	NA

Table A.9: Experiment and format same as in Table A.7, but only for short insertions at bases not defined as repetitive in GRCh37.

dbSNP Validation of Short Non-Repetitive Deletion Variations

Sample	T#	All			No wobble		
		TP	STP	SFN	TP	STP	SFN
NA12878	227	82	30	79	29	8	94
NA12892	222	81	34	89	30	13	96
NA19238	182	52	27	92	19	10	97
NA19239	219	74	33	90	24	11	96
NA19240	233	77	36	86	26	12	95
apd	89	92	NA	NA	37	NA	NA
cox	275	90	NA	NA	33	NA	NA
dbb	218	92	NA	NA	39	NA	NA
mann	248	89	NA	NA	32	NA	NA
mcf	243	88	NA	NA	36	NA	NA
nigerian	340	44	33	85	15	8	96
qbl	291	89	NA	NA	36	NA	NA
ssto	276	91	NA	NA	33	NA	NA
venter	265	84	0	100	34	0	100
yanhuang	227	63	NA	NA	23	NA	NA
panTro3	1236	19	NA	NA	6	NA	NA
aggregate	3555	78	NA	NA	30	NA	NA
C. Ref.	29	90	NA	NA	31	NA	NA

Table A.10: Experiment and format same as in Table A.8, but only for short deletions at bases not defined as repetitive in GRCh37.

Contiguity Statistics I

Samples	Reference	% M.	% M. & C.	% C. w. M.
NA12878	C. Ref.	98.13	98.13	100.00
	GRCh37	94.88	94.87	99.99
NA12892	C. Ref.	99.17	99.10	99.93
	GRCh37	98.95	98.88	99.93
NA19238	C. Ref.	98.84	98.84	99.99
	GRCh37	98.76	98.76	99.99
NA19239	C. Ref.	98.40	98.38	99.97
	GRCh37	98.08	98.06	99.98
NA19240	C. Ref.	98.44	98.42	99.98
	GRCh37	98.27	98.25	99.98
apd	C. Ref.	99.99	99.99	100.00
	GRCh37	98.64	98.61	99.97
cox	C. Ref.	99.84	99.84	100.00
	GRCh37	96.80	96.78	99.98
dbb	C. Ref.	99.92	99.92	100.00
	GRCh37	96.56	96.54	99.98
mann	C. Ref.	99.63	99.61	99.98
	GRCh37	93.47	93.47	100.00

Table A.11: Statistics on correct contiguity, comparing mapping through the Cactus alignment to either GRCh37 or C. Ref. (part I). '% M.': The proportion of randomly selected pairs that mapped to the reference. '% M. & C.': The proportion of all randomly selected pairs that mapped to the reference and were correctly contiguous. '% C. w. M.': The proportion of randomly selected pairs which mapped to the reference that were correctly contiguous. 'aggregate' row gives average over all samples.

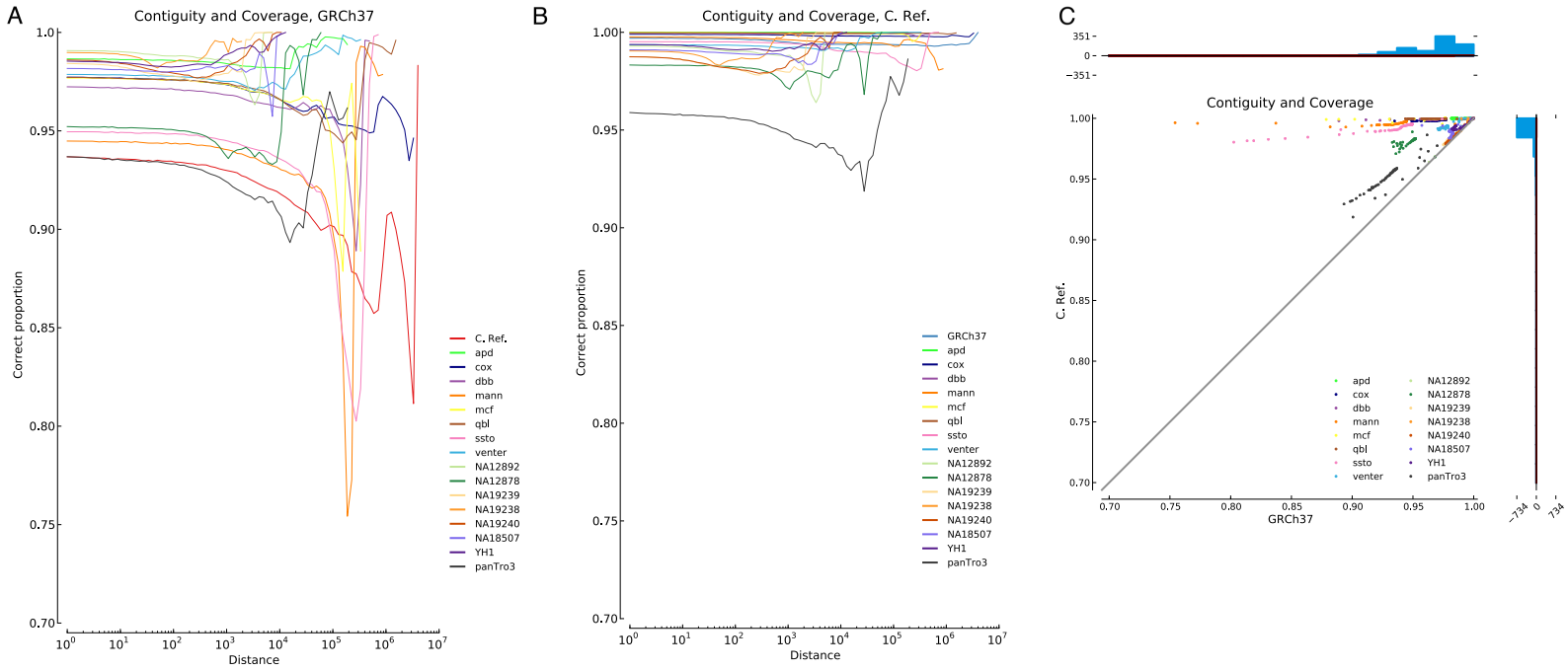

Figure A.4: The proportion of correctly contiguous pairs as a function of the pairs' separation. (A) For GRCh37. (B) For C. Ref. (C) A scatter plot of of correct contiguity in GRCh37 against correct contiguity in C. Ref. for all samples and separation distances. The histograms on the axes show densities of the comparisons. The blue bars indicate points not on the $x = y$ axis, with positive values indicating comparisons for which $y > x$ and negative values indicating comparisons for which $y < x$ in the case of the histogram on the x axis, and in reverse for the histogram on the y axis. The red bars indicate points on the $x = y$ axis.

Contiguity Statistics II

Samples	Reference	% M.	% M. & C.	% C. w. M.
mcf	C. Ref.	99.98	99.96	99.97
	GRCh37	97.04	97.04	100.00
nigerian	C. Ref.	98.96	98.96	100.00
	GRCh37	98.12	98.11	99.99
panTro3	C. Ref.	95.14	95.14	100.00
	GRCh37	92.80	92.78	99.99
qbl	C. Ref.	99.97	99.97	100.00
	GRCh37	97.08	97.06	99.98
ssto	C. Ref.	99.28	99.28	100.00
	GRCh37	93.49	93.47	99.98
venter	C. Ref.	99.34	99.34	100.00
	GRCh37	97.79	97.78	99.99
yanhuang	C. Ref.	99.25	99.24	99.99
	GRCh37	98.51	98.51	99.99
aggregate	C. Ref.	99.02	99.01	99.99
	GRCh37	96.83	96.81	99.98

Table A.12: Statistics on correct contiguity, comparing mapping through the Cactus alignment to either GRCh37 or C. Ref. (part II). '% M.': The proportion of randomly selected pairs that mapped to the reference. '% M. & C.': The proportion of all randomly selected pairs that mapped to the reference and were correctly contiguous. '% C. w. M.': The proportion of randomly selected pairs which mapped to the reference that were correctly contiguous. 'aggregate' row gives average over all samples.

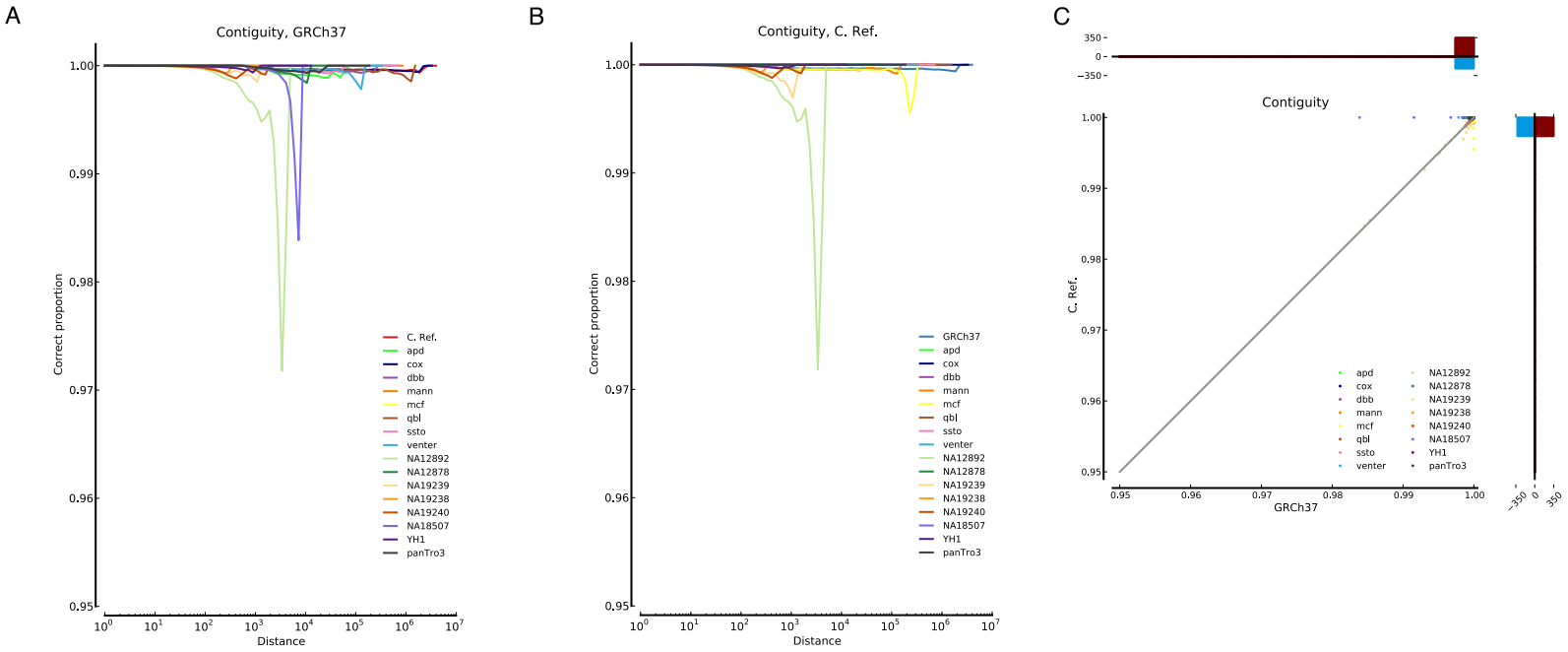

Figure A.5: The proportions of mapping pairs that are correctly contiguous. The figure is arranged otherwise identically to Figure A.4.

A.4 Mapping Large Indels

Predicted insertions with respect to GRCh37 larger than 1000 bases were aligned to the GRCh37 reference assembly (obtained from the UCSC Genome Browser database, excluding the alternative loci) using LASTZ. The LASTZ parameters used were *–identity=90 –chain* and both the query and target were set as *[unmask]*.

An insertion was labelled 'Mapped' if $\geq 50\%$ of its bases mapped and 'Unmapped' otherwise. A mapped insertion was classified as 'Mapped to MHC' if it mapped best within the MHC region and 'Mapped Outside MHC' vice versa. An insertion was marked as 'Multi-mapping' if $\geq 50\%$ of its bases mapped to multiple locations, 'Repeats' if $\geq 50\%$ of its bases classified by Repeat Masker as repetitive [Smit et al., 2010].

Predicted insertions of each sample were also aligned to the sample's sequence to assess the copy number changes, using the same alignment program and settings as described above. A 'copy number change' was recorded when $\geq 90\%$ of an insertion's bases mapped to multiple locations.

