Expanding the horizons of next generation sequencing with RUFUS

Author: Andrew R. Farrell

Boston College

The Graduate School of Arts and Sciences

Department of Biology

EXPANDING THE HORIZONS OF NEXT GENERATION SEQUENCING WITH RUFUS

A dissertation

by

JOHN ANDREW REISER FARRELL

submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

July 15 2014

Abstract

Expanding the horizons of next generation sequencing with RUFUS

J. Andrew R. Farrell

Dissertation advisor: Gabor T. Marth

To help improve the analysis of forward genetic screens, we have developed an efficient and automated pipeline for mutational profiling using our reference guided tools including MOSAIK and FREEBAYES. Studies using next generation sequencing technologies currently employ either reference guided alignment or *de novo* assembly to analyze the massive amount of short read data produced by second generation sequencing technologies; the far more common approach being reference guided alignment due to the massive computational and sequencing costs associated with *de novo* assembly. The success of reference guided alignment is dependent on three factors; the accuracy of the reference, the ability of the mapper to correctly place a read, and the degree to which a variant allele differs from the reference. Reference assemblies are not perfect and none are entirely complete. Moreover, read mappers can only map reads in genomic locations that are unique enough to confidently place reads; paralogous sections, such as related gene families, cannot be characterized and are often ignored. Further, variant alleles that drastically alter the subject's DNA, such as insertions or deletions (INDELs), will not map to the reference and are either entirely missed or require further downstream analysis to characterize. Most importantly, reference guided methods are restricted to organisms for which such reference genomes have been assembled. The current

alternative, *de novo* assembly of a genome, is prohibitively expensive for most labs requiring deep read coverage from numerous different library preparations as well as massive computing power.

To address the shortcomings of current methods, while eliminating the costs intrinsic to *de novo* sequence assembly, we developed RUFUS, a novel, completely reference-independent variant discovery tool. RUFUS directly compares raw sequence data from two or more samples and identifies groups of reads unique to one or the other sample. RUFUS has at least the same variant detection sensitivity as mapping methods, with greatly increased specificity for SNPs and INDEL variation events. RUFUS is also capable of extremely sensitive copy number detection, without any restriction on event length. By modeling the underlying k-mer distribution, RUFUS produces a specific copy number spectrum for each individual sample. Applying a Bayesian detection method to detect changes in k-mer content between two samples, RUFUS produces copy number calls that are equally as sensitive as traditional copy number detection methods with far fewer false positives. Our data suggest that RUFUS' reference-free approach to variant discovery is able to substantially improve upon existing variant detection methods: reducing reference biases, reducing false positive variants, and detecting copy number variants with excellent sensitivity and specificity.

Acknowledgements

I'd like to thank Marc Jan Gubbels and his entire lab for performing the laboratory work in this project; particularly Bradley Coleman, Keith Eidell, and Brian Benenati.

I'd like to thank everyone in the Marth lab, particularly Alistair Ward, as well Erik Garrison for help with FREEBAYES and for running experiments with Velvet.

A special thanks to Gabor Marth for allowing me to work in his lab and being my mentor. Thank you for the guidance and giving me the freedom to pursue my own ideas, no matter how crazy they were.

Finally, and most importantly, I would like to especially thank my parents for all of the support over the last 5 years, and in my entire scientific carrier.

Chapter 1:
Introduction and motivation

In 1908 Thomas H. Morgan published a paper challenging the theory that genetic information, or "factors", were passed from parent to offspring through germ cells. He believed that Mendelian inheritance theories were far too simplistic, were not based in fact, and invented to merely explained the data being observed[1]. To prove his ideas he set out working with *Drosophila melanogaster* and instead definitively confirmed Mendel's ideas, defined genetic linkage, and set the stage for modern-day genetic research. He chose *Drosophila melanogaster* for its extremely short generation time, and its numerous visible traits that could be easily screened in the lab for phenotypic differences. He used both laboratory selection experiments and various mutagens in an effort to create novel visible traits that could be easily tracked in the lab in order to study their inheritance patterns[2].

Historically, the most notable phenotype identified was the white eye trait, initially identified in a male fly. Crossing the male mutant with a wild type, red-eyed, female produced offspring will all red eyes. Subsequently, inbreeding of this generation (f1) produced a 3:1 ratio of red to white-eyed flies in the next generation (f2) exactly as expected based on Mendelian genetics. However, there were no white-eyed female flies; all white-eyed individuals were male. Subsequent crossing experiments showed that the white eye trait was not lethal to female flies, but instead was inherited in conjunction with sex determination (Figure 1). This suggested that the factors that determined sex and eye color were in some way

inherited together on the same molecule, supporting the chromosomal inheritance theory[3] .

Further work by Morgan on other sex linked factors showed that certain factors appeared to always be inherited together, thus linked. Other factors showed incomplete linkage; frequently occurring together but would occasionally separate. This led Morgan to theorize that genes were physically connected on chromosomes like "beads on a string", and further it had been observed that the chromosomes wrap around each other during division. He theorized that when pulled apart, these strings may break and rejoin the sister chromosome, thus allowing cross-over during sexual reproduction producing new combinations of traits. This suggested that the physical distance between genes would thus determine the probability, and thus the rate, that a crossover would occur[4]. Morgan's student, Alfred H. Sturtevant, used this information to create the first ever genetic map, showing the physical arrangement of traits on a chromosome and defining distances between the genes in map units (later renamed centimorgan) [5], shown in Figure 2.

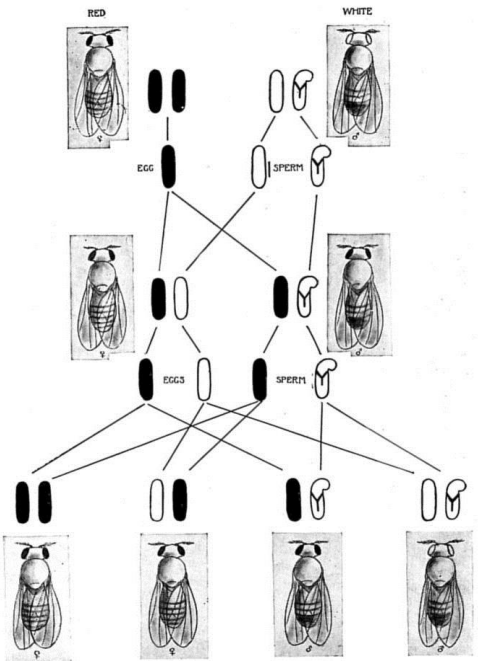


Figure 1: Depicts a cross between a white-eyed male and a red-eyed female of *D. melanogaster*. The sex chromosomes are indicated by the rods. A black rod indicates that the chromosome carries the factor for red; the open chromosome carries the factor for white eye color. Reprinted from Morgan, T. H. 1919[50]

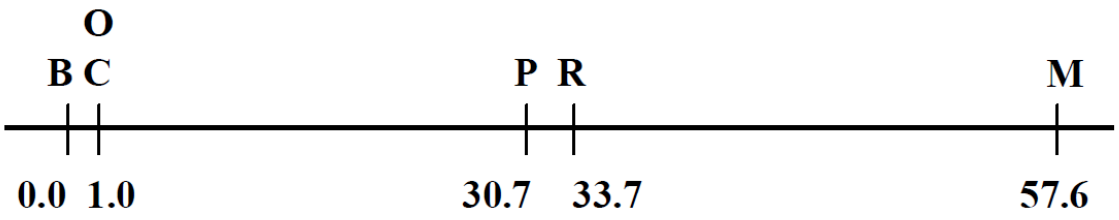


Figure 2: Sturtevant's original linkage map showing the relative ordering of 6 traits based on linkage frequency. Sturtevant defined the traits as follows: (*B*) black body color factor (black/yellow), *(C)* eye color factor (red/white*), (O)* second eye color variant (red/eosin*), (P)* third eye color (red/vermilion*). (R) and (M)* wing development producing either miniature or rudimentary wings. Reprinted from Sturtevant, A. H. 1913 [5].

Morgan's work not only proved the chromosomal theory of inheritance, it also laid the groundwork for what is now known as the classical, or forward genetic method. Simply stated, Forward Genetics links a trait (phenotype) to a specific DNA sequence (genotype). In order to link specific traits to a particular genotype, mutations are introduced into the genome of interest using various methods such as X-rays, chemical mutagens, or laboratory evolution. Individuals are then screened for a desired phenotype and mutants are identified and isolated. A variety of methods are then used to identify the exact mutation that caused the given phenotype, along with its location. One of the most direct methods, still used today, is to perform crosses to measure the given phenotype's linkage with known genetic markers, exactly as Sturtevant did. The difficulty with forward genetics is that the mutations are generated at random throughout the genome and thus identifying the mutations can be costly and extremely time consuming.

The more modern approach, reverse genetics, takes the opposite approach. Methods specifically developed for a given model organism are used to directly introduce targeted DNA mutations into an organism. The resulting phenotype is

then observed. These methods have the advantage in that the location of the mutation is known. However, it can be difficult to specifically target a given phenotype of interest. These methods are also often limited to model organisms for which specific genetic methods have previously been developed. The advantage of forward genetics is that it allows the discovery of genes that directly contribute to a specific trait of interest with little or no prior knowledge required. This makes forward genetics particularly attractive when studying non-model organisms, with unique traits that are driven by completely unknown genes. As our knowledge of basic cell processes has grown, it has become increasingly attractive to widen the scope of research and directly study more biologically relevant non-model organisms that have a direct human health impact.

For this work, we have focused on the parasite *Toxoplasma gondii* in collaboration with the Gubbels lab. Of particular interest is the invasion/egress phenotype which has no known homology to any other studied organism and is crucial to its virulence in humans.

1.2 *Toxoplasma gondii*

Toxoplasma gondii is a member of the protozoan phylum Apicomplexa, which includes numerous important human pathogens such as *Plasmodium* spp. (the causative agent of malaria) and *Cryptosporidium* spp. (severe enteritis). *Toxoplasma gondii* can infect and undergo asexual reproduction in any warm-blooded vertebrate, while only being able to undergo sexual reproduction in cats; its definitive host. Infection of intermediate hosts consists of a short acute phase

followed by a life-long maintained chronic phase[6,7,8]. *Toxoplasma* is wide-spread in the U.S. with a current prevalence of 22.5% [9]. Its high prevalence, with relatively low public knowledge, is due largely to the fact that most acute infections pass with mild or no symptoms and in healthy individuals the chronic phase shows no symptoms. However, severe disease can occur in immune-compromised patients[10-13]. Toxoplasmosis can also occur if a pregnant mother passes on an infection to the fetus. This may result is progressive vision loss or serious developmental damage to

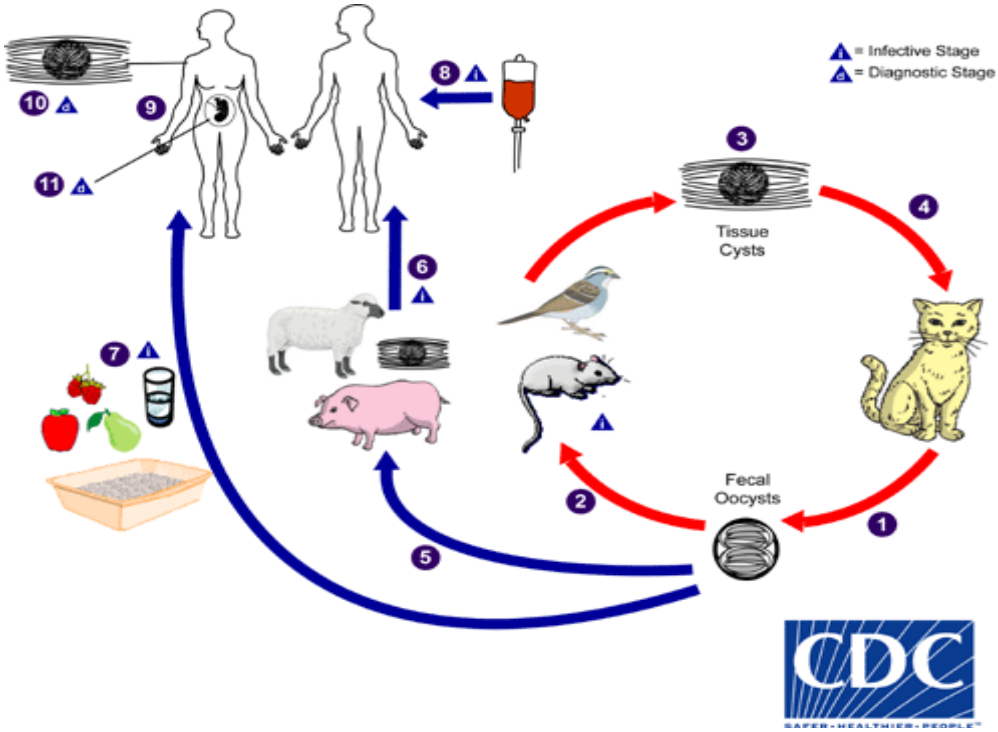


Figure 3: **Life Cycle of *Toxoplasma gondii***. The only known definitive hosts for *Toxoplasma gondii* are members of Felidae family such as the domestic cat. Oocysts are shed in the cat's feces (1). Intermediate hosts (including birds, rodents, as well as any other warm-blooded animal) become infected after ingesting contaminated soil, water or plant material (2). Oocysts transform into tachyzoites shortly after ingestion and localize in neural and muscle tissue and develop into cysts (3). Cats become infected after consuming intermediate hosts (4). Humans can become infected by direct contact with fecal oocysts (7), eating undercooked meat (6), organ transplantation or blood transfusion (8) or though transplacentally from mother to fetus (9). Reprinted from CDC.org

the unborn child[12]. In the U.S., one in 5,000 pregnancies shows complications due to *T. gondii* infections. In Europe, these numbers are slightly higher[7,8].

Upon injection by a host, *T. gondii* parasites differentiate into tachyzoites. Tachyzoites replicate rapidly causing the acute state of infection resulting in flu-like symptoms, along with tissue death. The tachyzoites invade a host cell by creating a vacuole containing a parasite Tachyzoites replicate quickly inside this vacuole until the cell is depleted. They then lyse the cell and escape, a process known as egress. In a healthy individual, the immune system will suppress the tachyzoites, stimulating them to differentiate into bradyzoites. Bradyzoites create tissue cysts, hiding from the host's immune system within the host cells, where they divide slowly, characterizing the chronic phase of the infection[14]. During this phase, *Toxoplasma* has little effect on the host and has generally been considered benign. However, recent research suggests that long term infection may be linked to schizophrenia and possibly other psychological disorders[15,16]. To date, pharmaceuticals developed to treat *Toxoplasma* strictly target the acute stage of infection[9,17]. There are currently no drugs targeting the chronic phases of the infection. Although generally considered benign, the correlations drawn between chronic *Toxoplasma* infection and severe psychological disorders demand a better understanding of the mechanisms involved in infection, such as host cell invasion. A more in depth understanding of *Toxoplasma's* life cycle could contribute to the development of novel wide-acting drugs.

1.3 Creation of the F-P2 mutant

Central to *T. gondii* infection and life cycle is its ability to invade a host cell, multiply, and egress from the host cell causing cell death (Figure 4). In order to identify genes specifically involved in this process, Dr. Gubbels created a temperature sensitive screen to detect mutants that showed limited growth due to mutations in various stages of the cell cycle[18]. The F-P2 mutant was identified using this screen in 2001. F-P2 lacks the ability to escape from a host cell (egress) at the restrictive temperature 40°C but is fully capable at 38°C[19]. Dr. Gubbels has previously shown that mutations identified in this screen can be genetically

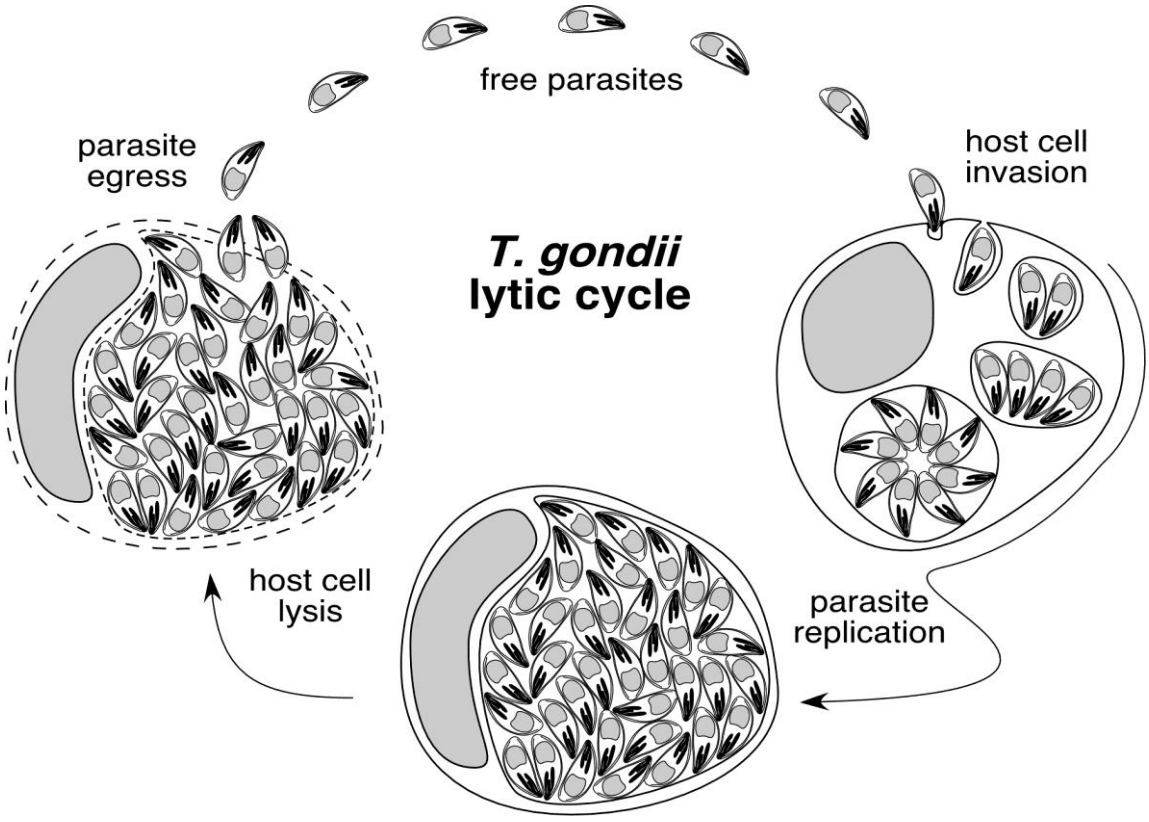


Figure 4: *T. gondii* lytic cycle. During infection, free tachyzoites invade host cells creating a vacuole. The parasite divides using the cells resources until the cell is depilated. The parasites then egress and the freed parasites begin the cycle again. Reprinted with permission from Dr. Marc Jan Gubbels

complemented and identified by transfecting using a wild-type cosmid library[18]. However, after 30 independent transfection experiments the F-P2 egress deficient phenotype could not be rescued, indicating that the mutation may be dominant. To test this, Dr. Gubbels constructed a 130-fold genomic coverage cosmid library from the F-P2 mutant. The library was successfully verified by transfecting it into a strain with a known deletion of the HXGPRT gene, with 4 out of 5 transfections successful. However, after ten transfections of the F-P2 cosmid library into the parent line the F-P2 phenotype could not be replicated.

The inability to complement this phenotype in the years since its identification led Dr. Gubbels to pursue mutational profiling in collaboration with the Marth lab. Mutational profiling is enabled by second generation whole genome sequencing technologies. The goal of the method is to completely sequence the genome of both the mutant and the parent, creating a complete catalog of all differences between the two samples. This has numerous benefits over traditional forward genetic methods of identifying causative mutations. Second generation sequencing is extremely fast, and it can be completed in under one week for almost any organism. This allows detection of mutations far faster than performing labor intensive back crossing and linkage analysis experiments. Further, sequencing can be performed on any organism, enabling forward genetic screens on a wide array of organisms that may not be practical for extended work in the laboratory. Additionally, by identifying specific regions of interest in the genome, researchers can focus work on specific sequences, which may have been missed due to possible biases in other methods.

The Marth Lab has extensive experience in second generation sequencing, with a focus on tool development for sequence analysis and alignment, variant calling, visualization, and sequence interpretation. We are involved in numerous projects that use and challenge our methods such as the 1000 Genomes and TCGA cancer genome atlas project. Using the F-P2 mutant, we have successfully used our tools and experience to develop a mutational profiling pipeline specifically designed to identify novel mutations in forward genetic studies.

Chapter 2:

Mutational Profiling

2.1 Introduction

Mutational profiling is used to identify all mutations that exist between two samples in an effort to identify possible causative mutation or mutations of a given phenotype. Mutational profiling is enabled by second generation whole genome sequencing technologies, which facilitates complete sequencing of both mutant and parent genomes. The resulting sequence data are then aligned to a reference, variants are called, and sequence differences between the two samples are identified. There are numerous challenges that must be overcome to accomplish this successfully. The variant calls must be of extremely high quality and confidence, in order to minimize time wasted by researchers pursuing false positives. Conversely, the calls must be extremely sensitive to ensure that no variants are missed that may be biologically relevant. We have used our expertise with computational methods to develop a complete pipeline for mutational profiling.

Here we focus on the initial development of our mutational profiling pipeline, in collaboration with the Gubbels lab at Boston College. This pipeline was developed while analyzing whole genome sequence data from the *Toxoplasma*

gondii mutant DOC2 (also referred to as F-P2), published in Farrell et al Science 2013 [19]. This method has since been used on over 20 laboratory samples including human, *Toxoplasma gondii*, and *Plasmodium berghei*.

2.1.1 Second generation sequencing

All next generation sequencing methods (more recently termed second generation) are characterized by massively parallel, short-read, shotgun sequencing, which enables relatively inexpensive sequencing of an organism's entire genome in a matter of days. The Illumina technology has emerged as the dominant second-generation sequencing technology and was utilized for this research. The current Illumina iteration is the HiSeq 2500, which offers up to 1 terabase of sequence in as little as 6 days using a read length of 125 bp[20]. Recently, read length has been extended out to 250 bp, further increasing the throughput.

The workflow of the Illumina has been designed to be as automated as possible, reducing both laboratory costs and human error. The process is comprised of 3 steps; library prep, cluster generation, and sequencing (Figure 5). Illumina offers numerous library prep kits for a wide range of applications, such as Whole genome, RNA, Chip-seq, and exome capture. Here, we will focus on whole genome sequencing; however the basic principles involved are shared between all kits. Purified genomic DNA is sheered using sonication and size selected to ensure a uniform fragment length. Illumina specific adapters are ligated to the ends of the

fragments, producing a library of labeled DNA fragments (Figure 5.1). From this point forward all work is automated. Sequencing is performed on flow cell; each cell contains 8 separate sequencing lanes. The flow cell is made from a glass microscope slide with a "lawn" of DNA oligos bound to the glass surface, which are complementary to the adapters ligated to the DNA strands. During cluster generation, the cluster station denatures the library and "flows" the DNA sample across the lawn. The DNA fragments anneal to random locations on the lawn via their ligated adapters. This specific arrangement of oligos is the heart of the Illumina sequencer, as it creates a lawn of spatially-separated, individual fragments that can then be sequenced in a single parallel run. This creates a situation analogous to parallel capillary sequencing in a 96 or 384 well plate, however with over 3 billion separate reactions. The individual fragments then undergo a step known as bridge amplification (Figure 5.2). At a specific temperature the fragments will bend and the free end will anneal to a second anchor bound to the slide, producing a "bridge". When this fragment is extended it creates a complementary strand, whose end is physically bound to the slide and

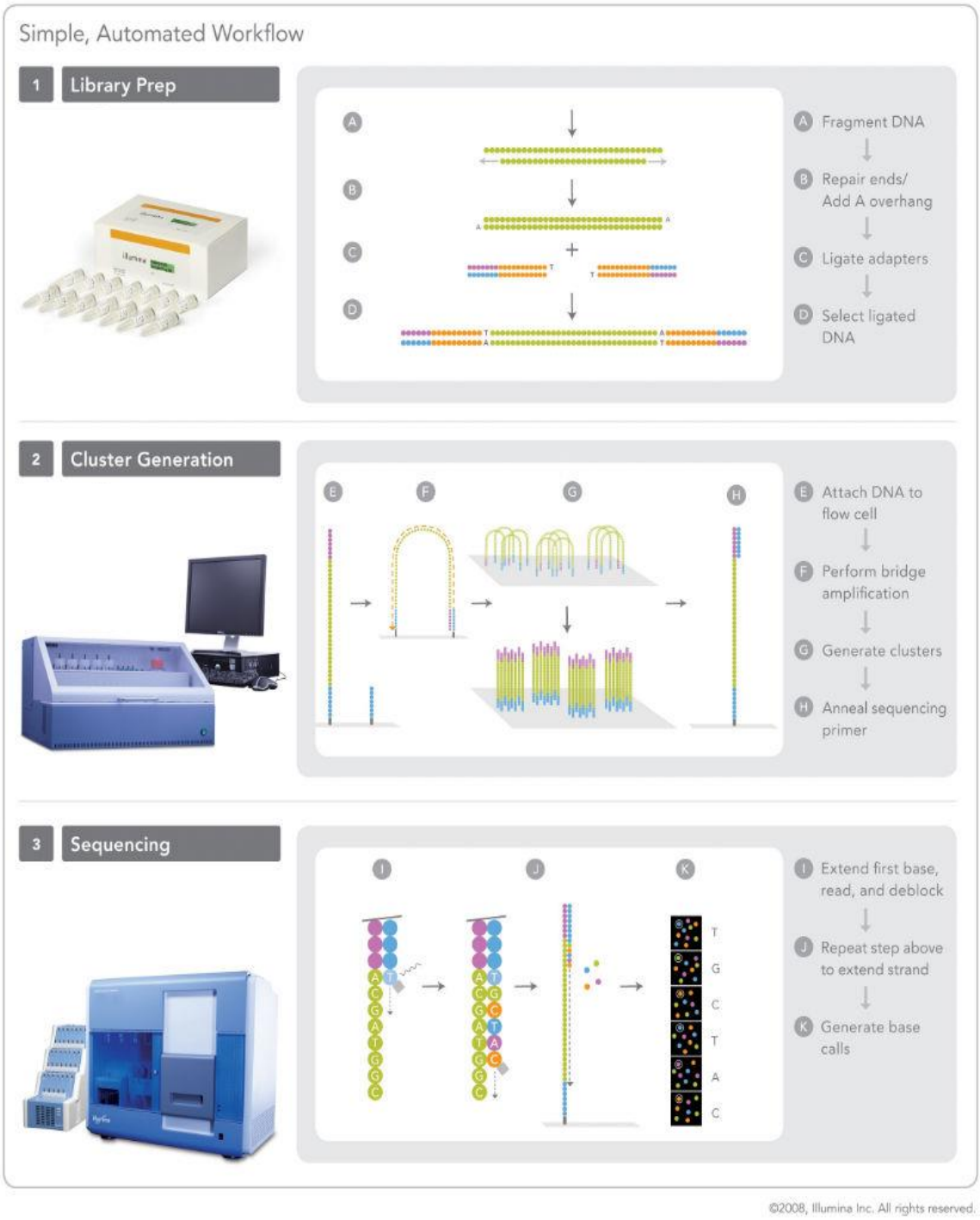


Figure 5: **The Illumina sequencing method**. (1) Libaray prep adds Illumina specific adapters to genomic DNA. (2) Cluster generation binds DNA fragments to a flow cell, and creates clustyers of clonal fragments to increase sequencing signal. (3) Sequencing is performed by extending each fragment with a fluorescently labeled reversibly terminated dNTP. Reprinted from Illumina Inc. Promotional materials 2008.