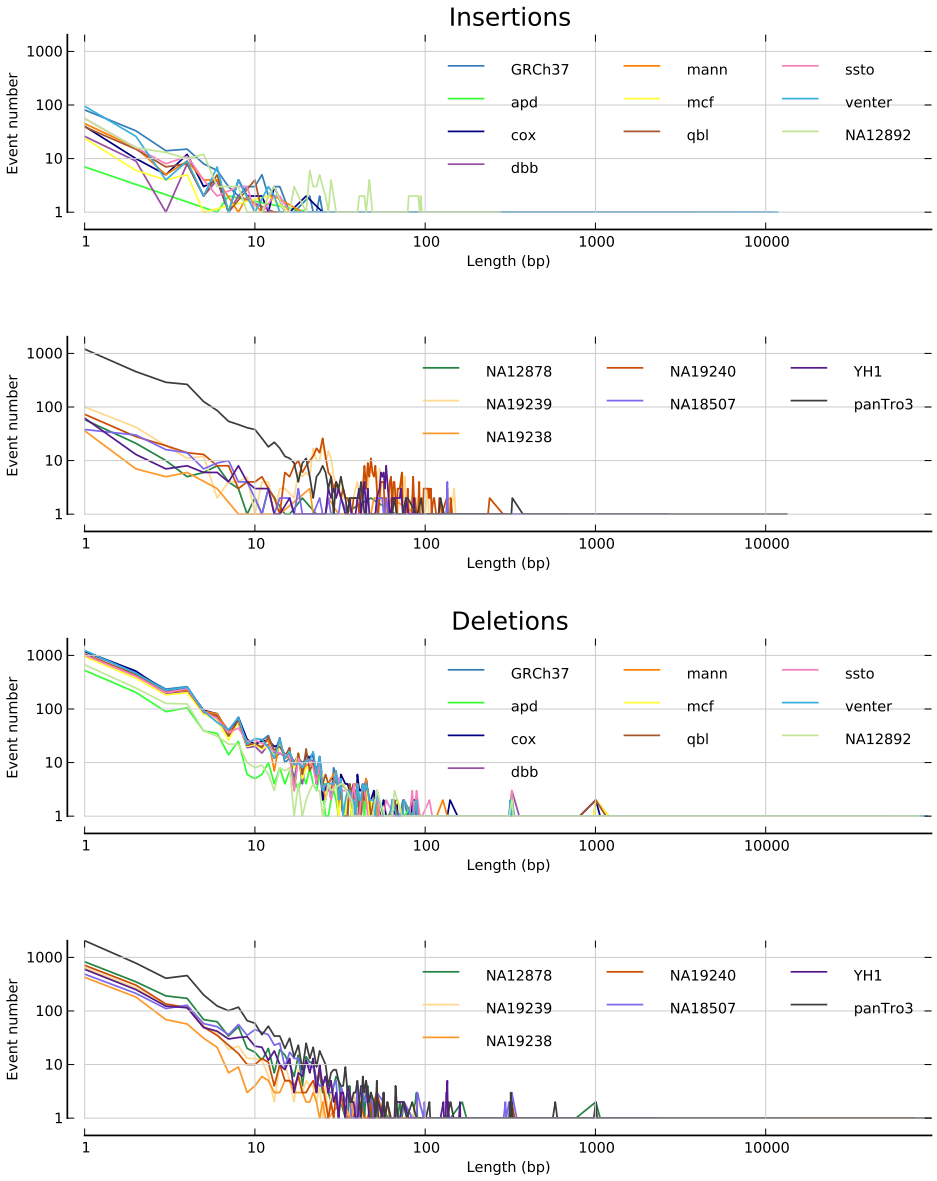

Figure A.6: Length distributions of insertion and deletion events with respect to C. Ref. The top two panels show insertion lengths for each sample and the bottom two panels show deletion lengths for each sample.

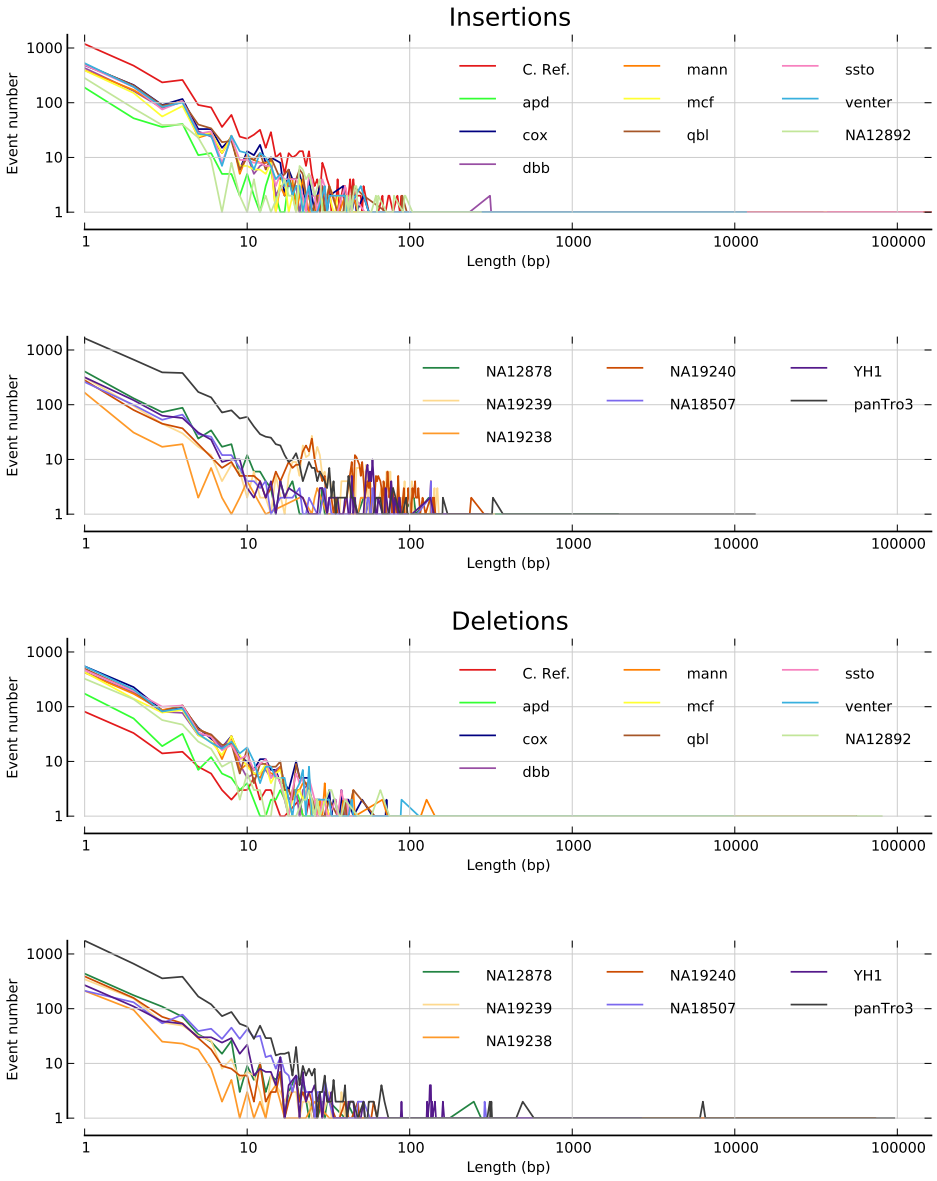

Figure A.7: Length distributions of insertion and deletion events with respect to GRCh37. The top two panels show insertion lengths for each sample and the bottom two panels show deletion lengths for each sample.

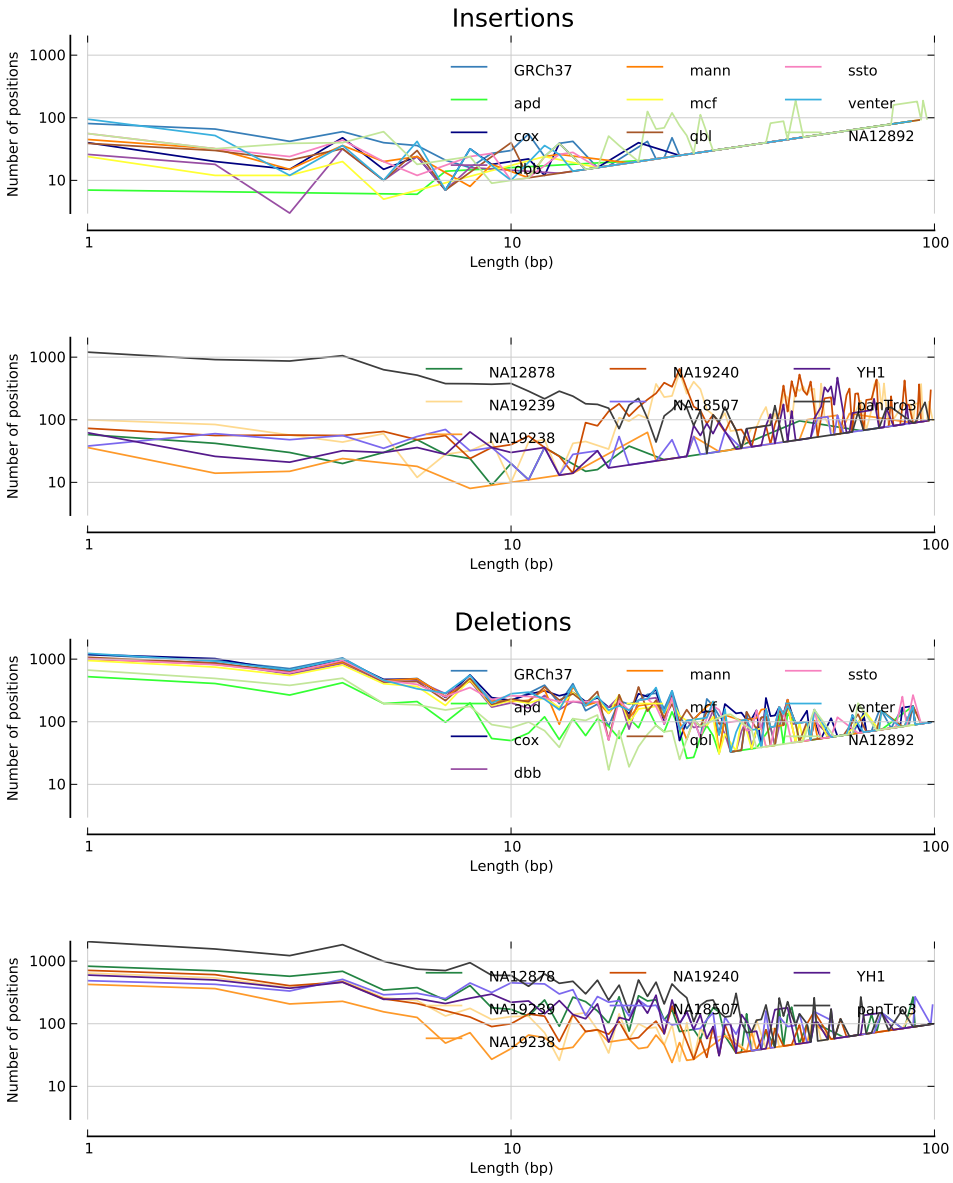

Figure A.8: The number of bases per sample effected by insertions and deletions of a given length as a function insertion/deletion length. Insertions and deletions are with respect to C. Ref. The top two panels show results for insertions and the bottom two panels show results for deletions.

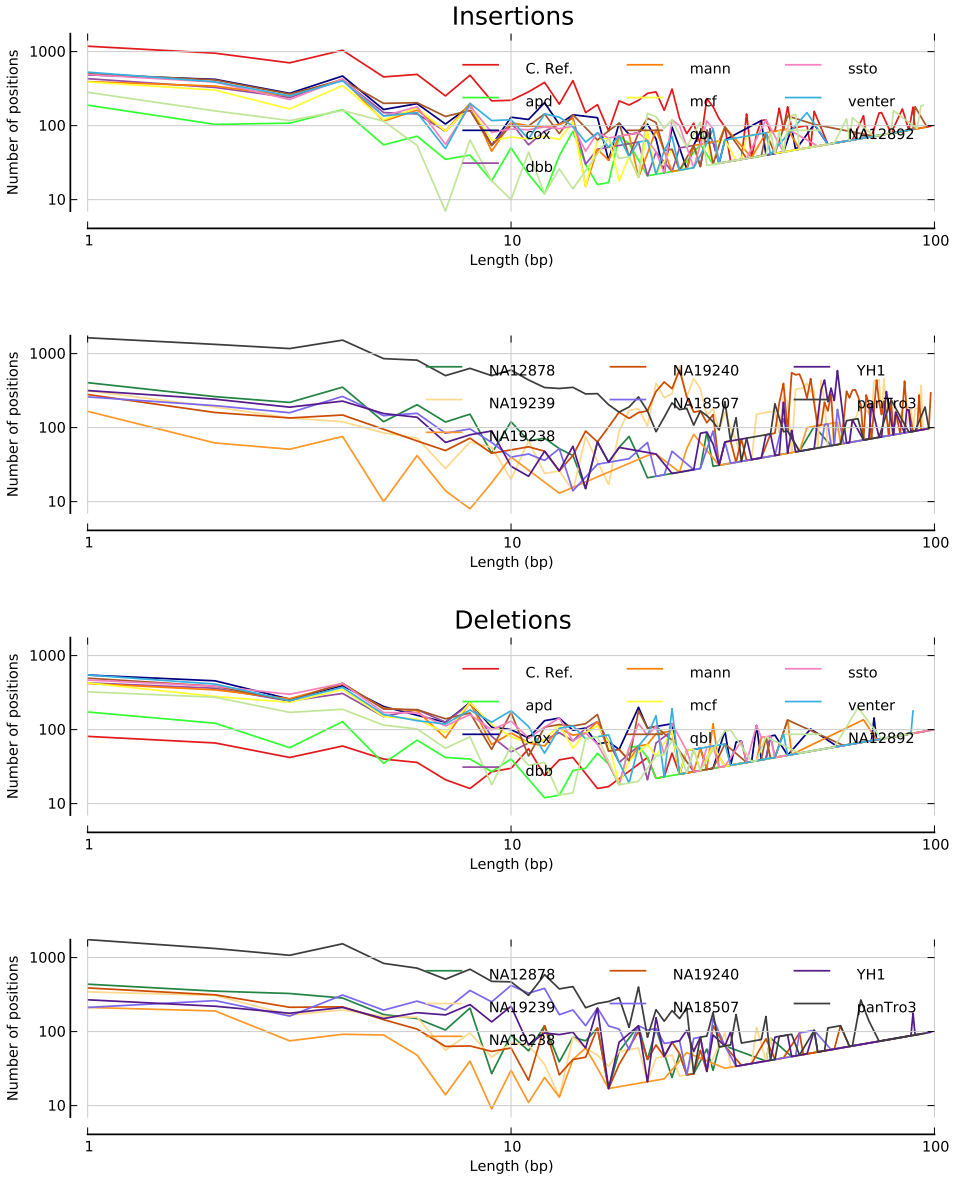

Figure A.9: The number of bases per sample effected by insertions and deletions of a given length as a function insertion/deletion length. Insertions and deletions are with respect to GRCh37. The top two panels show results for insertions and the bottom two panels show results for deletions.

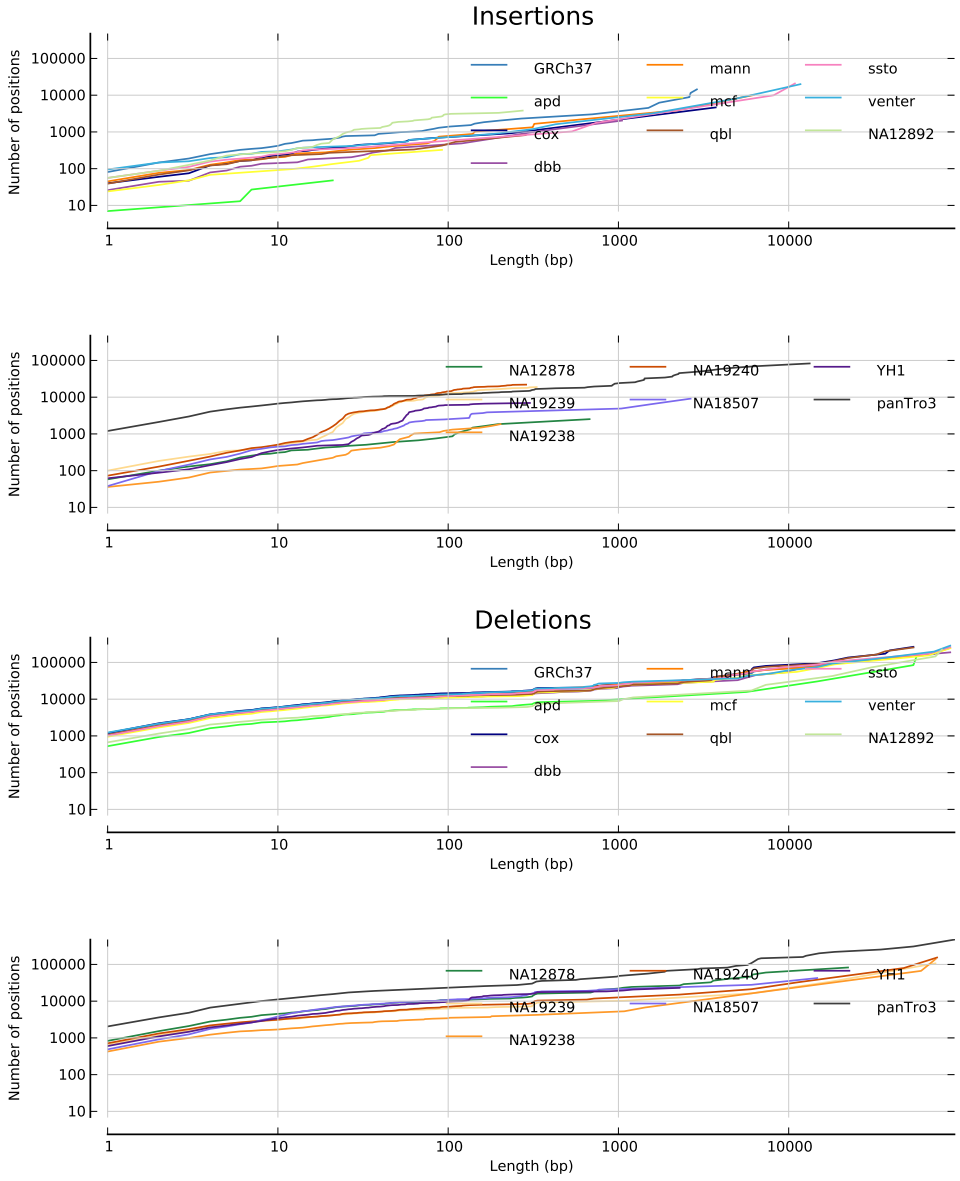

Figure A.10: The cumulative total length of insertion and deletions events as a function of indel event length, with respect to C. Ref. The top two panels show cumulative insertion lengths for each sample and the bottom two panels show cumulative deletion lengths for each sample.

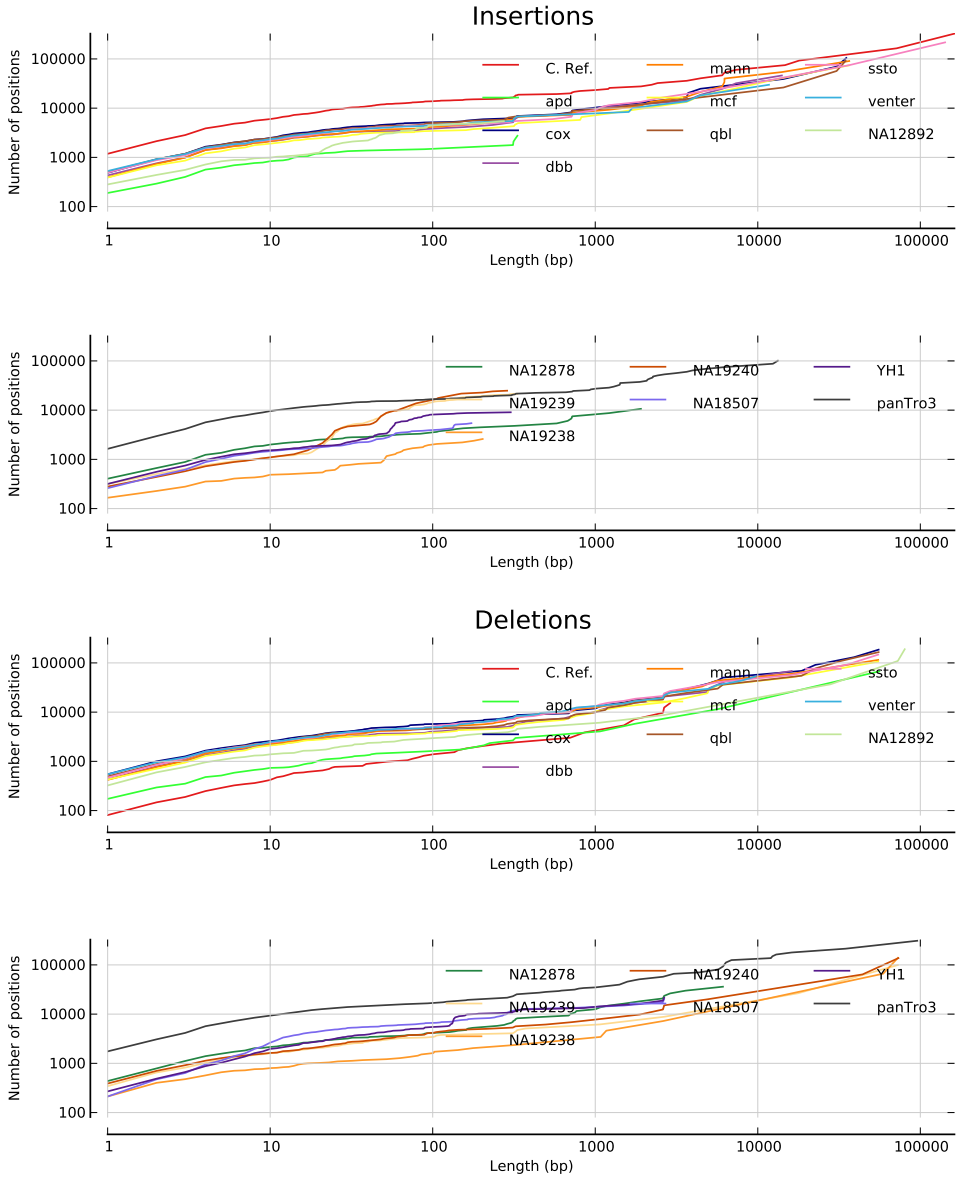

Figure A.11: The cumulative total length of insertion and deletions events as a function of indel event length, with respect to GRCh37. The top two panels show cumulative insertion lengths for each sample and the bottom two panels show cumulative deletion lengths for each sample.

	90%		95%	
	C. Ref.	GRCh37	C. Ref.	GRCh37
Genbank RNA	3209	2986	3095	2881
RefSeq transcript	371	368	370	367
RefSeq genes all tx mapped	213	210	212	209
RefSeq genes ≥ 1 tx mapped	213	211	212	210

Table A.13: Statistics on RNAs and RefSeq transcripts mapping to either references, GRCh37 or C. Ref. Columns: '90%': RNAs must have at least 90% bases aligned to the reference, '95%': RNAs must have at least 95% bases aligned to the reference, 'C. Ref. ': the reference is C. Ref., 'GRCh37': the reference is GRCh37 MHC main locus. Rows: 'Genbank RNA': number of Genbank RNAs mapped best to the appropriate reference with the appropriate base coverage. 'RefSeq transcript': similar to 'Genbank RNAs' but for RefSeq transcripts instead of Genbank RNAs. 'RefSeq genes all tx mapped': number of RefSeq genes that have all the transcripts mapped best to the appropriate reference with the appropriate base coverage. 'RefSeq genes ≥ 1 tx mapped': number of RefSeq genes that have at least one transcript mapped best to the appropriate reference with the appropriate base coverage.

A.5 Gene Mapping

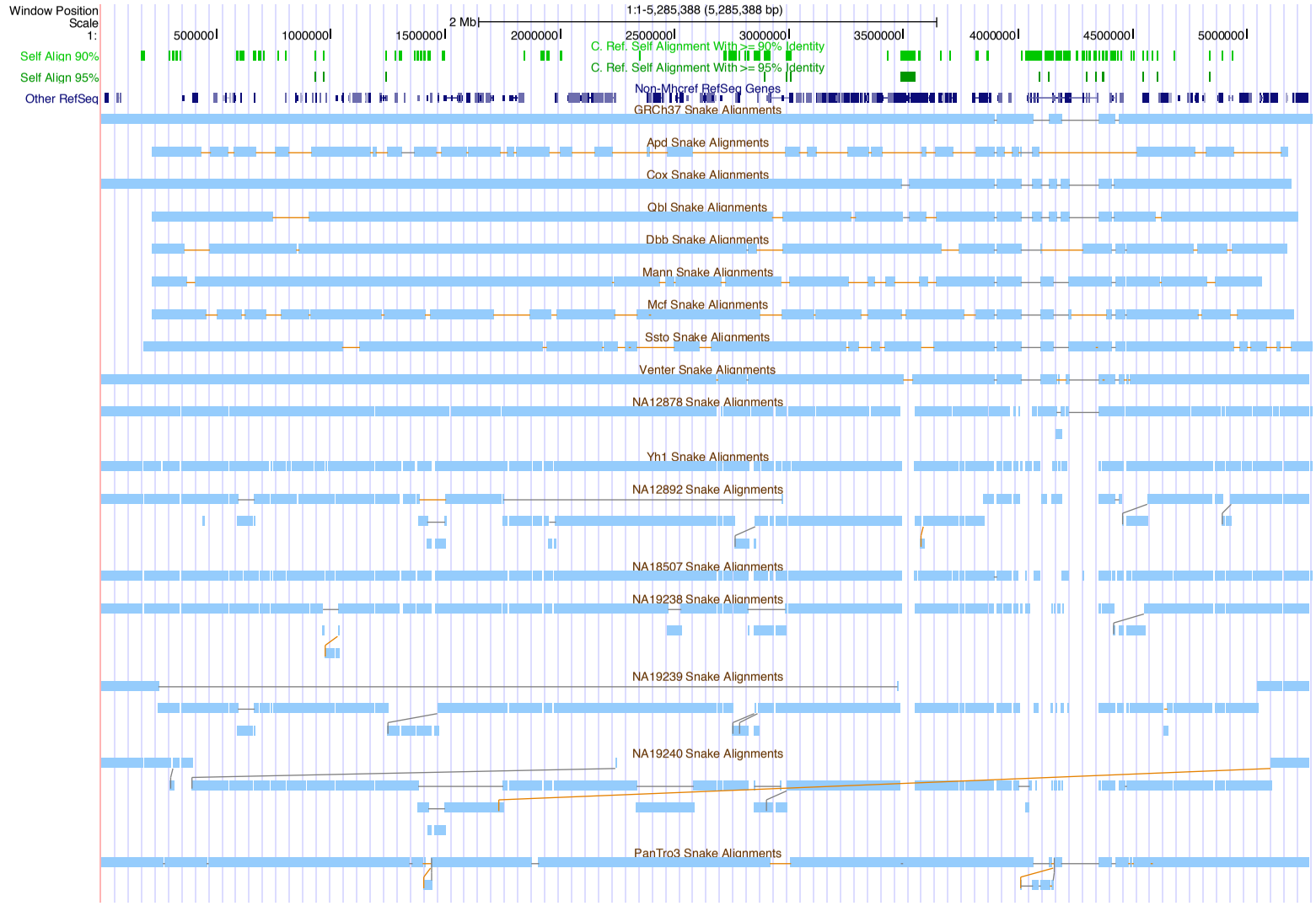

Figure A.12: A UCSC Browser screenshot showing a prototype C. Ref. MHC reference browser. The figure is arranged similarly to Figure A.14