cannot be washed off. Multiple rounds of PCR amplify the fragments, creating a spot of clonal DNA fragments, thereby amplifying the signal. Following amplification, the flow cell is loaded into the sequencer. Sequencing is performed using PCR with fluorescently labeled dNTPs with a chemically blocked 3'-OH group, preventing extension (Figure 5.3). Each nucleotide, A, C, T, and G are labeled with a different fluorophore. For each round of base incorporation, all 4 nucleotides are flowed across the cell simultaneously and for each strand a single base is extended. The amplified clonal clusters produce enough fluorescent signal that each spot's newly incorporated nucleotide can be read with a high resolution camera and the new base recorded. The fluorophore and block on the 3'-OH are cleaved off and a new round of nucleotides are added to read the second base. The cycle is repeated, reading each base until a set number of cycles is reached.

If paired-end sequencing is desired, all amplified DNA fragments produced in the first round of sequencing are denatured, and the process is started again from the other end of the fragment. This can be used to both increases the total volume of sequence, as well as increase confidence in both assembly and read mapping. By sequencing both ends of a fragment, it can be assumed with high confidence that the two mates, or mate pair, exist on the same fragment. The distance between the two fragments can be estimated by the fragment size

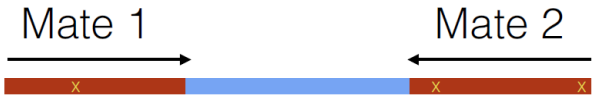


Figure 6: Mate pair sequencing. A single DNA molecule can be sequenced from both ends, labeled Mate 1 and Mate 2. Note the opposite orientations of the reads depicted by the direction of the arrows. The center of the reads sequence is unknown and is referred to as the insert sequence is depicted in blue. Image reprinted with permission from Dr. Alistair Ward

selected in library prep. If one sequences a library with a fragment size of 500bp, it can be assumed that the two 100bp sequenced mates have roughly 300bp of sequence between them. However, the method used to size select the DNA will affect the stringency of the insert size and may result in skew towards smaller or larger fragments.

The efficiency of nucleotide incorporation limits the read length of this technology, and defines the error profile. Nucleotide incorporation, like any chemical reaction, is neither 100% efficient or accurate. Each cluster on the slide, while representing a single sequence, is made up of approximately 1 million amplified fragments. Each round, a given proportion of the fragments will not incorporate a base and lag behind (phasing), while others may incorporate extra bases (pre-phasing), the greater the number of cycles in a run, the greater the number of strands that will have fallen out of sync. This will increase the

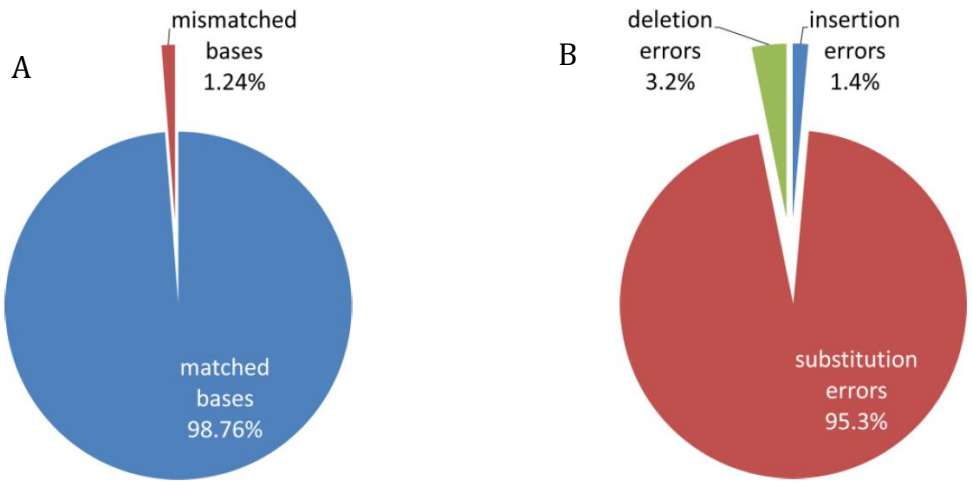


Figure 7: Error Profile of Illumina Reads. (A) Quantification of the total number of bases in an Illumina 36bp data set that did not match the reference when aligned. A negligible fraction may represent genetic variants. (B) Quantification of the error types from the mismatch set in A. Unpublished work performed by Derek Barnett (Boston College).

background noise effectively drowning out the sequencing signal. Errors tend to cluster at the end of sequence reads and tend to be in the form of substitutions, i.e. misread bases. Insertions or deletion errors are extremely rare (Figure 7). In addition to phase and pre-phasing, any PCR errors that occur early in cluster formation will propagate, resulting in base substitution errors as well.

The result of all second generation sequences is a FASTA formatted text file containing a single sequence, and base quality scores, for each cluster identified by the machine. The bioinformatician's task is to interpret these raw shotgun data and to build a complete genomic picture of the organism, accounting for all contamination and error. There are two methods currently employed to analyze whole genome shotgun data, de-novo assembly and reference guided assembly (also referred to as mapping).

2.1.2 *De novo* assembly

De novo assembly attempts to construct a complete genome using short sequence reads in the same way a puzzle is put together without knowing the complete picture beforehand. The earliest, and conceptually most straight-forward, methods for *de novo* assembly use an overlap-layout consensus algorithm. These methods were used to create the first human genomic assemblies[21,22]. Overlap based methods used the entire read and attempt to construct longer contigs by finding overlapping sequencing regions between reads and combining them to

produce longer contigs (Figure 8b). However, these methods are extremely computationally intensive as they must load the entire data set into memory and compute complete overlap scores between all read combinations. These methods were sufficient for assembling the hundreds of thousands of BAC sequences used to create the first drafts of the human genome. However, due to the fact that these methods increase massively in computational time as the numbers of reads are increased, it is physically impossible to use these earlier algorithms to assemble the billions of reads produced by a single Illumina run. [23]

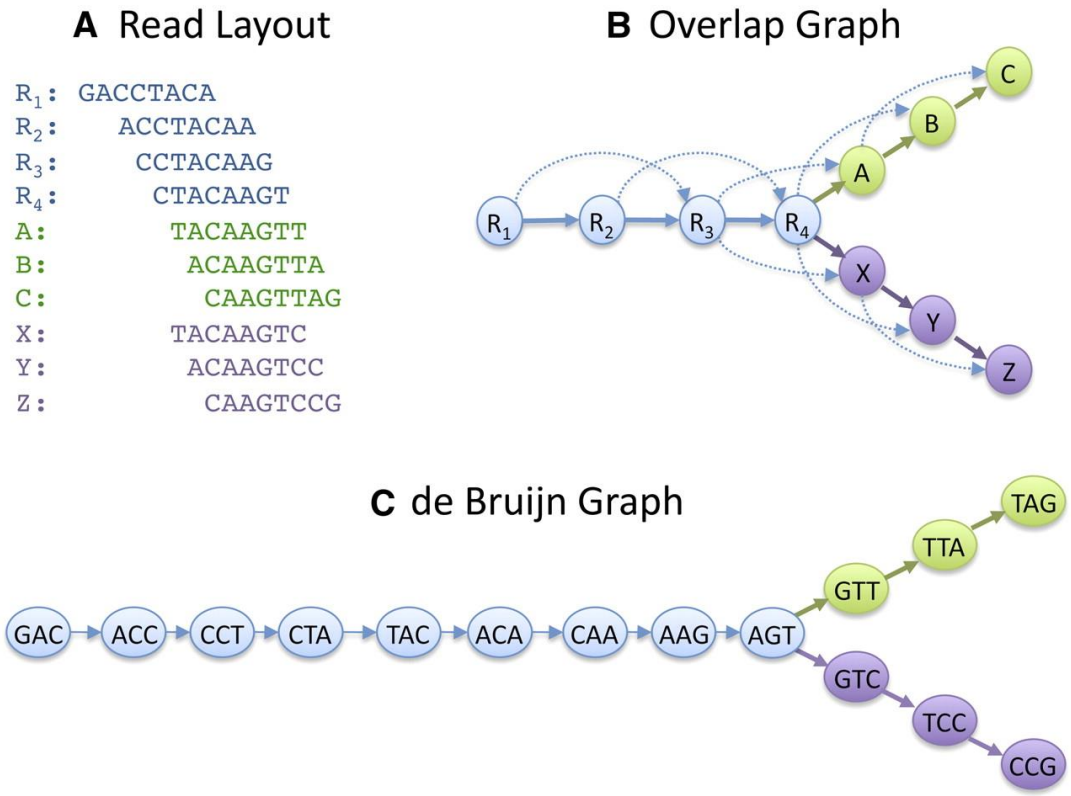


Figure 8: Assembly Algorithms. (A) 10 theoretical reads. (B) Example of an overlap assembly, each read is ordered based on its sequence overlap with the other reads. Solid arrows depict direct read overlaps, dotted arrows shows secondary read overlaps. (C) Example of de Bruihjn Graph assembly, the reads have been hashed into 3bp k-mers, and collapsed as nodes. A graph has been constructed though the nodes based on the relationship of the k-mers within the reads. Reprinted from Schatz,M.C. 2010 [23]

Assemblers designed specifically for next generation sequencing data tend to use a graph or tree structure[24,25]. These assemblers break reads into smaller sections called hashes or k-mers. Exact hashes are combined into nodes. A graph is then constructed linking the nodes based on their association in each original reads. The graph is then analyzed for the most likely path through the reads to produce the final assembly (Figure 8c). The factors that separate different graph assembly methods are usually how they handle repeats and errors.

Producing an accurate assembly for any organism is, unfortunately, not practical for most applications including mutational profiling. Assemblies are limited by the read length, the complexity of the genome, and the volume of data produced by second generation sequencers. Reads from second generation sequencers are relatively error prone which interfere with assembly; either confusing the assembler or producing spurious branches in the graph. In addition to the error rate, the coverage of second generation sequencing reads is not uniform across a genome. At best, the probability of a specific fragment of DNA annealing to the Illumina slide and producing a sequence is random, resulting in a Poisson coverage distribution. However, there are well known biases in PCR due to the fact that DNA polymerase will slow down based on certain DNA motifs and GC content, further skewing coverage across a genome[26]. To counter these two challenges, *de novo* assembly requires extremely deep coverage, often totaling over 300 fold coverage[27] in order to ensure each region of the genome has adequate coverage to both reduce the impact of random errors and to construct a complete contig through the region. Further, genomes are not comprised of completely random

sequence; they are instead enriched for highly repetitive stretches of DNA, such as pseudo-genes, transposable elements, gene families, and centromeres. The short-reads produced by second generation sequencing technologies do not provide enough information to span these regions and confidently place them in a genome. Data from paired end sequencing can be used to alleviate this problem. By using multiple sequencing libraries with varied insert sizes, large repeats can be spanned and the assembly improved. The Allpaths method outlines a specific recipe to assemble various genome sizes, though they all require up to 4 separate library preparations and coverage greater than 400 fold[27] (Table 1). In addition to the sequencing expense, assembly is massively computationally expensive. Even the most effect graph based assemblers take over 500GB of ram and days of runtime to assemble a 1GB genome[25]. Though *de novo* assembly does not require an assembled genome, they are unfortunately far too expensive for most applications. [27]

Provisional sequencing model for de novo assembly

Libraries, insert types*	Fragment size, bp	Read length, bases	Sequence coverage, ×	Required
Fragment	180†	≥100	45	Yes
Short jump	3,000	≥100 preferable	45	Yes
Long jump	6,000	≥100 preferable	5	No‡
Fosmid jump	40,000	≥26	1	No‡

*Inserts are sequenced from both ends, to provide the specified coverage.