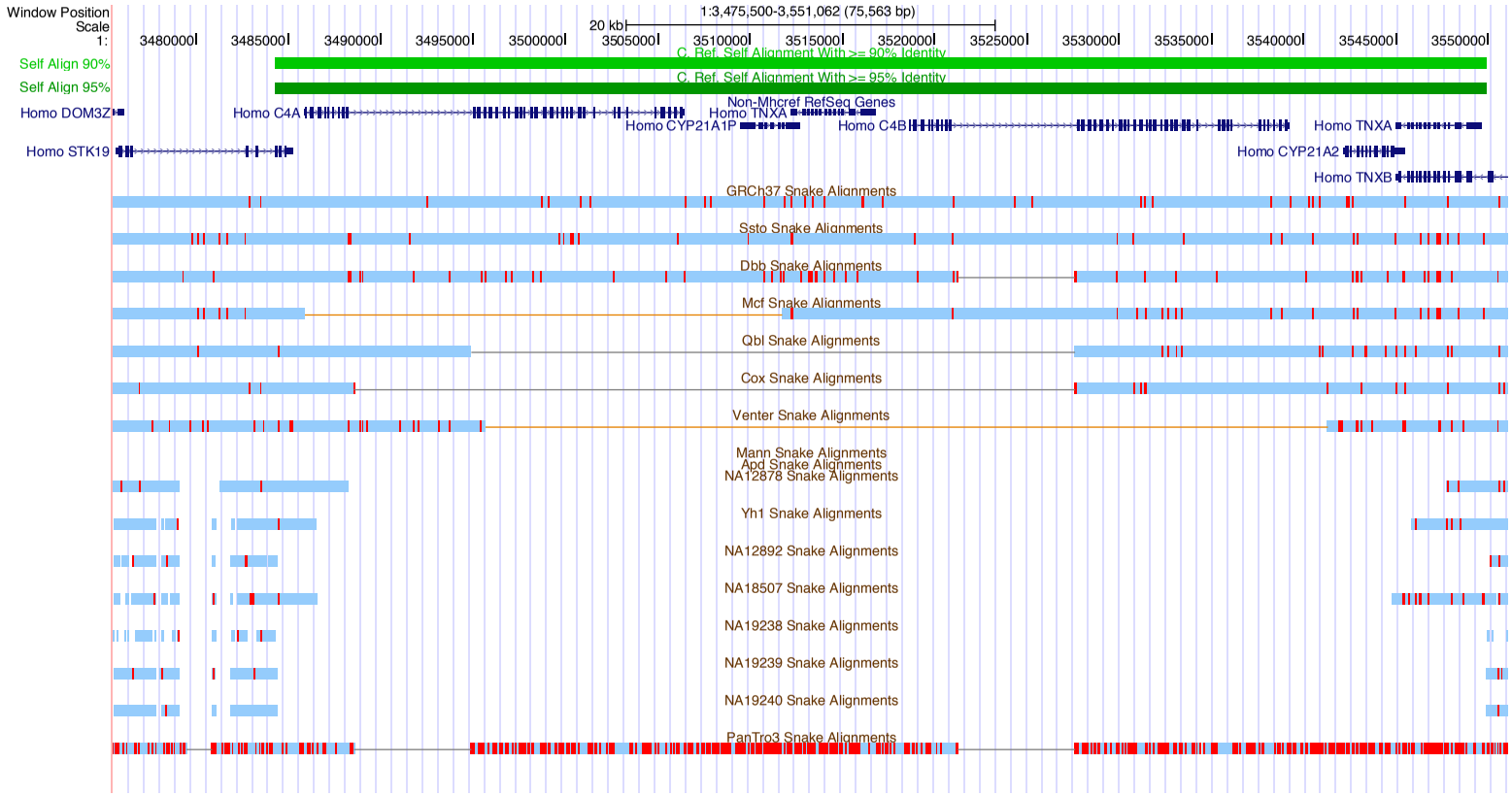

Figure A.13: A UCSC Browser display showing the RCCX gene region in a prototype C. Ref. MHC reference browser. The figure is arranged similarly to Figure A.14

A.6 Short Read Mapping

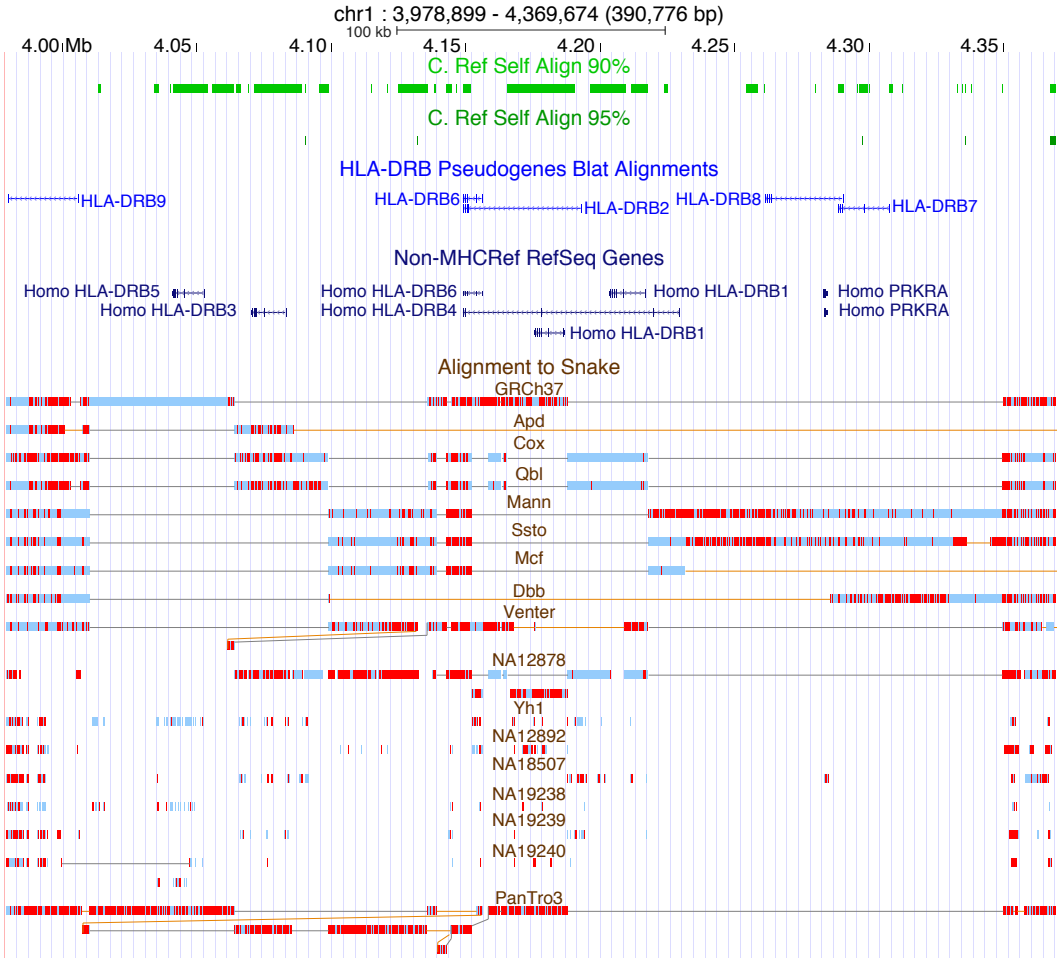

Figure A.14: A UCSC Browser display [Fujita et al., 2011] of the MHC HLA-DRB hypervariable region in a prototype C. Ref. MHC reference browser. Self Align tracks: Alignment of the region against itself with a 90% and 95% minimum identity threshold. It demonstrates that much of the region is homologous to itself at 90% identity, but very little at 95%, which is substantially below the threshold required in the MSA to create homology. Gene tracks: Genes identified by alignment and using RefSeq annotations (see Supplementary Section *A*.5). Snake tracks: Subsequences of contiguous bases aligned to the reference are shown as rectangles. SNVs with respect to the reference are coloured red, otherwise bases are coloured light blue. The lines connecting the rectangles show adjacencies between the bases. In addition to genes that are presnt in GRCh37 (known genes HLA-DRB5, HLA-DRB1 and pseudogenes HLA-DRB9, HLA-DRB6), C. Ref. also contains genes that are recurrent in the input samples (HLA-DRB3, HLA-DRB4 and pseudogenes HLA-DRB2, HLA-DRB7, HLA-DRB8). The MSA shows clearly the relationship of the samples in the region, e.g COX and QBL have the same DRB group and are grouped together. Lines coloured orange indicate adjacencies that contain unaligned bases only present in one sample.

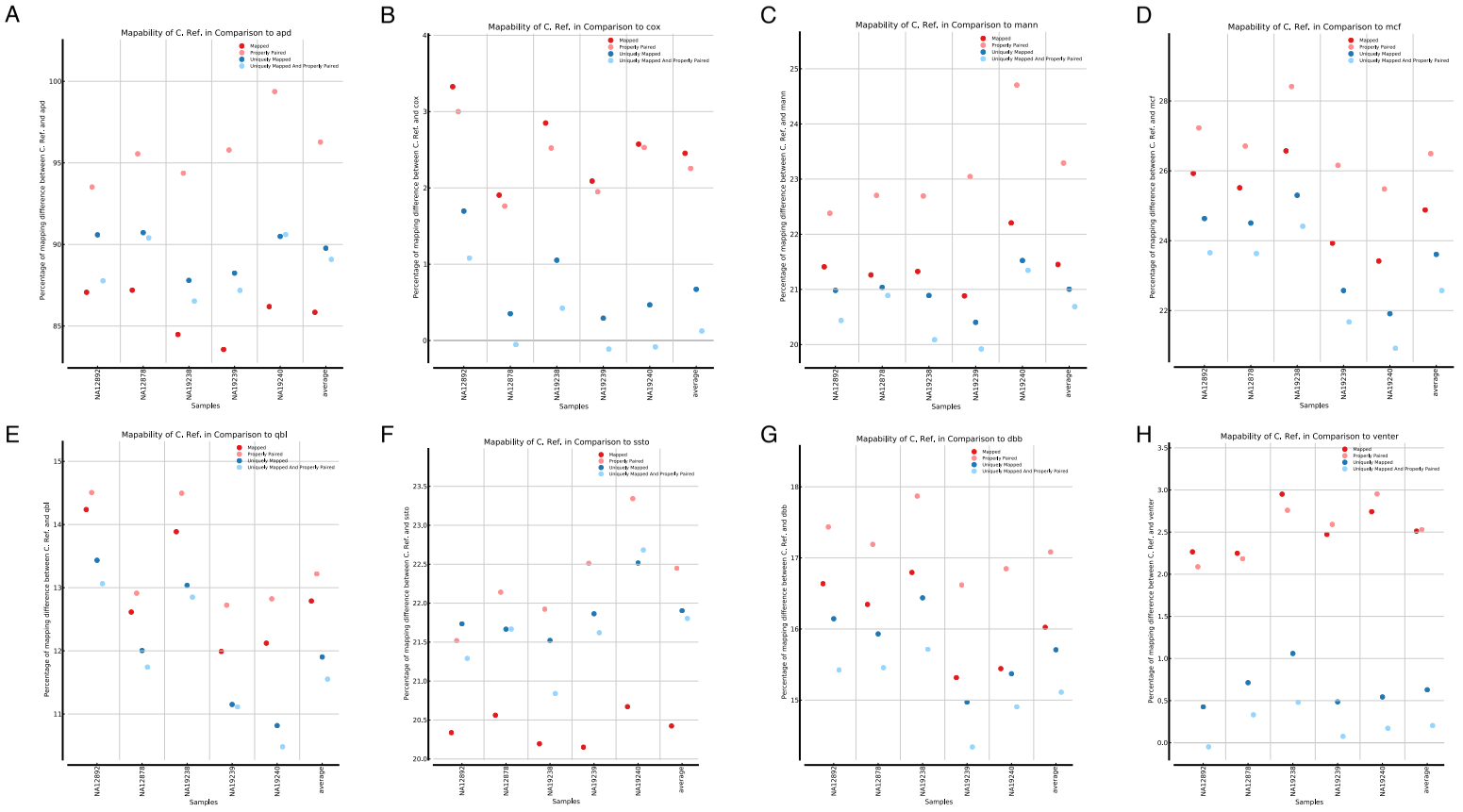

Figure A.15: Comparing mapping to C. Ref. against mapping to each of the 7 alternative haplotypes in GRCh37 and the Venter assembly. Each panel represents an experiment is as described in Figure 2.15 of the main text, but instead of using the GRCh37 sequence uses the alternative sequence.

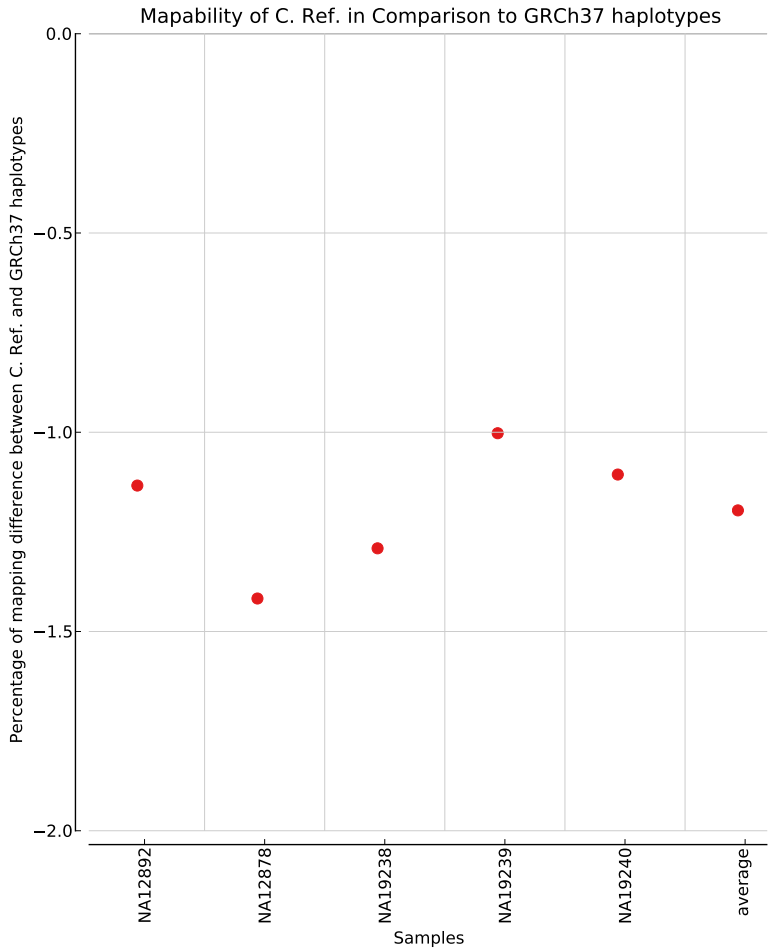

Figure A.16: Comparing mapping to C. Ref. against mapping to all 8 haplotypes in GRCh37. Figure as described in Figure 2.15 of the main text, but only showing the overall mapping ('Mapped') of the reads.

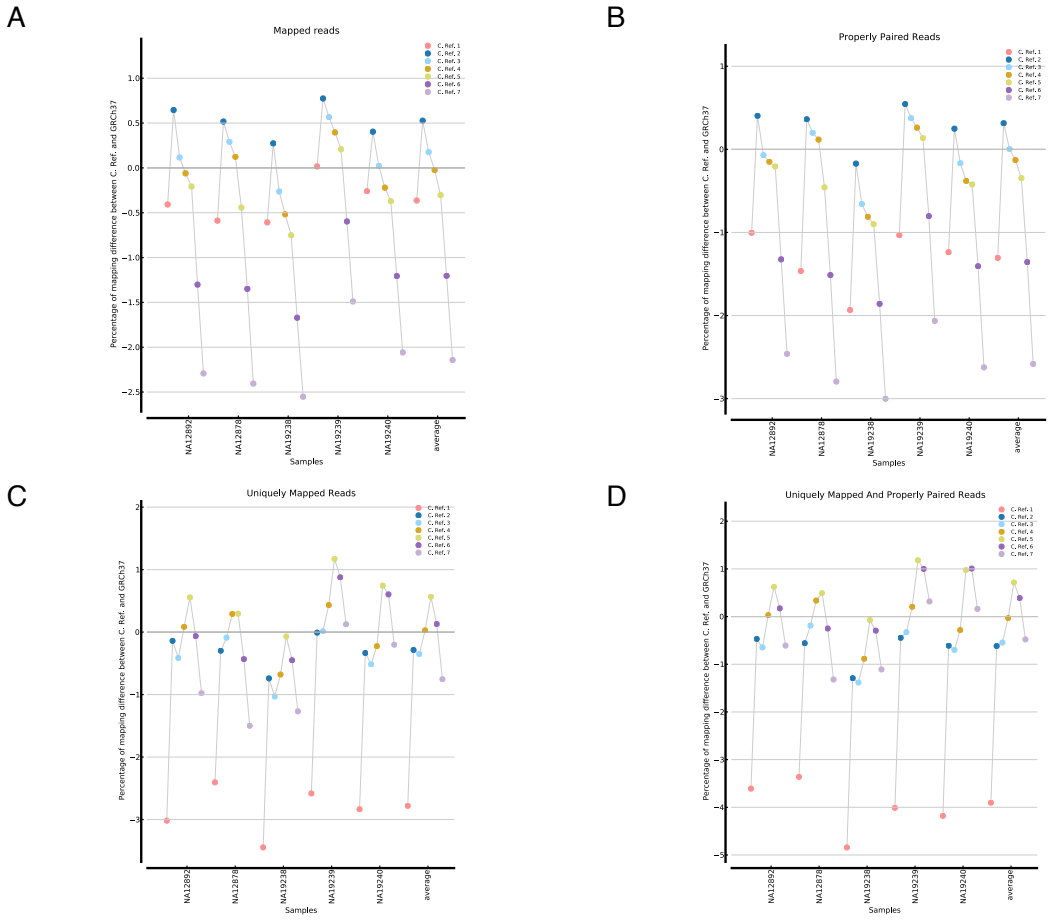

Figure A.17: Comparing mapping to GRCh37 against mapping to consensus references with different α values. The panels represent experiments as described in Figure 2.15 of the main text, but the experiment is rerun separately with multiple different consensus references each constructed with a different α parameter (see Supplementary Methods for α definition).

Mapping Stats I

Sample	Ref	Mapped	UniqMapped	PropPaired	UniqPropPaired	Snp
NA12878	C. Ref. 1	4,964,962	4,458,844	3,811,844	3,299,812	25,571
	C. Ref. 2	5,020,181	4,555,066	3,882,466	3,395,542	18,523
	C. Ref. 3	5,008,893	4,564,653	3,876,196	3,408,160	18,880
	C. Ref. 4	5,000,528	4,581,964	3,873,002	3,426,090	18,949
	C. Ref. 5	4,972,288	4,582,196	3,850,758	3,431,382	18,707
	C. Ref. 6	4,926,976	4,548,921	3,809,922	3,406,056	18,221
	C. Ref. 7	4,874,187	4,500,244	3,760,360	3,369,542	18,399
	GRCh37	4,994,346	4,568,699	3,868,460	3,414,600	22,137
NA12892	C. Ref. 1	2,923,764	2,575,580	1,738,820	1,472,756	19,252
	C. Ref. 2	2,954,688	2,651,991	1,763,542	1,520,720	15,664
	C. Ref. 3	2,939,156	2,644,708	1,755,224	1,518,000	15,640
	C. Ref. 4	2,933,959	2,657,935	1,753,834	1,528,358	16,304
	C. Ref. 5	2,929,642	2,670,457	1,752,836	1,537,422	16,615
	C. Ref. 6	2,897,518	2,654,008	1,733,200	1,530,522	16,206
	C. Ref. 7	2,868,441	2,629,738	1,713,216	1,518,590	15,823
	GRCh37	2,935,739	2,655,709	1,756,452	1,527,884	17,907
NA19238	C. Ref. 1	2,522,686	2,188,623	1,405,436	1,170,794	18,095
	C. Ref. 2	2,545,061	2,249,935	1,430,686	1,214,496	15,490
	C. Ref. 3	2,531,437	2,243,349	1,423,710	1,213,366	15,735
	C. Ref. 4	2,524,981	2,251,325	1,421,518	1,219,496	16,085
	C. Ref. 5	2,519,049	2,265,111	1,420,232	1,229,462	16,425
	C. Ref. 6	2,495,696	2,256,519	1,406,498	1,226,752	16,381
	C. Ref. 7	2,473,352	2,237,964	1,390,130	1,216,742	15,939
	GRCh37	2,538,112	2,266,716	1,433,142	1,230,378	16,378

Table A.14: Comparing mapping to GRCh37 against mapping to consensus references with different α values (part 1). The panels represent experiments as described in Figure 2.15 of the main text, but the experiment is rerun separately with multiple different consensus references each constructed with a different α parameter.

Mapping Stats II

Sample	Ref	Mapped	UniqMapped	PropPaired	UniqPropPaired	Snp
NA19239	C. Ref. 1	3,665,300	3,159,440	2,609,902	2,115,094	21,220
	C. Ref. 2	3,693,064	3,242,893	2,651,538	2,193,666	16,631
	C. Ref. 3	3,685,454	3,243,577	2,647,064	2,196,286	16,629
	C. Ref. 4	3,679,201	3,257,217	2,644,056	2,208,004	17,081
	C. Ref. 5	3,672,312	3,281,159	2,640,758	2,229,518	17,580
	C. Ref. 6	3,642,797	3,271,596	2,615,962	2,225,512	17,899
	C. Ref. 7	3,610,092	3,247,236	2,582,716	2,210,470	17,533
	GRCh37	3,664,690	3,243,169	2,637,162	2,203,484	22,409
NA19240	C. Ref. 1	4,648,881	4,031,496	3,816,132	3,125,620	25,579
	C. Ref. 2	4,679,817	4,135,223	3,873,576	3,242,018	19,293
	C. Ref. 3	4,662,098	4,127,718	3,857,512	3,239,130	19,712
	C. Ref. 4	4,650,728	4,139,751	3,849,304	3,252,728	20,118
	C. Ref. 5	4,643,662	4,179,960	3,847,652	3,293,806	20,540
	C. Ref. 6	4,604,785	4,174,124	3,809,620	3,294,794	20,273
	C. Ref. 7	4,565,070	4,140,714	3,762,602	3,267,200	19,486
	GRCh37	4,660,996	4,149,095	3,863,974	3,261,942	21,676
average	C. Ref. 1	3,745,118	3,282,796	2,676,426	2,236,815	25,571
	C. Ref. 2	3,778,562	3,367,021	2,720,361	2,313,288	18,523
	C. Ref. 3	3,765,407	3,364,801	2,711,941	2,314,988	18,880
	C. Ref. 4	3,757,879	3,377,638	2,708,342	2,326,935	18,949
	C. Ref. 5	3,747,390	3,395,776	2,702,447	2,344,318	18,707
	C. Ref. 6	3,713,554	3,381,033	2,675,040	2,336,727	18,221
	C. Ref. 7	3,678,228	3,351,179	2,641,804	2,316,508	18,399
	GRCh37	3,758,776	3,376,677	2,711,838	2,327,657	22,137

Table A.15: Comparing mapping to GRCh37 against mapping to consensus references with different α values (part 2). The panels represent experiments as described in Figure 2.15 of the main text, but the experiment is rerun separately with multiple different consensus references each constructed with a different α parameter.

Sample	MD Reads	Total MD BP	% Repeats	SNV Rate	bcftools ESR
NA12878	16,914	212,110	71.61	0.0080	2.1992
NA12892	10,049	140,475	72.42	0.0071	2.3092
NA19238	8,867	120,511	72.29	0.0068	2.2446
NA19239	15,285	158,813	70.18	0.0067	2.0362
NA19240	20,016	198,664	72.27	0.0076	2.0203
average	14,226	166,114	71.75	0.0072	2.1557

Table A.16: An analysis of reads that mapped uniquely to C. Ref. but non-uniquely to GRCh37. Table has same format as Table 2.2 in the main text.

Appendix B

Supplement for: "Comparative assembly hubs: web accessible browsers for comparative genomics"

B.1 *E. coli* KO11FL 162099 KO11_***** genes

List of the KO11_***** genes in Figure 3.6:

KO11_18970, KO11_18930, KO11_18890, KO11_18850, KO11_18810, KO11_18770, KO11_18730, KO11_18690, KO11_18650, KO11_18610, KO11_18570, KO11_18530, KO11_18490, KO11_18450, KO11_18410, KO11_18370, KO11_18330, KO11_18290, KO11_18250.

B.2 Gene annotation corrections

In a number of genomes, I observed and corrected obvious errors, such as genes with positions that were out of range of the sequence length and genes with multiple exons that overlapped with each other (self-overlapped). The corrections are listed below.

B.2.1 Out of range genes

There are a number of genes that have the annotated start and end positions lying out of range of the corresponding genome assembly (Sup. Table B.2.2). I removed those genes out of the genome gene annotations.

B.2.2 Self-folded genes

I have noticed and reported here a list of coding genes with multiple exons that overlapped with each other (Sup. Table B.2.2). For example, gene NP_288053.1 of genome O157 H7 EDL933, sequence NC_002655, has two CDS entries: one ranges from 2369166 to 2370296, and one ranges from 2370296 to 2370979. These two regions overlap with each other by one base (2370296). The concatenated sequence of these two regions (with base 2370296 appearring 2 times) has a length of 1815 bp and translates into the correct protein sequence. I suspect that this error is likely an error of the assembly. There are 6 genomes with this type of self-folded error.