†More generally, the inserts for the fragment libraries should be equal to ~1.8 times the sequencing read length. In this way, the reads from the two ends overlap by ~20% and can be merged to create a single longer read. The current sequencing read length is ~100 bases.

‡Long and Fosmid jumps are a recommended option to create greater continuity.

Table 1: Allpahts recipe. The sequencing recipe suggested by Allpaths to produce the best possible *de novo* assemblies [27].

2.1.3 Reference guided assembly/mapping

Reference guided assembly is a far more practical approach to whole genome analysis, more frequently termed alignment or mapping. Mapping methods are enabled by the published reference sequences of numerous genomes, using them as a guide. Like assembly, mapping methods are analogues to a puzzle, though now the original picture is known and can be used as a guide. Hundreds of aligners have been developed by different groups and almost all of them follow the same basic two steps. The aim of mapping is to take each sequence read and finds its most likely placement in the reference genome. To find the true highest scoring alignment for a read, one could perform a full Smith Waterman[28] alignment for each read against the reference genome; however this would be prohibitively slow on an

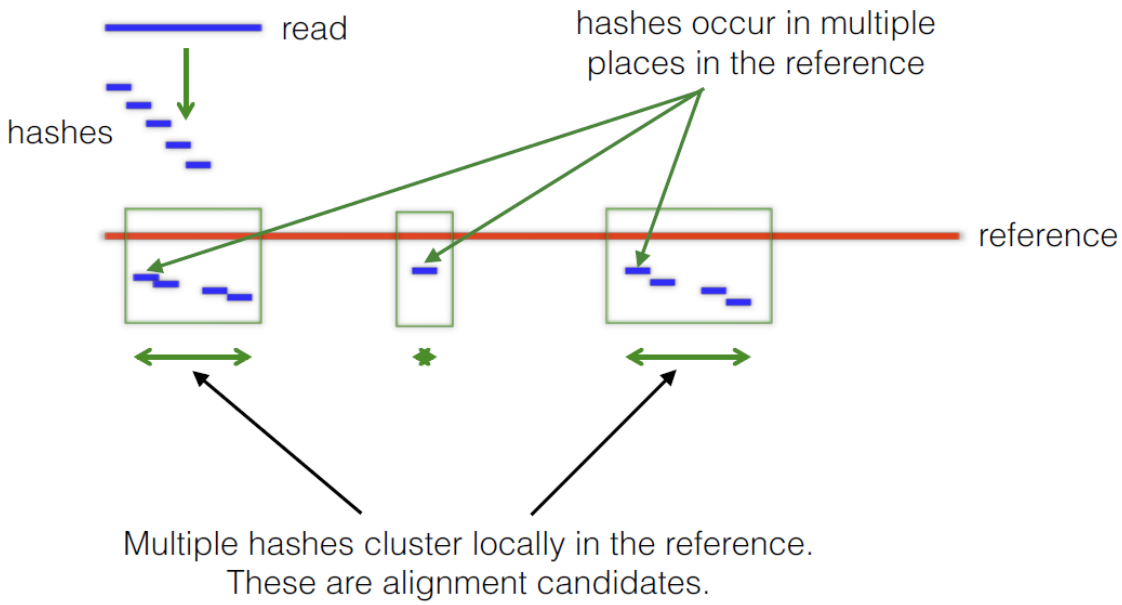


Figure 9: Reference guided alignment. Reference guided alignment (mapping) is completed in two steps. Depicted here, a read is hashed up into k-mers. Exact matches to the k-mers are identified in the genome to identify candidate locations. Image reprinted with permission from Dr. Alistair Ward

entire Illumina data set. Instead, modern aligners use a two-step process to limit the use of expensive pairwise alignment. Candidate regions in the reference are identified using an exact string match (Figure 9). Aligners differ on how they accomplish this, the two main methods being to perform a Burrows-Wheeler transform, or to create a hash Table of the reference. Both of these methods will produce the same final result. For both, a given k-mer is chosen that is shorter than the read length. For a given read, every k-mer is generated and their exact matches are identified in the genome and the location of each match is recorded. These hits are then used as a seed, to begin a more sensitive local alignment, such as a Smith-Waterman[28] or Needleman–Wunsch[29]. In this manner, sensitive alignment can be quickly focused to highly similar regions of the genome, vastly reducing the time required over a global alignment. For this work we have used the hash-based aligner MOSAIK[30], developed in our lab by Wan-Ping Lee and Michael Stromberg and extensively used in the 1000 genomes project.

Mapping has numerous advantages over *de novo* assembly. Firstly, it is far less computationally intensive; MOSAIK uses less than 15GB of ram to align an entire human dataset. For the purposes of variant detection, there is no need for multiple libraries, or hundreds of fold coverage. Confident variant detection can be achieved in regions with as little as 1 read, though 10X per chromosome is usually considered a conservative minimum to ensure completely confident genotyping[31]. These characteristics make mapping based methods a far more attractive option for most labs, and make whole genome sequencing experiments a viable option for almost any lab.

2.1.4 Variant discovery

There have been numerous software packages developed to detect variants in reference guided alignment data. For this work we have used FREEBAYES[32,33], developed by Gabor Marth and Erik Garrison. FREEBAYES uses a Bayesian algorithm that takes into account previous knowledge on the rate of mutation with a population to detect small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of the reads.

2.1.5 *Toxoplasma. gondii* genome

Toxoplasma has a haploid genome consisting of 14 chromosomes totaling 64 MB with 52.1% GC content. Ninety percent of all *Toxoplasma* isolates fall within three genotypes, Type I, II and III[34]. ToxoDB has draft-quality assembled genome sequences available for representatives of all three lines[35]. For this work we used the Type I strain GT1, whose reference has been sequenced at 8-fold capillary read coverage and has been assembled into 402 sequence scaffolds. GT1 is the closest related reference strain to the laboratory strain used for these experiments, RH. The estimated divergence between GT1 and RH is estimated to be only 0.01%[36]. It is predicted that *Toxoplasma* contains roughly one gene per 7.4 kb, totaling 7,817 genes with an average of 4 introns per gene.

2.2 Mutational profiling pipeline development

F-P2 exhibits a temperature sensitive egress-deficient phenotype with normal intracellular growth. The mutant was identified in a screen of ENU treated mutants in 2001, and subsequent experiments to complement the phenotype have proved unsuccessful (described in section 1.3 and appendix A). The analysis of this sample was used to develop the methods implemented in our final mutational profiling analysis pipeline.

2.2.1 Illumina sequencing

Purified DNA from the F-P2 strain and the parent were sent to The Broad Institute for Illumina whole genome sequencing. The samples were sequenced at the Broad by their technicians using the following protocol. Genomic DNA (3µg) was sheared to ~400 bp in size using the Covaris E210 instrument (Covaris, MA). Fragmented DNA was end-repaired with T4 DNA polymerase (NEB, MA), phosphorylated with T4 polynucleotide kinase (NEB, MA) and 3' adenylated with Klenow fragment (NEB, MA) using standard protocols. DNA fragments were ligated with Illumina paired-end adaptors according to the manufacturer's protocol (Illumina, CA). All enzymatic steps were cleaned up using Qiagen min-elute columns (Qiagen, CA). Adapter ligated fragments were purified via gel electrophoresis (4% agarose, 85 volts, 3 hours) and a single band in 500-550 bp size range was excised

resulting in a library with insert averaging 400 bp in size. DNA was extracted from the gel using Qiagen min-elute columns (Qiagen, CA). Purified library fragments were enriched via PCR amplification using Illumina paired-end PCR primers (Illumina, CA) and Phusion polymerase (NEB, MA). Enriched libraries were quantified using standard SYBR green qPCR protocols, using primers specific to the Illumina paired-end adapters. Libraries were normalized to 2nM and denatured using 0.1 N NaOH. Denatured libraries were cluster amplified on V2 flowcells using V2 chemistry according to manufacturer's protocol (Illumina, CA). Flowcells were sequenced on Genome Analyzer II's, using V3 Sequencing-by-Synthesis kits and analyzed with the Illumina's v1.3.4 pipeline following manufacturer's protocol producing paired end 75bp reads (Illumina, CA).

2.2.2 Reference guided alignment

MOSAIK was used to align the Illumina reads to a reference genome containing both the *Toxoplasma gondii* GT1 Genomic reference v5.0 and the Human genome reference build 37. Due to the fact that the parasites are cultured in human cells, including the human reference in the alignment is a necessary step to remove any possible contamination. Both of the samples did show significant human sequence. For the F-P2 sample 90.1% of the sequenced reads successfully aligned to the combined reference genome. Of those reads, 32.2% aligned to the human reference and 67.8% aligned to *Toxoplasma gondii*. The Parent sample had a total read alignment of 90.5%. Of the aligned reads, 49.4% aligned to the Human

reference and 50.6% of the reads aligned to *Toxoplasma gondii*. Despite the large proportion of human DNA in the sample, there was still adequate coverage across the *Toxoplasma gondii* genome. The average coverage was over 30X for both samples, and the alignments covered greater than 96% of the known genome in both samples at a minimum coverage of 5x.

2.2.3 Variant calling and filter development

Variants were called separately in each of the two samples using BAMBAYES, and the resulting calls were intersected to identify variant calls unique to the mutant. Using a minimum coverage threshold of 5x, BAMBAYES reported a total of 1841 SNVs in the F-P2 sample and 1728 in the Parent sample. The minimum allele coverage of 5X is lower than the standard 10X[31]. We felt using a lower coverage was an acceptable tradeoff to increase sensitivity and avoid missing any valuable possibly SNVs. Using SNVs with lower coverage will increase the chances of calling a false positive, but reduce the false negative rate. A simple comparison of the SNVs identified 1469 SNVs shared between the two samples and 292 unique to F-P2. Upon further inspection, it became apparent that the majority of the SNVs were spurious calls, and appeared in locations where the alignments appeared to be heterozygous (Figure 10). Given that *Toxoplasma gondii* is haploid in the life cycle used in this research, this is not possible. We believe this suggests that reads from multiple sections of the genome are piling up in the same area on the reference. This can be caused by either gene duplications that have arisen between RH and

GT1, though more likely, these are errors in the GT1 reference assembly where similar genes, or gene families, have been incorrectly placed on top of each other. This is a common problem in assembly algorithms. No allelic variation was seen between parent and mutant at these sites, so we elected to filter these areas out for this study. To clean up the calls we developed a filtering algorithm that excluded any variants whose major allele call comprised less than 70% the total number of bases at that position. This removes 0.008% of the genome (4,976 bp), but accounted for over 88% of the unique SNVs. After filtering we showed 997 SNVs shared between the two samples (which are likely Single Nucleotide Polymorphisms or SNPs) and 33 SNVs unique to F-P2, which are all candidates as the causative mutation resulting in the F-P2 phenotype.