***E. coli* and *Shigella* spp. Genome Information (Part I)**

NCBI Genome Name	Genome Short Name	Phylogroup	Pathotype
EscherichiaColiSe15Uid161939	SE15	B2	Commensal
EscherichiaColiNa114Uid162139	NA114	B2	UPEC
EscherichiaColiEd1aUid59379	ED1a	B2	Commensal
EscherichiaColiO83H1Nrg857cUid161987	O83 H1 NRG 857C	B2	AIEC
EscherichiaColiLf82Uid161965	LF82	B2	AIEC
EscherichiaColiAbu83972Uid161975	ABU 83972	B2	ABU
EscherichiaColiCft073Uid57915	CFT073	B2	UPEC
EscherichiaColiCloneDI14Uid162049	D i14	B2	UPEC
EscherichiaColiCloneDI2Uid162047	D i2	B2	UPEC
EscherichiaColi536Uid58531	536	B2	UPEC
EscherichiaColiUti89Uid58541	UTI89	B2	UPEC
EscherichiaColiUm146Uid162043	UM146	B2	AIEC
EscherichiaColiIhe3034Uid162007	IHE3034	B2	ExPEC
EscherichiaColiS88Uid62979	S88	B2	ExPEC
EscherichiaColiApecO1Uid58623	APEC O1	B2	APEC
EscherichiaColiO127H6E234869Uid59343	O127 H6 E2348 69	B2	EPEC
EscherichiaColiSms35Uid58919	SMS 3 5	D2	ExPEC
EscherichiaColiO7K1Ce10Uid162115	O7 K1 CE10	D2	NMEC
EscherichiaColiIai39Uid59381	IAI39	D2	ExPEC
EscherichiaColi042Uid161985	042	D1	EAEC
EscherichiaColiUmn026Uid62981	UMN026	D1	ExPEC
ShigellaDysenteriaeSd197Uid58213	Sd197	S	S

Table B.1: Summary information of *E. coli* and *Shigella* spp. genomes (part I). 'NCBI Genome Name': the genome name as they appeared on the NCBI ftp website ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ in January 2013. 'Genome Short Name': genome names used in this report. 'Phylogroup' and 'Pathotype': annotated genome phylogroup and pathotype from literature.

***E. coli* and *Shigella* spp. Genome Information (Part II)**

NCBI Genome Name	Genome Short Name	Phylogroup	Pathotype
EscherichiaColiXuzhou21Uid163995	Xuzhou21	E	EHEC
EscherichiaColiO157H7SakaiUid57781	O157 H7 Sakai	E	EHEC
EscherichiaColiO157H7Edl933Uid57831	O157 H7 EDL933	E	EHEC
EscherichiaColiO157H7Ec4115Uid59091	O157 H7 EC4115	E	EHEC
EscherichiaColiO157H7Tw14359Uid59235	O157 H7 TW14359	E	EHEC
EscherichiaColiO55H7Cb9615Uid46655	O55 H7 CB9615	E	EPEC
EscherichiaColiO55H7Rm12579Uid162153	O55 H7 RM12579	E	EPEC
EscherichiaColiAtcc8739Uid58783	ATCC 8739	A	Commensal
EscherichiaColiHsUid58393	HS	A	Commensal
EscherichiaColiUmnk88Uid161991	UMNK88	A	ETEC
EscherichiaColiEtecH10407Uid161993	ETEC H10407	A	ETEC
EscherichiaColiP12bUid162061	P12b	A	Lab
EscherichiaColiBw2952Uid59391	BW2952	A	Commensal
EscherichiaColiK12SubstrDh10bUid58979	K12 DH10B	A	Commensal
EscherichiaColiK12SubstrMg1655Uid57779	K12 MG1655	A	Commensal
EscherichiaColiK12SubstrW3110Uid161931	K12 W3110	A	Commensal
EscherichiaColiDh1Uid161951	DH1 161951	A	Lab
EscherichiaColiDh1Uid162051	DH1 162051	A	Lab
EscherichiaColiBRel606Uid58803	B REL606	A	Commensal
EscherichiaColiBl21De3Uid161947	BL21 DE3 161947	A	Commensal
EscherichiaColiBl21De3Uid161949	BL21 DE3 161949	A	Commensal
EscherichiaColiBl21GoldDe3PlyssAgUid59245	BL21 Gold DE3 pLysS AG	A	Commensal

Table B.2: Summary information of *E. coli* and *Shigella* spp. genomes (part II). 'NCBI Genome Name': the genome name as they appeared on the NCBI ftp website ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ in January 2013. 'Genome Short Name': genome names used in this report. 'Phylogroup' and 'Pathotype': annotated genome phylogroup and pathotype from literature.

***E. coli* and *Shigella* spp. Genome Information (Part III)**

NCBI Genome Name	Genome Short Name	Phylogroup	Pathotype
EscherichiaColiWUid162011	W	B1	Commensal
EscherichiaColiWUid162101	W	B1	Commensal
EscherichiaColiKo11flUid162099	KO11FL 162099	B1	Commensal
EscherichiaColiKo11flUid52593	KO11FL 52593	B1	Commensal
EscherichiaColiSe11Uid59425	SE11	B1	Commensal
EscherichiaColiIai1Uid59377	IAI1	B1	Commensal
EscherichiaColi55989Uid59383	55989	B1	EAEC
EscherichiaColiO104H42009el2050Uid175905	O104 H4 2009EL 2050	B1	EAEC
EscherichiaColiO104H42009el2071Uid176128	O104 H4 2009EL 2071	B1	EAEC
EscherichiaColiO104H42011c3493Uid176127	O104 H4 2011C 3493	B1	EAEC
EscherichiaColiO103H212009Uid41013	O103 H2 12009	B1	EHEC
EscherichiaColiE24377aUid58395	E24377A	B1	ETEC
EscherichiaColiO111H11128Uid41023	O111 H 11128	B1	EHEC
EscherichiaColiO26H1111368Uid41021	O26 H11 11368	B1	EHEC
ShigellaBoydiiCdc308394Uid58415	SbCDC 3083 94	S	S
ShigellaBoydiiSb227Uid58215	Sb227	S	S
ShigellaSonnei53gUid84383	Ss53G	S	S
ShigellaSonneiSs046Uid58217	Ss046	S	S
ShigellaFlexneri2002017Uid159233	Sf2002017	S	S
ShigellaFlexneri2a2457tUid57991	Sf2a 2457T	S	S
ShigellaFlexneri2a301Uid62907	Sf2a 301	S	S
ShigellaFlexneri58401Uid58583	Sf5 8401	S	S

Table B.3: Summary information of *E. coli* and *Shigella* spp. genomes (part III). 'NCBI Genome Name': the genome name as they appeared on the NCBI ftp website ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ in January 2013. 'Genome Short Name': genome names used in this report. 'Phylogroup' and 'Pathotype': annotated genome phylogroup and pathotype from literature.

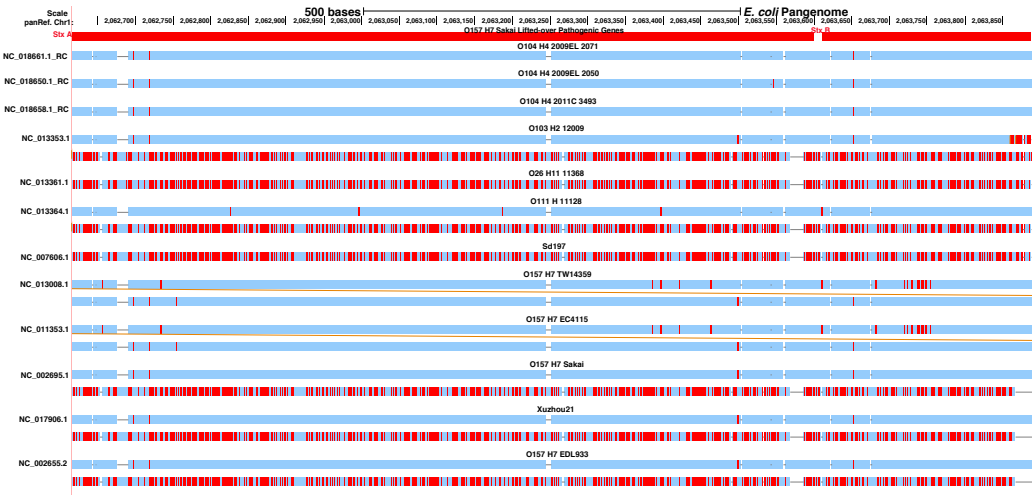

Figure B.1: The Shiga toxin region displayed along the pangenome browser, showing all genomes containining the *Stx* genes. The "O157 H7 Sakai Lifted-over Pathogenic Genes" track shows the lifted-over pathogenic gene annotations of strain O157 H7 Sakai, corresponding to the *Stx* subunit A on the left and subunit B on the right. There are two major groups of *Stx*, *Stx1* and *Stx2*. Different genomes contain different numbers of *Stx* genes as well as different *Stx* groups. The pangenome view allows for the presentation of these variations, showing that Sd197 and O26 H11 11368 have one copy of *Stx1*, Xuzhou21, O157 H7 Sakai, O157 H7 EDL933, O103 H2 12009 and O111 H11 128 have one copy of *Stx1* and one copy of *Stx2*, O104 H4 2009 EL2050, O104 H4 2009 EL2071 and O104 H4 2011 C3493 have one copy of *Stx2*, and O157 H7 EC4115 and O157 H7 TW14359 have two copies of*Stx2*. As there are more copies of *Stx2* than *Stx1* (12 versus 7), the pangenome, which is a consensus sequence, is more similar to *Stx2* than *Stx1*, visibly by many SNPS on the *Stx1* copies. Variations (SNPs and indels) between the two groups *Stx1* and *Stx2* and among different genomes are shown. The texts (the labels) on the screenshot were adjusted for better readability.

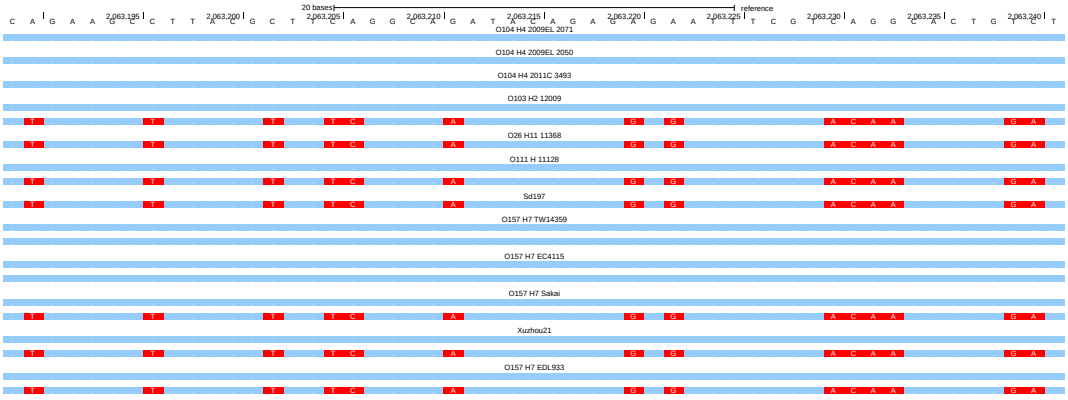

Figure B.2: A zoomed-in, base-level browser screenshoot of the Shiga toxin region.

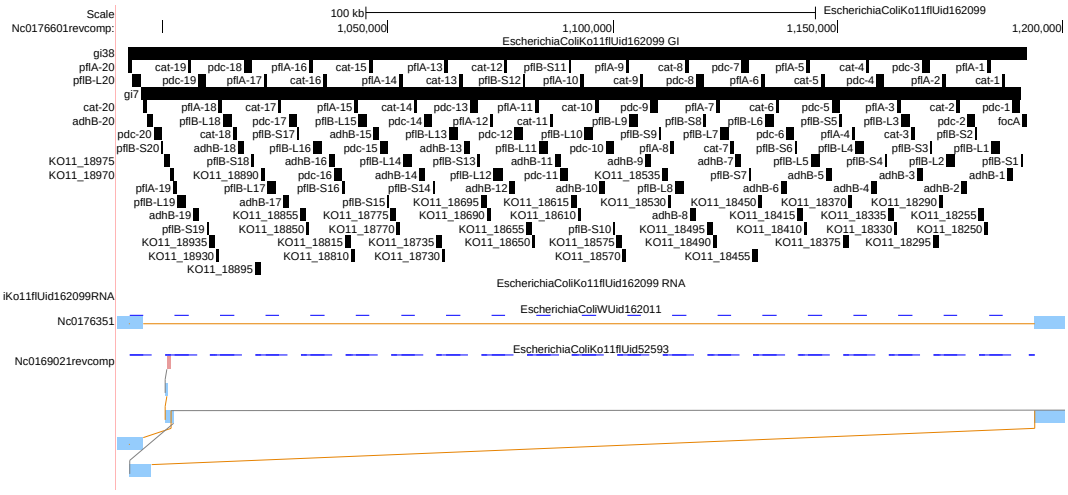

Figure B.3: A browser screenshot showing the *pdc-adhB-cat* tandem repeat region of *E. coli* KO11FL 162099. Here I show another view (with KO11FL 162099 as the reference) of the same region as in Figure 3.6 (with KO11FL 52593 as the reference).

Out of Range Genes

Genome	Sequence	Sequence Length	Gene Name	Gene Start	Gene End	Status
ETEC H10407	NC_017722.1	66681	YP_006203743.1	66646	66897	Removed
ETEC H10407	NC_017724.1	94797	YP_006203823.1	93910	95481	Removed
O104 H4 2009EL 2050	NC_018651.1	109274	YP_006772656.1	108212	112774	Removed
O104 H4 2009EL 2050	NC_018654.1	74213	YP_006772784.1	73478	74284	Removed
O104 H4 2011C 3493	NC_018659.1	88544	YP_006781712.1	86977	90021	Removed
O104 H4 2011C 3493	NC_018660.1	1549	YP_006781806.1	1313	1789	Removed
O157 H7 Sakai	NC_002128.1	92721	NP_052607.1	92527	95223	Removed
Xuzhou21	NC_017907.1	92728	YP_006315901.1	92534	95230	Removed

Table B.4: Genes that had the annotated start and end positions lying out of range of the corresponding genome assembly.