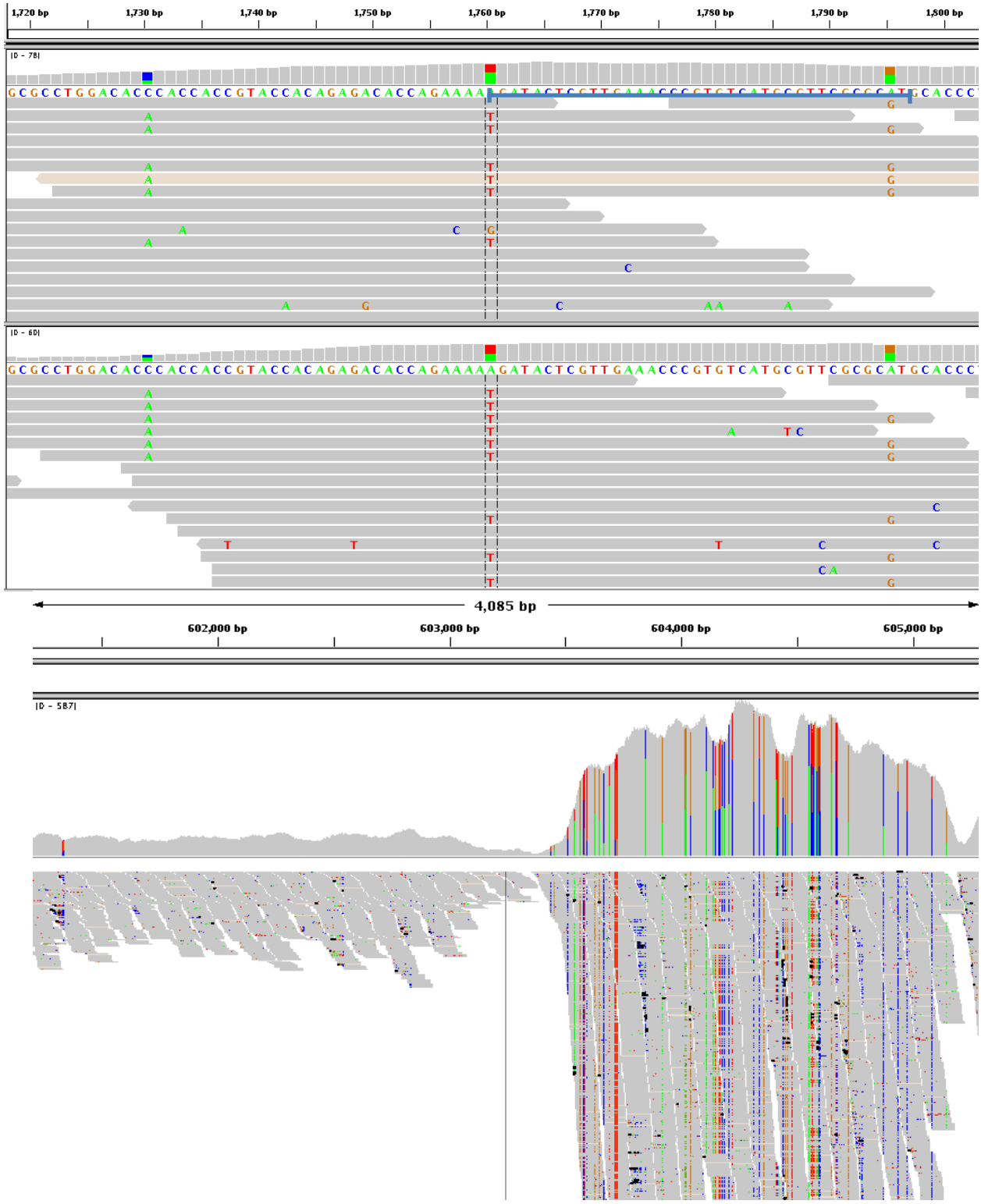


Figure 10: Incorrect genome assembly regions. (A) IGV screen shot showing alignment of a region with heterozygous sequence. (B) IGV screen shot showing clear pileup of reads over the location of the ROP5 gene which is known to exist in multiple copies (GT1 reference V6)

2.2.4 Final variant calls

We developed and applied an automated program to categorize the 33 candidate SNVs, using information from the annotated reference available through ToxoDB (Table 2). There were 8 SNVs in coding regions; 7 missense changes, and one silent mutation. Additionally, there were 6 intronic SNVs, and 19 SNVs that had no reference annotations. All 33 SNVs, as well as 14 of the filtered-out SNVs, were confirmed using PCR and Sanger sequencing. 31 of the SNVs unique to F-P2 were positively confirmed. Of the two that were not confirmed, both showed less than 10X coverage; one of these was the result of a duplication in a microsatellite that caused a nonexistent SNV to be called at the end of that region. The second was the result of a heterozygous area that was skewed enough to pass our filter at low coverage. This gives a false positive rate of 6.06%. If we had raised our cutoff to 10X it would have excluded these two calls, but would have removed one of our

A

Sequence feature	parent	F-P2
Total bp sequenced	6,074,293,500	6,527,102,400
Total paired end reads	40,495,290	43,514,016
Total read aligned	36,504,440	39,362,485
Total reads aligned (%)	90.1	90,5
Genome covered (%)	96,5	96,9
Human reads (%)	32.2	49,4
Reads in unassembled contigs (%)	0.9	0.7
Total SNPs called	982	1015

B

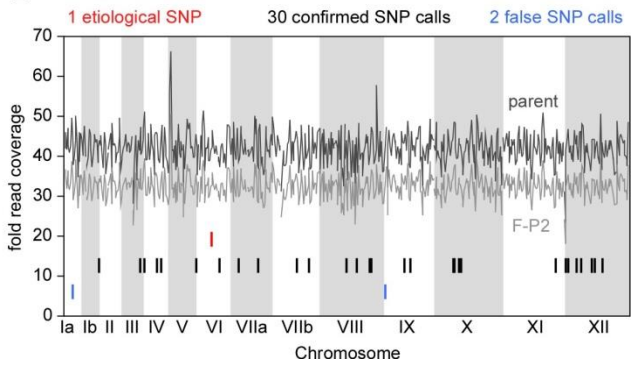


Figure 11: Results of paired-end Illumina re-sequencing of parent and F-P2 genomes. (A) Summary statistics from sequencing, alignment and SNV calling. Total reads aligned refers to total reads that aligned to both the human and GT1 references. All other stats are in reference to the GT1 reference only. (B) Genome coverage of Illumina reads across the chromosomes of parent and F-P2. Fold coverage is averaged over a 100 kb window. The chromosomal localization of the 33 called SNVs SNVs between parent and F-P2 are shown, differentiated by confirmed and false SNV calls. The validated, causative, SNV is highlighted in red. Reprinted from [19].

confirmed SNVs. As we expected, none of the 14 filtered-out SNVs were confirmed as true variations. Of the 14, 12 produced good sequence and correctly confirmed showing the reference allele in both samples, supporting our decision to filter these out. The two SNVs that could not be sequenced were in unaligned contigs and did not produce PCR products. This is likely due to the fact that these extra contigs have not been reliably assembled and may not represent truly contiguous sequence in the genome.

The causative mutation was confirmed by the Gubbels lab as an A to G mutation at chrVI:1579375. It resulted in a Phenylalanine to Serine substitution in the theoretical protein, TGGT1_049850, annotated as a C2 domain-containing protein, (for specific details see Appendix A). It was concluded that cosmid-library complementation mediated rescue of the F-P2 mutation failed due to the fact, that by random chance, there was only a single 40kb cosmid that contained the entire gene sequence. This experiment shows the power that mutational profiling can offer in instances where traditional means have failed.

Table 2: Genome localization of SNP calls, annotation and validation.

Chromosomal Location	SNP	Gene Name	AA change	Validation	Gene Annotation
Missense					
chrIb-1893131	A to T	TGGT1_064370	Asp to Val	Confirmed	conserved hypothetical protein
chrVI-1579375	A to G	TGGT1_049850	Phe to Ser	Confirmed	C2 domain-containing protein 2C putative
chrVIII-1544385	T to C	TGGT1_115970	Ser to Pro	Confirmed	DEAD_2 domain-containing protein
chrVIII-4001335	G to A	TGGT1_111590	Asp to Asn	Confirmed	casein kinase II beta j chain 2C putative
chrIX-2830729	A to T	TGGT1_032520	Ile to Phe	Confirmed	hypothetical protein
chrX-2661681	C to T	TGGT1_079790	Gly to Asp	Confirmed	kinesin motor domain-containing protein putative
chrXII-2880600	G to T	TGGT1_026590	Glu to Asp	Confirmed	phospholipase D active site motif domain-containing protein 2C putative
Sense					
chrIV-1438411	A to T	TGGT1_122960	Thr to Thr	Confirmed	conserved hypothetical protein
Intronic					
chrVIIa-847497	G to A	TGGT1_062210	N/A	Confirmed	conserved hypothetical protein
chrX-2171807	C to T	TGGT1_080660	N/A	Confirmed	conserved hypothetical protein
chrX-2889629	A to G	TGGT1_079320	N/A	Confirmed	GTP-binding protein 2C putative
chrXI-5635171	A to T	TGGT1_097900	N/A	Confirmed	hypothetical protein
chrXII-1242572	A to G	TGGT1_029900	N/A	Confirmed	conserved hypothetical protein
chrXII-1736436	A to G	TGGT1_028920	N/A	Confirmed	kinesin 2C putative
No Information (intergenic)					
chrIa-921699	A to G	N/A	N/A	False call	
chrIII-1992039	A to T	N/A	N/A	Confirmed	
chrIV-79513	A to G	N/A	N/A	Confirmed	
chrIV-1886744	A to T	N/A	N/A	Confirmed	
chrV-3081061	T to C	N/A	N/A	Confirmed	
chrVI-2430153	T to A	N/A	N/A	Confirmed	
chrVIIa-2939045	G to C	N/A	N/A	Confirmed	
chrVIIb-2603496	A to C	N/A	N/A	Confirmed	
chrVIIb-3906142	C to T	N/A	N/A	Confirmed	
chrVIII-2892685	T to C	N/A	N/A	Confirmed	
chrVIII-5399013	T to C	N/A	N/A	Confirmed	
chrVIII-5569348	T to C	N/A	N/A	Confirmed	
chrIX-127509	G to A	N/A	N/A	False call	
chrIX-2188760	A to G	N/A	N/A	Confirmed	
chrX-2057265	A to G	N/A	N/A	Confirmed	
chrXII-76471	A to C	N/A	N/A	Confirmed	
chrXII-345599	A to C	N/A	N/A	Confirmed	

2.2.5 Computational controls

We have two primary computational controls which support the accuracy of our SNV calls; the frequency of unique SNVs in the parent line as compared to F-P2, and the nucleotide frequency of the ENU induced mutations. The mutant F-P2 line was created by treating the parent strain with ENU, thus we would expect to find numerous point mutations in the mutant F-P2 that are not in the parent. Conversely, we would not expect to find any SNVs that are unique to the parent strain. The parent strain only showed 2 SNV calls that pass our filters and are unique to the parent with respect to F-P2. These calls are located adjacent to each other at the beginning Chromosome 1A at positions 7601 and 7602. These SNVs exhibit the same characteristics as the other heterozygous calls that passed our filter and we do not believe them to be real. Unfortunately, this area is highly repetitive, and we cannot successfully PCR this location to confirm this SNV. In subsequent whole genome sequencing runs on this sample, these two SNPs no longer appear. Based on this, we conclude that our parent strain does contain zero novel SNVs when compared to the F-P2 strain, as expected.

Additionally, we looked at the allele frequency of the ENU induced mutations. It has been shown in the literature that in other species ENU tends to preferentially mutate A/T nucleotides, but not exclusively[37]. Taking this into account, we would expect our list of 31 confirmed SNVs in the F-P2 background to be skewed towards mutation events at A/T bases. Of the 31 confirmed SNVs, 77% were a mutated A/T base and 23% were C/G, suggesting that ENU was likely the cause of these SNVs.

We compared this number to the list of shared SNVs between the 2 samples. These SNVs accrued over time due to random mutation events, and thus should not have the same skew in allele balance. As expected, these SNVs exhibit a significantly more even distribution; A 23%, T 22%, C 26%, G29%. These findings support the accuracy of our SNV calls and the effectiveness of this method at accurately identifying SNVs between a parent and mutant offspring line.

2.3 Final mutational profiling pipeline

The final pipeline for mutational profiling is outlined in Figure 12. The pipeline was designed to be as hands off as possible to enable a lab technician, with limited computational training, to create an accurate set of annotated variant calls from raw Illumina data. The user simply provides paired-end FASTQ files for two samples, a mutant and parent, as well as a Genbank (GFF) format reference for the organism of interest. If desired they can supply a list of FASTA formatted sequence files of known contamination sequences that will be included in the alignment to reduce mis-mappings. At this point all future steps are hands-off and automated. The pipeline will align reads with MOSAIK, call variants with FREEBAYES, filter the resulting variant calls as described above, annotate, and prioritize them with no user input. The resulting report is a tab separated Table that contains the list of filtered and annotated variants, separated into groups based on their possible impact. The pipeline has been updated to include our most recent software versions, MOSAIK 2.0[30] and FREEBAYES[32].

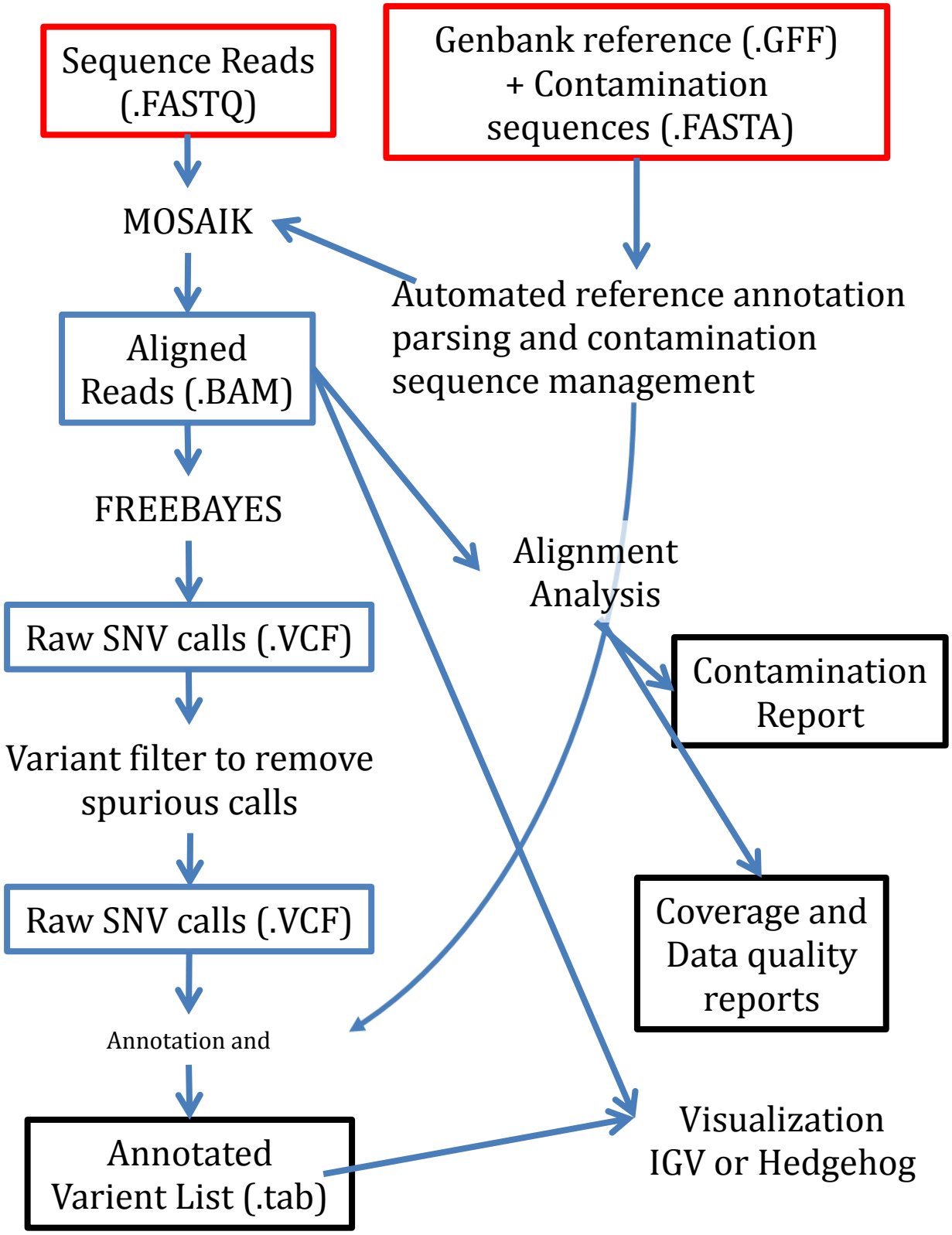


Figure 12: Final automated mutational profiling pipeline. Boxes indicate data files; red boxes indicate are input files, blue intermediate files, and black represent final output files.

2.4 Reference guided mutational profiling conclusions

The pipeline developed here has been used successfully to analyze over 20 samples. We have recently published an analysis of 15 of these samples along with a detailed analysis and description of the laboratory methods used to aid future research in *Toxoplasma gondii*[38], included in appendix B. This pipeline has also been used in 3 other publications; the DOC2 paper discussed extensively in this section and included in appendix A, as well as two papers published with collaborators Ira J Blader "Forward Genetic Screening Identifies a Small Molecule That Blocks *T. gondii* Growth by Inhibiting Both Host- and Parasite-Encoded Kinases." and Jeroen Pj Saeij "Genetic basis for phenotypic differences between different *T. gondii* type I strains."

The mutational profiling methods outlined here offer the ability to identify mutations in samples identified in forward genetic screen. The entire method, from DNA extraction to final variant calls can be completed in as little as 3 weeks, far faster and cheaper than many standard laboratory methods. Further, as was the case with DOC2, these methods are able to identify mutations that have been missed by traditional methods. Every researcher working in genetics has heard of research where a mutant screen has produced a strain with interesting phenotypes, whose causative mutation has eluded identification with traditional means. There are, unfortunately, no conclusive data on how many such cases are out there as these go largely unpublished. Even so, the methods and pipeline we have outlined in this Chapter could breathe new life into these studies, giving researchers a new angle to attack stubborn mutations and possibly identify the causative variants.

Chapter 3:

RUFUS

3.1 Motivation

Traditionally, studies using whole genome sequencing, such as mutational profiling, employ reference based mapping; sequencing reads are aligned to a reference genome, and variant callers are used to determine variations between the sample and a reference. As discussed in the previous Chapter, we developed such a mapping-based method to identify causative variations in strains of *Toxoplasma gondii* derived by chemical mutagenesis, to identify genes involve in the poorly-understood apicomplixan parasitic cell invasion and egress process[119,38] (Appendices A and B). This pipeline is based on the reference guided aligner MOSAIK[30] and variant caller FREEBAYES[32,33], developed in our lab. This method has proven very effective; we have used it successfully to analyze mutations in over 15 *Toxoplasma gondii* mutants[19,38].

Despite the success of these methods, we have encountered numerous challenges associated with mapping-based approaches that may limit the success of these methods in future projects. These methods are inherently limited by three factors: the existence and accuracy of a reference, the ability of the mapper to correctly place a read, and the degree to which a variant allele differs from the reference. These methods are extremely effective when analyzing small mutations

in organisms with extremely high quality reference sequences. This restricts whole genome analysis to well-studied organisms with pre-assembled high quality reference sequences, and can potentially cause valuable mutations to be missed.

3.1.1 Reference Limitations

The most significant limitation of mapping-based mutation detection methods is the requirement of a reference genome. This restricts genomic studies to well-studied model organisms with previously assembled reference genomes. Model organisms are generally chosen, not for their biological significance, but for their ease of use in the lab. This makes it impossible to directly study many important biological processes that are unique to other species, as they lack the basic genetic tools that are required to perform such analyses.

If a reference is available, the quality of the reference will have a direct impact on the success of mapping methods. No reference sequence is entirely complete, and as discussed in section 2.1.2, assemblers can have difficulty assembling contigs in, and around, repetitive regions of the genome. Many of these repetitive regions are often omitted from the genome, either intentionally to improve alignments, or through assembly artifacts. These regions are not comprised solely of uninformative repetitive DNA. One such case that has been well-characterized is the *Toxoplasma gondii* ROP5 gene, which is known to exist in numerous copies and has been linked to virulence[39,40]. The GT1 reference sequence

has been assembled with all copies of ROP5 collapsed to a single locus. Therefore, all of the sequence reads for each copy of ROP5 pile up on this single locus, pooling their respective sequence, making variant discovery in this important gene family impossible with normal alignment and SNV calling methods (discussed in Chapter 2 and shown in Figure 10). For mapping based methods it is important that the reference contains the entire genome, as any genomic regions omitted will be completely ignored and lead to false negative calls.

We can attempt to quantify the amount of genome sequence in our RH strain that has been either omitted from the GT1 reference, or unique to RH. As discussed in section 2.2.1, the Illumina sequencer's power is its ability to indiscriminately sequence all DNA in a sample. Therefore, the full set of reads sequenced on the instrument will represent the total genome of the sample, plus some amount of machine-specific error and human contamination. We can attempt to quantify the amount of the *T. gondii* genome that is not in the reference by analyzing the proportion of reads that have not aligned to the assembled sections of the GT1 reference or the human genome (Figure 14). Using one of our highest quality sequencing samples (nF-P2 described in section 3.5 and Appendix B), 90.9% of the total reads sequenced align to the 14 assembled chromosomes, covering greater than 99.87% at a depth of 10 or greater. The GT1 reference contains an additional 328 un-aligned contigs, who have not yet been confidently placed in one of the chromosomes. For this sample, 5.7% of the reads align to these unaligned contigs, indicating these contigs may account for a significant proportion of the genome. In addition, the aligner could not place 2.03% of the reads in either the GT1 *T. gondii* or

human hg19 reference. These reads could belong to omitted regions of the genome, or alternatively, may surface from simple machine or library prep errors. Some proportion of unaligned reads are expected, previous studies report alignment percentages from 70% to 94% [41-44]. In an effort to separate random error from high quality reads, we assembled the unaligned reads with RUFUS.overlap (described in section 3.4.5), producing 2084 contigs with a read depth greater than 30 fold. 423 contigs are longer than 1 kb, the longest being 4.9 kb, with a total length of 712 kb. This indicates a significant amount of unaligned reads that may represent DNA from contiguous genomic regions and not due to random error. Using BLASTN to align the reads longer than 1 kb against the NCBI nucleotide database, 329 contigs have significant sequence homology to know regions of other *T. gondii* strains, including numerous known genes (Appendix C). These results suggest that the current reference has omitted sections of genomic DNA that could be of biological importance.

In addition to genomic sequence that is missing from the reference, misassembled regions within the genome will further hinder variant detection. Of the 14 assembled chromosomes, there are numerous regions where our RH samples alignment show greater than the expected coverage. This indicates that either RH contains additional copies of these sequences, or more likely, these regions represent assembly errors in the reference where multiple gene families have been collapsed on top of each other. These regions have been shown to contribute a vast number of false positive variant calls, and must be filtered out in order to reveal true variants. In our DOC2 paper, 259 of the 292 called variants were false positives

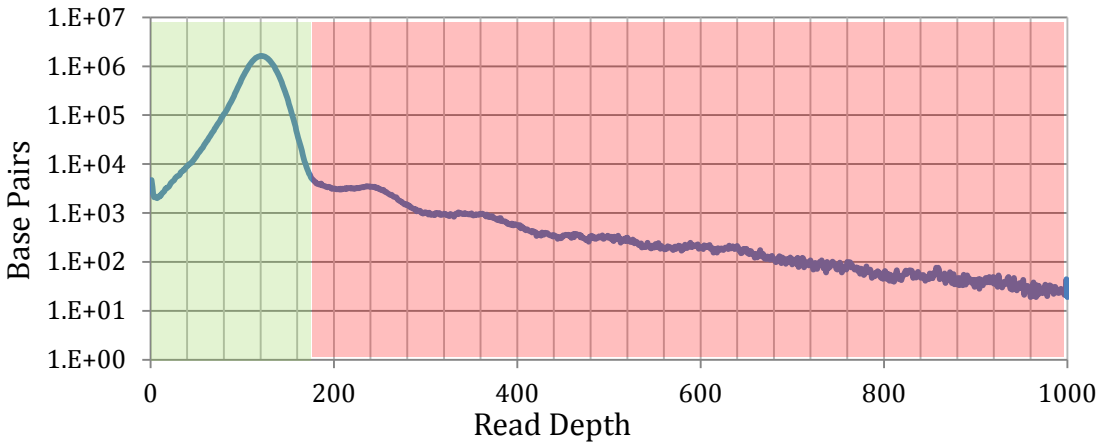


Figure 13: *T. gondii* Coverage Frequency Plot. Distribution of coverage depths per base in a MOSAIK alignment of F-P2 against the *Toxoplasma gondii* GT1 genome. Regions of the genome predicted to be correctly represented as haploid DNA are highlighted in green, regions that appear at higher copy number highlighted in red. Sample used was nF-P2, total reads 83,487,194.

contributed by these regions. We can attempt to quantify the amount of genetic information that is lost to these areas by analyzing the alignment coverage, illustrated within the graph shown in Figure 13. For this sample, the average coverage for unique regions of the genome is 120 fold. A peak of duplicated DNA can be seen at an average coverage of 240X. As a rough estimate, we can take the point between these two peaks, at 180X, and consider any point below this as correctly assembled as 1X sequence, and all points above it to represent duplicated sequence. This yields an estimation of 97.4 % of the assembled genome that is correctly represented as haploid DNA, and 2.59% as collapsed higher copy regions. In total, of the original 83 million reads, 5.75% are in unassembled contigs, 2.02% are from regions missing

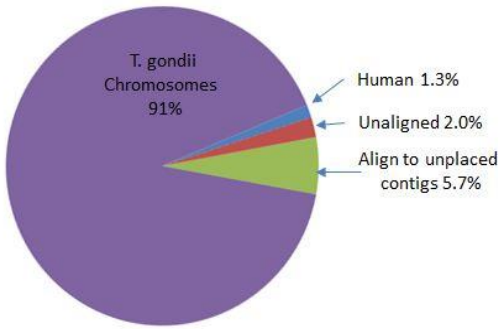


Figure 14: Sequence Read Placement. Quantification of the alignment location of Illumina whole genome reads aligned to the T. gondii GT1 reference. Sample used was nF-P2, total reads 83,487,194.

from the genome, and another 2.35% are mapped to duplicated regions. This gives a total of up to 10.2% of the genome that may be missed by current mapping based methods.

3.1.2 Mapping algorithm limitations

Reference-based mapping methods themselves introduce both limitations and bias to sequence analysis. As discussed in Chapter 2, all current mapping-based methods used in second generation sequencing use a similar two-step alignment method. First, an exact string match using a fixed-length hash to the reference is used to define candidate regions. After candidate regions are defined, a more sensitive local alignment is used to place the read, allowing for substitutions and insertions or deletions. This method works extremely well when placing unique reads in the genome with few to no errors or mutations. However, if a read is non-unique, or contains significant sequence variation from the reference, it will hinder mapping. If a read cannot be placed confidently in the genome, it cannot be used in variant calling and thus its sequence is ignored.

Many regions of the genome have extremely similar or identical sequence, as is the case with many gene families. Sequence reads from these regions will have almost identical DNA, and when aligned, will contain multiple regions with similar, or exactly the same, alignment scores. Most aligners, including MOSAIK, will exclude reads that have more than one equally-possible alignment. This leads to many

regions of the genome where variant discovery is simply impossible. Additionally, the hashing step may return thousands of candidate locations for a given read. To prevent these reads from bogging down alignment, most aligners will randomly pick a subset of candidate regions to perform a full alignment. This can cause a higher scoring alignment to be missed entirely, causing the read to be misplaced or excluded. Two other methods used to handle the multiple alignment problem are to place the read randomly to one of the possible locations, or to place a copy of the read in every location to which it maps. All three of these methods have their advantages, but all of them will ultimately make variant discovery in these regions impossible. Frequently, in some projects (such as the 1000 Genomes), repetitive regions are simply masked-out and completely ignored to improve confidence in the final call set.