Self-foled Genes

Genome	Sequence	Genes
042	NC_017626	YP_006096541.1, YP_006098554.1, YP_006098853.1, YP_006099007.1
APEC O1	NC_009837	YP_001481211.1
ATCC 8739	NC_010468	YP_001723276.1, YP_001723852.1, YP_001723974.1, YP_001724556.1, YP_001724597.1, YP_001724792.1, YP_001725165.1, YP_001725413.1, YP_001725669.1, YP_001725826.1, YP_001725837.1, YP_001726120.1, YP_001726121.1, YP_001726135.1, YP_001726293.1, YP_001726572.1, YP_001726810.1, YP_001726822.1
BL21 DE3 161949	NC_012892	YP_006094151.1, YP_006094155.1, YP_006094156.1, YP_006094162.1, YP_006094165.1, YP_006094170.1, YP_006094174.1, YP_006094175.1, YP_006094176.1, YP_006094180.1, YP_002998455.2, YP_006094196.1, YP_006094200.1, YP_006094206.1, YP_006094214.1, YP_006094218.1, YP_006094226.1, YP_006094239.1, YP_006094241.1, YP_006094247.1, YP_006094249.1, YP_006094256.1, YP_006094259.1, YP_006094260.1, YP_006094264.1, YP_006094266.1, YP_006094269.1, YP_006094273.1, YP_006094276.1, YP_003000422.2, YP_006094277.1, YP_006094286.1, YP_006094298.1, YP_006094307.1, YP_003001107.2, YP_006094309.1, YP_006094314.1, YP_003001318.2, YP_006094328.1, YP_006094329.1, YP_006094330.1, YP_006094331.1, YP_006094333.1
ETEC H10407	NC_017722	YP_006203795.1, YP_006203796.1, YP_006203807.1
O157 H7 EDL933	NC_002655	NP_286815.1, NP_286819.1, NP_287133.1, NP_287666.1, NP_287879.1, NP_288941.1, NP_290110.1, NP_290960.1

Table B.5: Coding genes with multiple exons that overlapped with each other.

Genes Present in the Core Genome

MinCoverage	Total	Core	% Core/Total
100%	4872	1879	38.57
99%	4872	2253	46.24
98%	4872	2348	48.19
95%	4872	2433	49.94
90%	4872	2507	51.46

Table B.6: The average number of genes of each genome that are present in the core genome. "MinCoverage": minimum proportion of a gene that overlaps with the core genome to be counted. "Total": average number of total genes of an *E. coliShigella* genome. "Core": average number of genes of an *E. coliShigella* genome that are present in the core genome (core genes). "%CoreTotal": average percentage of total genes that are core genes. Note that reported numbers are average number of 66 genomes.

Appendix C

Supplement for: "Comprehensive assessment of T-cell receptor repertoires"

C.1 Clonal expansions of healthy samples in published high-throughput TCR sequencing studies

To confirm our results of the literature search for the sample clonal expansions (Section 4.3.6.1), we repeated the search for clonal expansions of healthy samples of previously published high-throughput TCR sequencing studies. The healthy samples included in the analyses consisted of three samples from Warren *et al.* [Warren et al., 2011], one sample from Wang *et al.* [Wang et al., 2010], one sample from Robins *et al.* [Robins et al., 2012], five samples from unpublished data provided

by Adaptive Biotechnologies.

The Warren *et al.* data were obtained from `ftp://ftp.bcgsc.ca/supplementary/TCRb2010`. The samples included were *male1_blooddraw1*, *male2* and *female*. The Wang *et al.* data were downloaded from NCBI SRA archive `ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/litesra/SRP/SRP001/SRP001441`. The Robins *et al.* and the Adaptive Biotechnologies data were provided by Adaptive Biotechnologies.

C.2 Identification of ERAP1 risk allele status

The sample ERAP1 alleles were extracted from the exome sequencing results of the samples. The two ERAP1 AS-associated SNPs were rs30187 and rs10050860 [Evans et al., 2011b]. The exome sequencing methods were as followed: DNA was extracted from approximately $4x10^6$ CD8 depleted PBMCs using a Purelink Genome DNA Extraction Kit (Invitrogen Kit# K1820-01). 5 ug of genomic DNA was submitted to Otogenetics (Atlanta, Georgia, USA) for exome sequencing with Agilent Sure Select V4. Resulting data was submitted to DNANexus (Mountain View, California, USA) for SNV calling.

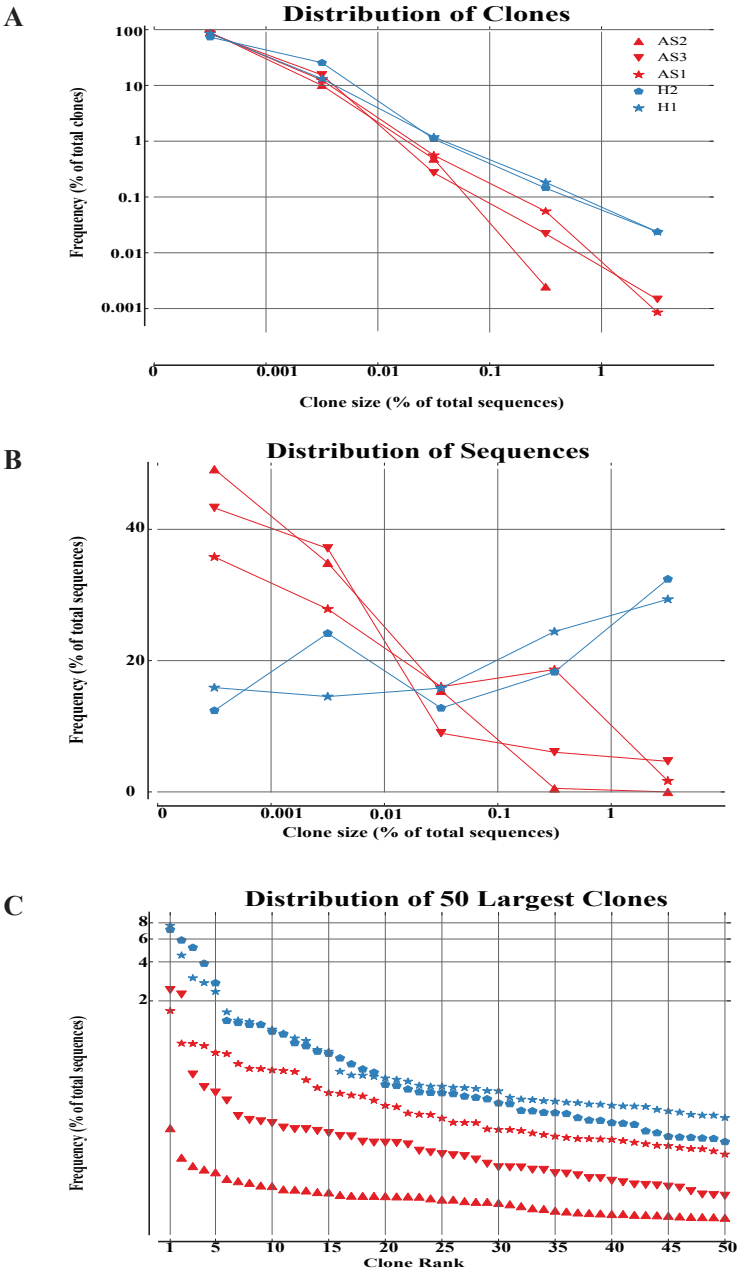

Figure C.1: Clone size distributions of one million sequence samplings. (A) Distribution of clones. (B) Distribution of sequences. (C) Distribution of sample 50 largest clones.

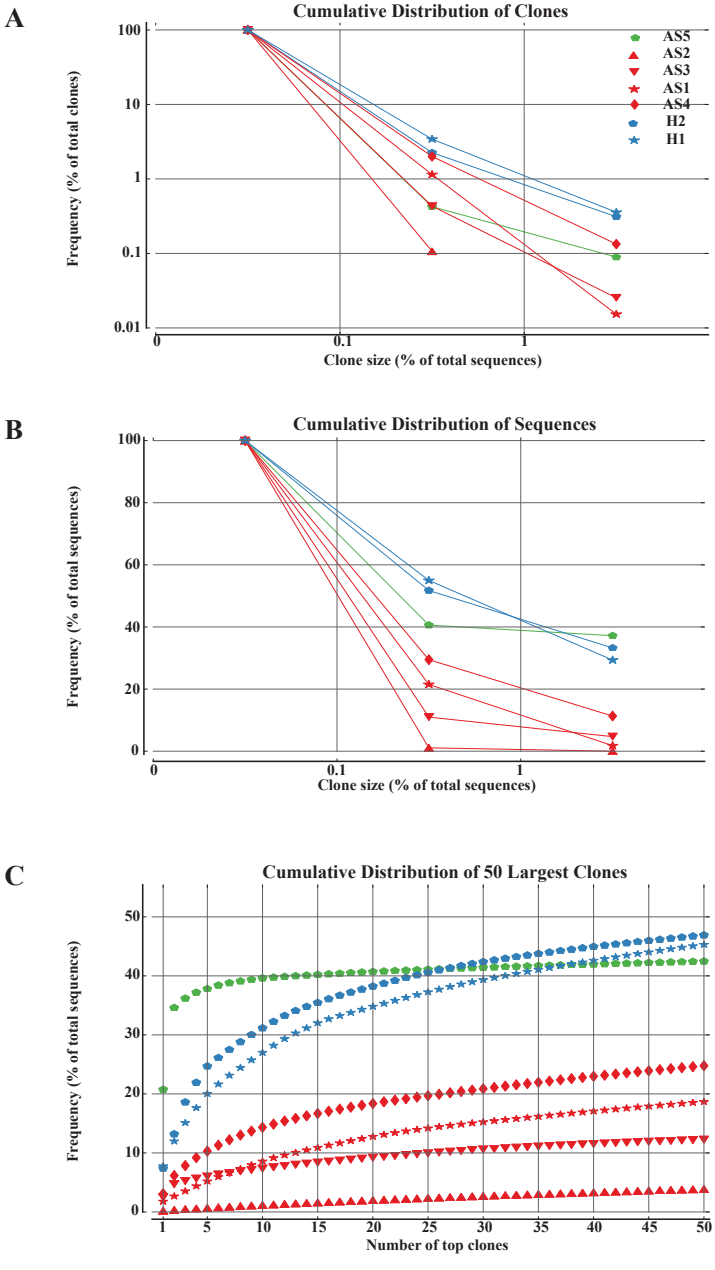

Figure C.2: Cumulative clone size distributions of ten thousand sequence samplings. (A) Cumulative distribution of clones. (B) Cumulative distribution of sequences. (C) Cumulative distribution of the sample 50 largest clones.

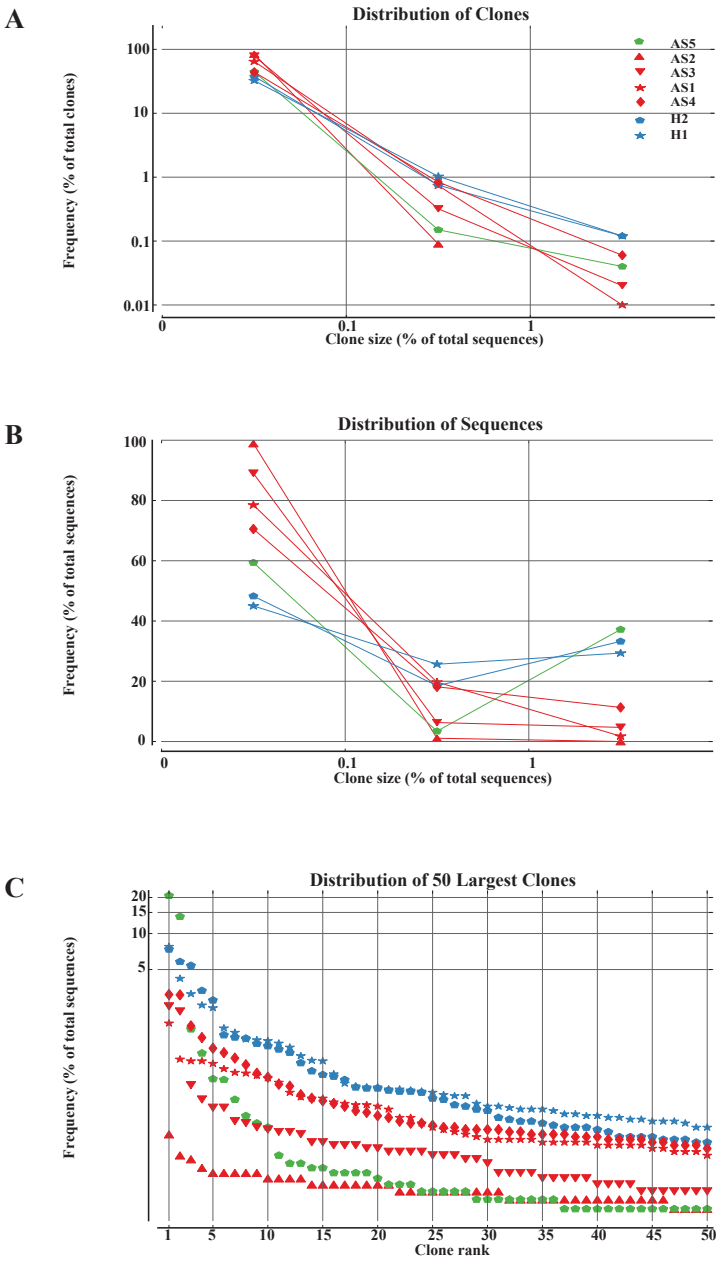

Figure C.3: Clone size distributions of ten thousand sequence samplings. (A) Distribution of clones. (B) Distribution of sequences. (C) Distribution of the sample 50 largest clones.