Using a reference introduces a bias towards the reference in both steps of alignment that can hinder mutation detection. The first bias occurs in the hashing step, where alignment candidate regions are identified. In this step, the read is divided up into fixed length k-mers, and each k-mer's exact matches in the reference are recorded. If a read contains a variation, even a single nucleotide change, all k-mers that overlap that region will no longer match to their appropriate location in the genome. This immediately reduces the chances of finding the true location where the read belongs in the genome. The most drastic case of this is when two SNPs occur spaced apart equal to the k-mer size. In this case all k-mers in this read would be completely missed and the read would never be aligned, even though its alignment score would be quite high. Furthermore, once candidate regions are

determined, alignment algorithms introduce another level of bias. When aligning a read to the genome, the best score possible is a perfect match. Any difference in a read compared to the reference sequence will reduce the overall alignment score, and decreases the chance that the read's map score will be high enough to consider it an acceptable alignment. If the sample has true variation with respect to the reference, reads that represent this variation will inherently have a reduced probability of mapping. The bias against mapping is increased even further the larger the variation from the reference making it difficult to find insertions or deletions in mapping based methods.

These limitations in reference-guided assembly introduce inherit bias against finding variation. In our opinion, aligning to a reverence genome and then calling variants is a case of putting the cart before the horse. Alignment is inherently biased against mutations, it would be far better to first identify variations in a data set, and only *after* that, determine the sequence context of that variation. This would remove bias that limits mutation discovery and restricts analysis to the mappable genome.

3.2 RUFUS

We set out to develop a variant discovery tool to address the limitations associated with reference guided alignment by creating a tool that does not use an assembled reference sequence, and, as opposed to whole genome assembly, can be

run on a computer with similar hardware and time requirements as required for current mapping methods. To fulfill these goals, we developed RUFUS; a k-mer based reference-free method for identifying variants in Next Generation sequencing data. RUFUS compares raw Illumina whole-genome sequence data from multiple samples in order to identify specific sequences that differ between the samples, thereby identifying reads that represent mutation events between the two samples. Unbeknownst to us, a similar tool, NIKS, was simultaneously developed at the Max Plank Institute in Germany[45] which uses a similar k-mer based approach to identify mutations between closely-related samples. Both of these methods have numerous advantages over mapping based methods; there is no need for a reference sequence so variation can be identified in any organism, in any sequence. There is no possibility of misalignment, which is the major contributor of false positive and false negative variant calls. Additionally, variant detection is not limited to a mappable genome; variants can be identified in missing and hard to map sequence. The lack of reference bias increases the chances of identifying rare variants and removes bias against large events such as insertions and deletions. While NIKS method uses a coverage cutoff approach, and is applicable to homozygous/haploid sequence mutations in unique (i.e. non-repetitive) sequence, RUFUS goes further using a Bayesian detection method to analyzing changes in specific k-mer frequencies between the samples. This allows RUFUS to detect all types of variation including SNVs, insertion/deletions, translocations and mobile elements, copy number variants, and will find both homozygous and heterozygous events. The greatest collective benefit of these reference-free methods is that they will enable

sequencing research in organisms that have either a very poor reference or none at all. It will also improve detection of highly variable and repetitive genomes that are difficult to analyze with current methods, such as *Plasmodium spp*.

3.2.1 RUFUS concept explained

If one were to compare all of the DNA sequence between two closely related samples, the vast majority of sequence will be identical between them, and thus relatively uninformative with respect to variant detection. Using expensive computation, such as alignment or assembly, on the majority of the data is a waste of resources, and will only serve to confuse the final results. However, any mutation between samples will create a unique stretch of DNA sequence between the two samples. Comparing the raw reads would not work, as the read may contain errors preventing simple string matching, and the probability that a complementary read starts at the same location in the second sample is extremely low. Instead the reads can be broken into k-mers. These k-mers can then be compared between the samples to identify specific sequences unique to either sample. Using this basic principle, RUFUS is able to identify variations without the need to map reads to a reference, or assemble the entire genome. We have extended this idea to detect changes in the abundance of sequences, and applied a Bayesian algorithm able to detect copy number and heterozygosis variations as well.

3.2.2 K-mer histogram analysis

To generate k-mer counts, we use Jellyfish[46], an extremely fast and memory efficient program developed at the University of Maryland. Jellyfish will count the occurrence of every k-mer in a FASTQ file of a given length. To visualize the data in a manner, k-mer counts can be represented as histogram showing the abundance of each k-mer count in the sample, Figure 15. This histogram is the basis of RUFUS. Due to the overlapping nature of whole genome sequence reads, k-mers derived from error free sequence will pile up, producing high frequency k-mers that reflect the genome of the sample. Illumina sequencing has a pseudo-normal coverage distribution, and genomic k-mers will largely exist as a single peak, seen in Figure 15A between 50 and 150. The shape and location of this peak will change between different sequencing samples, its center determined by the average coverage of the sequence reads, and the width of the peak will reflect the variability of the coverage. Conversely, sequencing errors will produce rare k-mers that exist at low frequency, particularly singleton k-mers that comprise over 76% of the k-mers sequences in Figure 15A.

The k-mer histogram also contains copy number information about the sample sequenced. Unique, single copy, regions within the genome produce k-mers with similar depths, duplicated regions of the genome will produce k-mers with twice the depth of unique sequence. This effect is evident in the k-mer frequency histogram when viewed on a log scale (Figure 15B). Each peak in this graph

represents k-mers from DNA sequences at successively higher copy number within the sample.

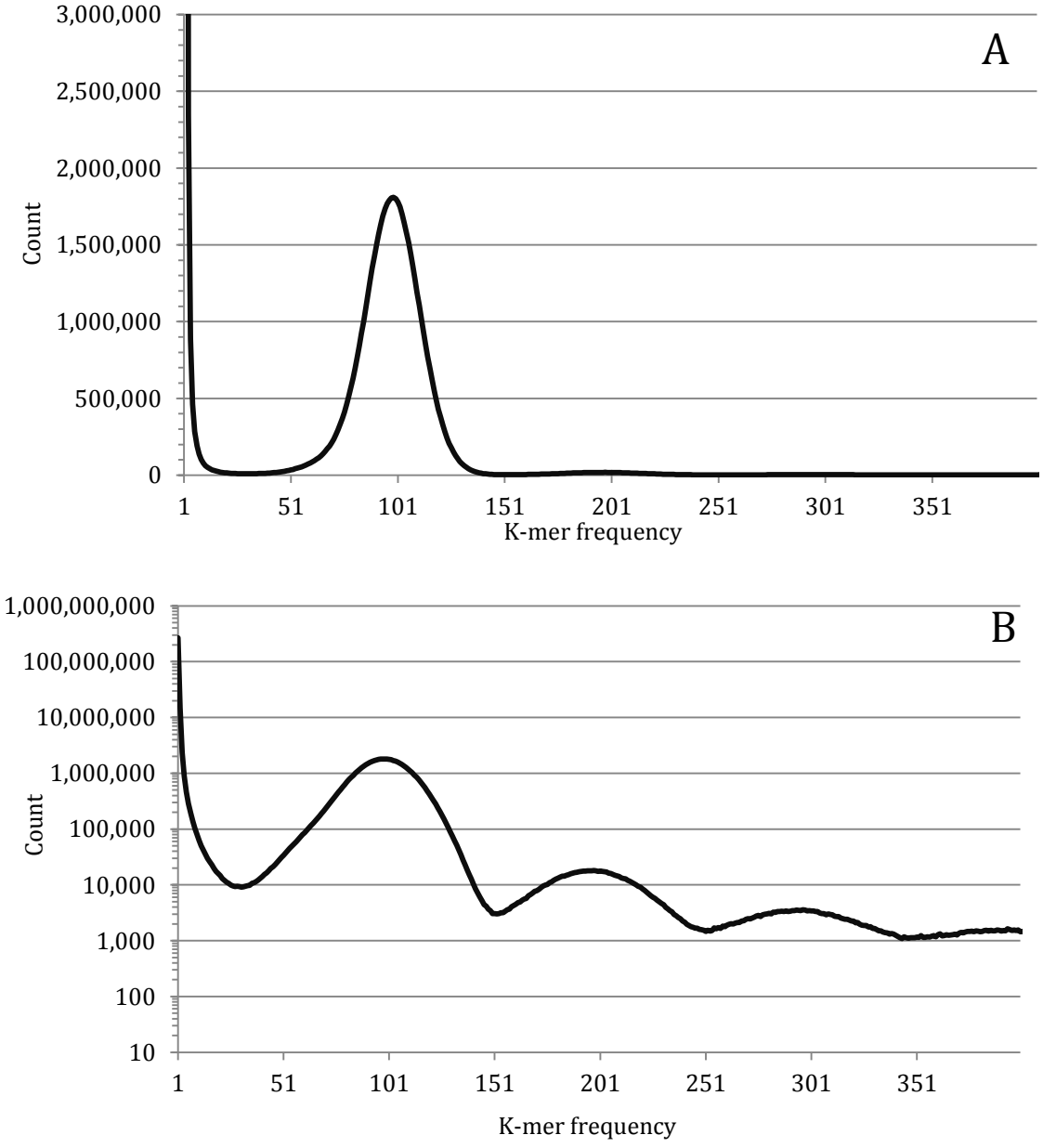


Figure 15: K-mer Histogram. For all graphs, Y axis represents the number of k-mers that have a frequency (depth) in the sample equal to the value on the X axis. (A) Raw k-mer frequencies. (B) Log scale representation of the k-mer frequencies in A.

To demonstrate these principles experimentally, the draft reference sequence for *Toxoplasma gondii* GT1 can be used to bin the k-mers based on their occurrence in the reference genome (Figure 16A). Using the first inflection point as a reference point (depth = 50), over 99.95% of the k-mers to the right of this point exist in the reference sequence. Conversely, over 99.53% of the k-mers that are not present in the reference genome are below this point, suggesting that low frequency k-mers are in fact derived from sequence errors and contamination. Of the k-mers that are unique in the reference sequence, and thus represent single copy DNA, 99.2% exist within an interval of 50 to 150. This supports our claim that unique DNA will produce k-mers with similar counts and produce a contiguous single peak. As the coverage increases, this trend continues, though there is a clear increase in variance as the copy number increases. 2 copy k-mers show 77.2% of the reads within 150 to 250, and 3 copy k-mers show 58.2% of their reads in the interval 250 to 350. Peaks in the wrong location indicate differences between the reference and the sample sequenced. For example, in Figure 16A the bin of single copy k-mers shows a second peak centered at 200 where there should only be duplicated k-mers. This indicates sequences where the reference contains a single copy, though in the sample this k-mer is in fact duplicated. Additionally, the Novel K-mers show a peak at 100. This is due to variants between the sample and the reference, such as SNPs which will produce k-mers that are not in the reference. Thus, the full k-mer frequency histogram gets its' periodic shape from the fact it is actually the sum of a series of independent distributions; one for each copy number state in the sample as well as a distribution that describes the error of the machine.