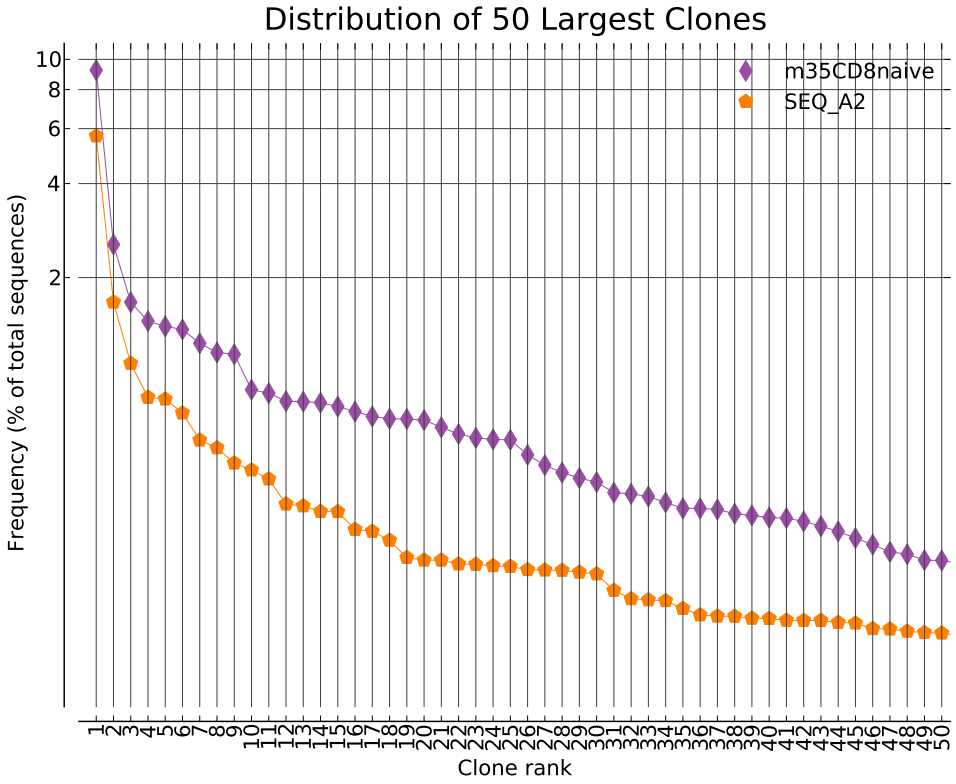

Figure C.4: Frequencies of the top 50 largest clones of healthy TRBV (DNA) repertoires sequenced by Adaptive Biotechnologies. 'SEQ_A2' was from [Robins et al., 2012] and 'm35CD8naive' was from unpublished Adaptive Biotechnologies data.

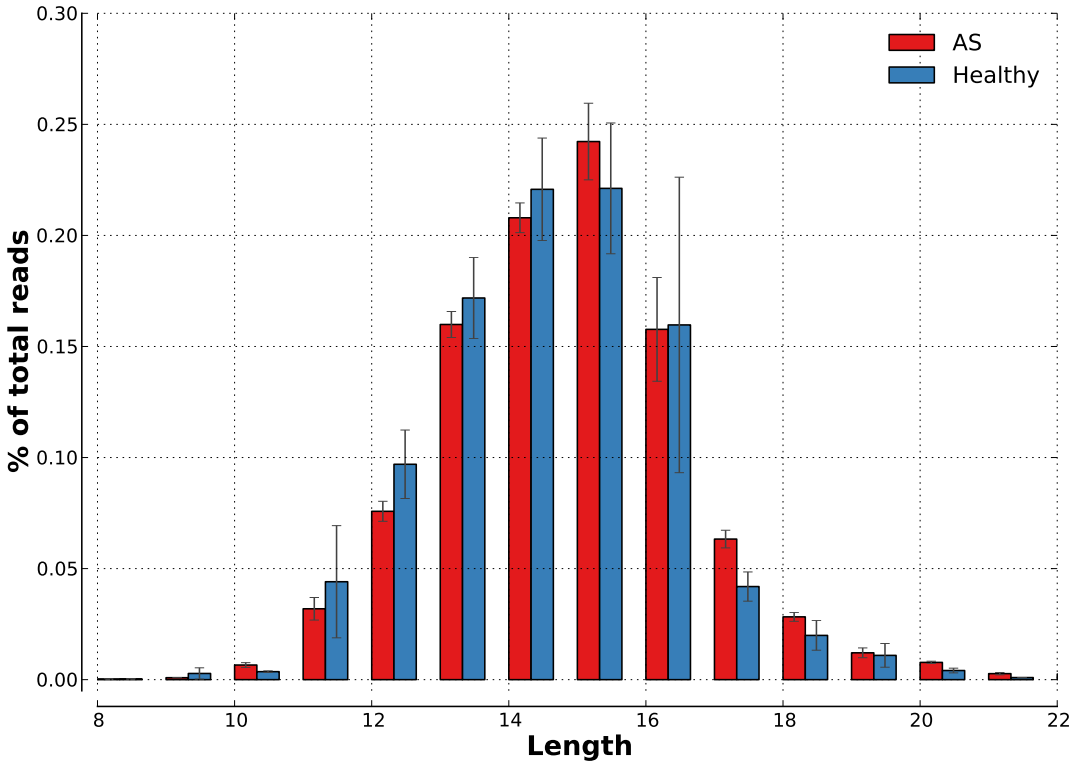

Figure C.5: CDR3 length distributions of total sequences.

Sample	Unique Clones	Simpson	Shannon	Fisher Alpha
AS1	6,544 ± 41	0.999 ± 0.000	8.174 ± 0.018	8,230 ± 168
AS2	8,446 ± 35	1.000 ± 0.000	8.939 ± 0.008	25,722 ± 740
AS3	7,900 ± 39	0.999 ± 0.000	8.582 ± 0.017	17,403 ± 441
AS4	4,531 ± 38	0.997 ± 0.000	7.563 ± 0.020	3,196 ± 58
AS5	4,500 ± 37	0.939 ± 0.002	6.169 ± 0.036	3,148 ± 55
H1	3,413 ± 42	0.988 ± 0.000	6.431 ± 0.026	1,828 ± 41
H2	3,815 ± 45	0.985 ± 0.000	6.415 ± 0.029	2,253 ± 51

Table C.1: Sample diversity indices of ten thousand sequence samplings. Rows and columns are similar to Supplemental Table 4.2

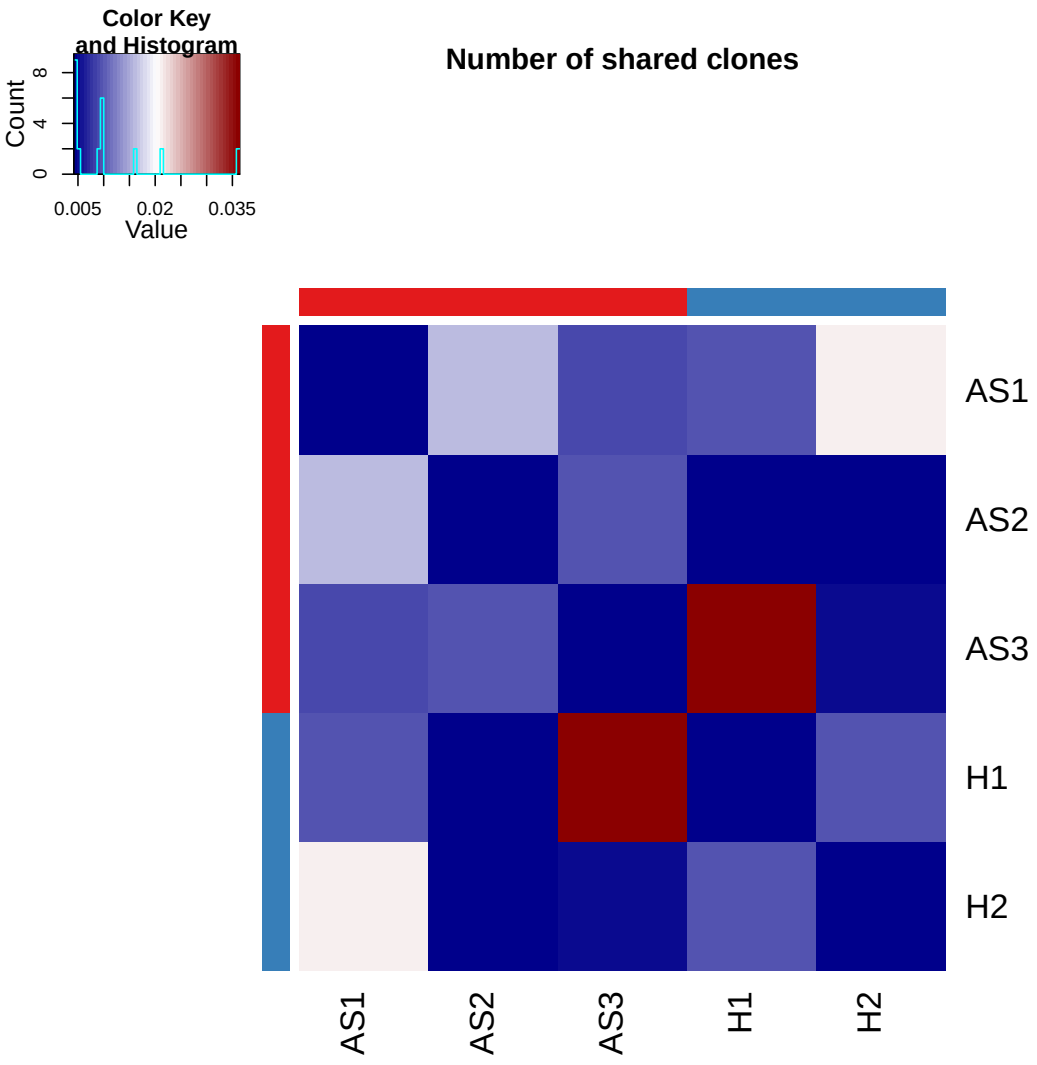

Figure C.6: Chao similarity index.

Sample	D		J		V		D-J		V-J		V-D-J	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Total	2	100.00	13	100.00	59	100.00	26	100.00	767	100.00	1534	100.00
AS1	2	100.00	13	100.00	52	88.14	26	100.00	709	92.44	1367	89.11
AS2	2	100.00	13	100.00	52	88.14	26	100.00	716	93.35	1401	91.33
AS3	2	100.00	13	100.00	51	86.44	26	100.00	709	92.44	1372	89.44
H1	2	100.00	13	100.00	51	86.44	26	100.00	699	91.13	1328	86.57
H2	2	100.00	13	100.00	51	86.44	26	100.00	701	91.40	1308	85.27

Table C.2: Summary of the sample total TRBV, TRBJ, TRBD genes and their recombinations of one million sequence samplings. 'Total': the numbers of human germ line TRBV, TRBD, TRBJ genes [Lefranc et al., 2009] and the numbers of all possible D-J, V-J and V-D-J recombinations. 'Count': number of genes or recombinations observed in the samples. '%': 'Count'/'Total'.

Sample	D		J		V		D-J		V-J		V-D-J	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Total	2	100.00	13	100.00	59	100.00	26	100.00	767	100.00	1534	100.00
AS1	2	100.00	13	100.00	52	88.14	26	100.00	700	91.26	1311	85.46
AS2	2	100.00	13	100.00	52	88.14	26	100.00	710	92.57	1364	88.92
AS3	2	100.00	13	100.00	51	86.44	26	100.00	699	91.13	1326	86.44
AS4	2	100.00	13	100.00	52	88.14	26	100.00	587	76.53	1019	66.43
AS5	2	100.00	13	100.00	52	88.14	26	100.00	631	82.27	1149	74.90
H1	2	100.00	13	100.00	51	86.44	26	100.00	674	87.87	1240	80.83
H2	2	100.00	13	100.00	51	86.44	26	100.00	673	87.74	1227	79.99

Table C.3: Summary of the sample total TRBV, TRBJ, TRBD genes and their recombinations of ten thousand sequence samplings. Rows and columns are similar to Supplemental Table C.2.

Sample	D		J		V		D-J		V-J		V-D-J	
	Count	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Total	2	100.00	13	100.00	59	100.00	26	100.00	767	100.00	1534	100.00
AS1	2	100.00	13	100.00	52	88.14	26	100.00	709	92.44	1367	89.11
AS2	2	100.00	13	100.00	52	88.14	26	100.00	716	93.35	1401	91.33
AS3	2	100.00	13	100.00	51	86.44	26	100.00	709	92.44	1372	89.44
AS4	2	100.00	13	100.00	52	88.14	26	100.00	590	76.92	1023	66.69
AS5	2	100.00	13	100.00	52	88.14	26	100.00	641	83.57	1176	76.66
H1	2	100.00	13	100.00	51	86.44	26	100.00	699	91.13	1328	86.57
H2	2	100.00	13	100.00	51	86.44	26	100.00	701	91.40	1308	85.27

Table C.4: Summary of the sample total TRBV, TRBJ, TRBD genes and their recombinations, no sampling. Rows and columns are similar to Supplemental Table C.2.