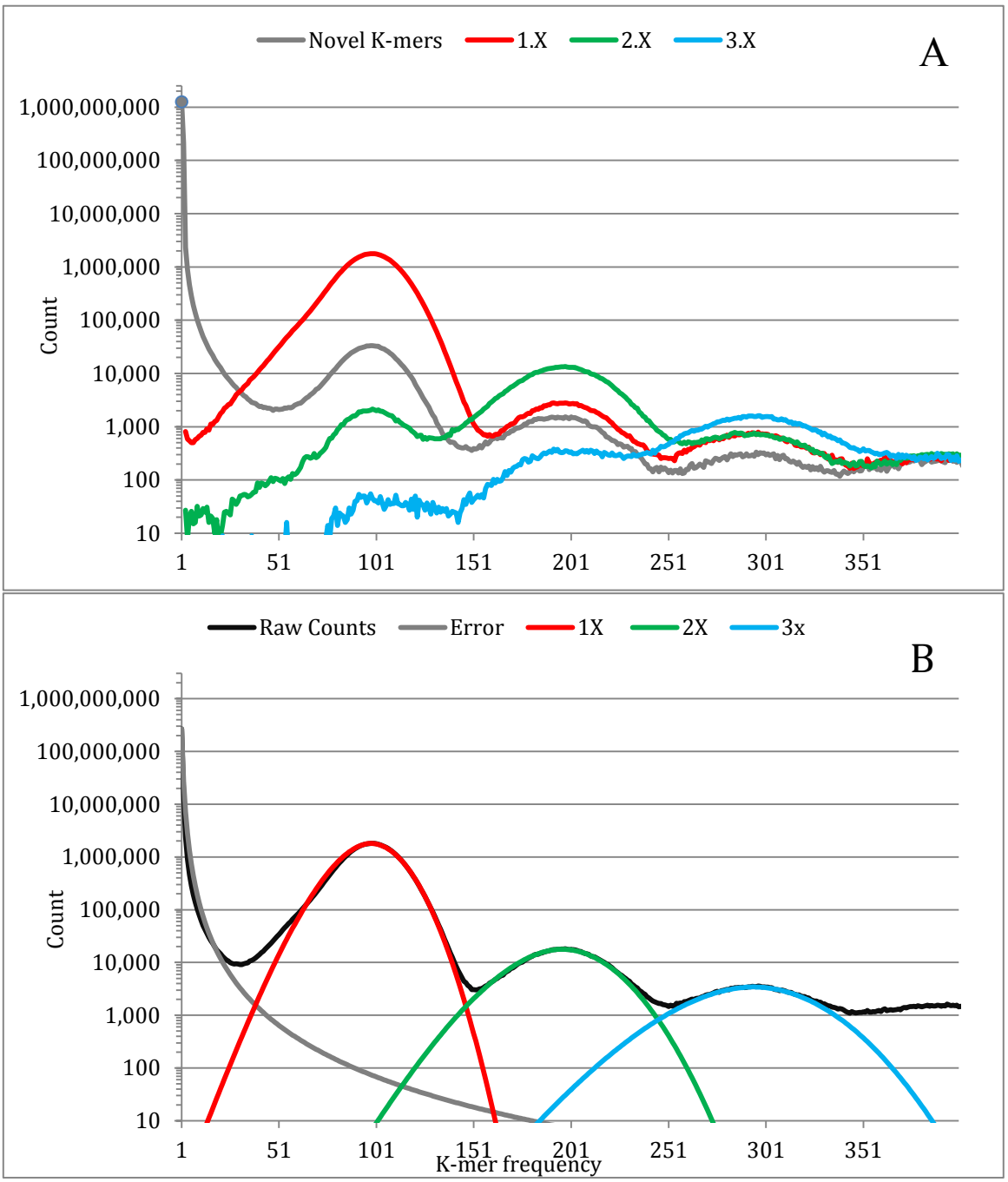


Figure 16: Copy Number K-mer Histograms. The K-mer frequency histogram reflects the copy number composition of the genome. (A) k-mers from *T. gondii* separated by their occurrence in the reference genome; Novel k-mers are never seen in the reference, 1X have a single occurrence, 2X are seen twice, and 3X have 3 occurrences. (B) Example model depicting the separate copy number distributions present in the k-mer histogram. Raw Counts represent the original data, Error is the model of the error, copy number distributions are labeled 1X, 2X, and 3X.

We can model the underlying distributions that create the k-mer histogram (Figure 16B), and leverage this model to detect variants between samples. By detecting specific k-mers whose count in each sample lie on different distributions, we can detect both mutations, such as SNVs and insertion/deletion events, as well as copy number changes in a single pass variant detection method.

3.3 RUFUS methods

An overview of the RUFUS pipeline for mutation detection is outlined in Figure 17. This pipeline is designed to perform mutation detection on a pair of samples, such as mutant and its parent. Initially jellyfish is run on each sample to produce k-mer counts for each sample, as well as a histogram of the k-mer

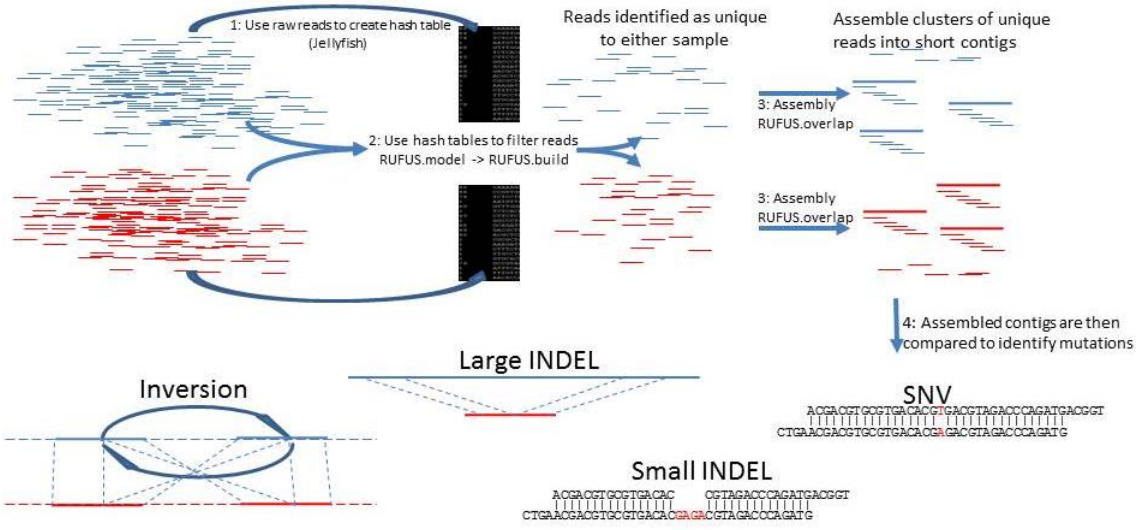


Figure 17: Overview of the RUFUS Workflow. Jellyfish is used to create k-mer frequency Tables for both samples (1). RUFUS.model is used to model each sample, and RUFUS.build uses these models to filter the original FASTQ reads (2). RUFUS.overlap is then used to assemble the filtered reads into contigs that represent variation (3). The assembled contigs are compared in to detect variation.

frequencies. Here we will cover the concepts behind each step in RUFUS. For detailed information on the algorithms the source code is available on the Marth Lab website (http://bioinformatics.bc.edu/marthlab/).

3.3.1 RUFUS.model

RUFUS.model takes the raw k-mer frequency histogram, produced by Jellyfish, and fits models of the underlying copy number distributions to the raw k-mer histogram for each sample. First, contamination is estimated by fitting a curve using equation 1 to the portion of the frequency histogram to the left of the first inflection point. As discussed in section 3.2.2, the majority of the reads in this region will be due to contamination and errors. The error model is subtracted from the original k-mer histogram and the resulting histogram is used to fit the genomic k-mer model. Genomic k-mers are modeled as a series of asymmetrical normal distributions with

Equation 1: Error model. i is equal to the k-mer count, x is the fitted constant, and Sk is the count for unique k-mers in the sample taken from the histogram.

$$\text{ErrorModel} = \frac{1}{i^x} * Sk$$

Equation 2: Genomic k-mer model. N indicates the normal distribution with parameters mean and variance, μ average coverage for the single copy peak, σ standard deviation for the single copy peak, i is the copy number state from 1 to ∞. sk, f and p are variable factors that are fit by RUFUS.model to account for the skew of Illumina data toward lower coverage. f determines the rate at which the standard deviation increase with each copy number increase, sk determines the magnitude of left hand skew on the normal distribution, and p corrects the slope of that skew. K is the total number of k-mers present at the given copy number denoted by i.

$$\text{GenomicKmerModel} = \sum_{i=1}^{cnum} N(\mu * i, \sigma + ((i-1) * f)) * k_i$$
$$if\ (x < \mu * i)\{\sigma = \sigma + ((\mu - x) * sk)^p$$

increasing variance for each successive copy number (equation 2). RUFUS iteratively fits each of the 5 parameters in equation 2 to produce the final model, examples shown in Figure 20B.

Initially a random walk was used to fit the model, however this often resulted in parameters becoming stuck in local minima and producing poor final fit. Instead the following method was developed which improves parallelization and avoids local minima. This method uses a 3 pass algorithm, where in each pass values are fit for each of the values in equation 2 in this order, f, μ, σ, sk, and p. The order of parameter fitting is based on their perceived impact on the model, from greatest to smallest. For the given parameter, a low and high range is set. Then in parallel, 10 independent models are generated, using 10 evenly spaced values for the given parameter between the high and low values. The value with the lowest sum of squares is chosen. New high and low values are selected using the newly selected best value ± 10%. The loop is run until the high and low values are within 0.1% of each other. This both allows extremely efficient multithreading, and reduces the possibility of becoming stuck in a local minima by testing a wide range of values each pass and by allowing the values to range both above and below the original high and low values. The initial values for μ and σ are calculated from the k-mer histogram. μ is taken as the k-mer depth with the largest value after the first inflection point. σ is calculated by finding the k-mer depth, whose value is $e^{-1/2}$ the value of μ. The values for the shape factors range as follows; $1 \leq f < \infty$ with 1 and 20 as the initial high and low, $0 < sk < \infty$ with .000001 and 2 as initial values, and $1 \leq p < \infty$ with 1 and 5 set as initial values.

Using the best fit model, RUFUS.model produces an estimate of the genome size, as well as the proportion of the genome that exists at each copy number. Data likelihoods are saved for each of the distributions separately to be used by RUFUS.filter. A model is generated for each sample independently, allowing sample specific estimations of the error, expected coverage's for each copy number, and the variation of copy number.

3.4.2 RUFUS.build

Using data likelihoods created by RUFUS.model, RUFUS.build intersects k-mer frequency Tables for a subject and reference sample, to identify k-mers that represent either mutations or copy number variants between the two samples. For each sample, RUFUS.build takes a Jellyfish k-mer count Table and data likelihoods generated by RUFUS.model, as well as prior probability for mutation events, a prior probability for a copy number change, and a P-value cutoff. To reduce memory, the two Jellyfish Tables are read in simultaneously and each k-mer is only stored in memory until its match is seen in the other sample. The memory for this step can be reduced to almost zero by pre-sorting the hash Tables. However as RUFUS.build often takes far less memory than jellyfish this is usually not done for *Toxoplasma gondii*. For each k-mer, a Bayesian algorithm is applied that calculates the posterior probability that the k-mers count in each sample indicate that the k-mer represents either a copy number event or a mutation, equations 3 and 4. If the posterior probability that the k-mer represents either type of variation is greater than the

supplied Pvalue, the k-mer is saved as either a mutation event or copy number event. RUFUS.build creates two files; a file containing k-mers that represent mutation events, and a file containing k-mers that represent copy number events.

$$P(C_a \neq C_b \mid K_a, K_b) = \frac{\sum_{a=1,b=1}^{for\ all\ a\neq b} (K_a,K_b \mid C_a,C_b) * P(C_a|K_a) * P(C_b|K_b) * PriorCopy}{P(K_a,K_b)}$$

Equation 3: Probability of Copy Number Change. The equation below calculates the posterior probability that for a given k-mer in samples a and b, the copy number state (Ca and Cb) for the k-mer is not the same in the two samples given the k-mer's count in those each sample (Ka and Kb). This is calculated for all copy number states not equal 0. PriorCopy is the prior probability that a copy number event has occurred, provided by the user.

$$P(C_{a=0} \text{ and } C_{b\geq 1} \mid K_a, K_b) = \frac{\sum_{b=1}^{\infty} (K_a,K_b \mid C_a,C_b) * P(C_a|K_a) * P(C_b|K_b) * PriorMut}{P(K_a,K_b)}$$

Equation 3: Probability of a Mutation Change. The equation below calculates the posterior probability that a k-mer in sample b represents sequence unique to that sample, thus a mutation. Posterior probability that the copy number state in sample a (Ca) is equal to 0 (does not exist in the sample) and the state in sample b (Cb) is equal to or greater than one, given the k-mer's count in each sample (Ka and Kb). PriorMut is the prior probability that a mutaion event has occurred, provided by the user

3.4.4 RUFUS.filter

RUFUS.filter identifies reads in the original FASTQ that represent either mutations or copy number events between the two samples. RUFUS.filter requires a FASTQ file for the sample to be filtered, as well as the Mutation.HashTable and CopyNumber.HashTable files produced by RUFUS.build. For each read in the sample FASTQ, RUFUS.filter compares each k-mer in that read with the Tables identified by RUFUS.build to determine if the read contains variant k-mers. If any k-mer matches are detected the read is a candidate for a mutation. To increase stringency a minimum number of variant hashes can be set for both mutations (default 3) and copy number (default 30), within a given window (default 10bp for mutations and 30 for copy number). If greater than the minimum number of hashes is identified as variant, within the given window in a read, the read is saved as either a mutation or copy number event read. RUFUS.build and RUFUS.filter are run as separate steps to improve parallelization and distribution of filtering to a cluster with minimal memory.

3.4.5 RUFUS.overlap

Once the set of reads that represent variation have been separated from the total genomic reads by RUFUS.filter, they are assembled into contigs that represent variant sequence between the two samples. The assembly can be performed using any available assembler; however the sensitivity of the assembly will determine the sensitivity of the final mutation discovery. The current state of the art graph based assemblers, Velvet[24] and Minimus[47] , generated assemblies in a matter of seconds on

the reduced set of filtered reads. However, they dropped out contigs from known variants during assembly. In order to generate assemblies in useful timeframes sophisticated assemblers, such as these, use heuristics to increase the assembly speed. However, many of these heuristics are no longer needed due to the fact that the overwhelming majority of reads have been filtered out by RUFUS. In many cases these heuristics cause valuable reads to be rejected leading to a reduction in variant detection sensitivity. To address this we developed RUFUS.overlap, a simple greedy overlap assembler, which uses an iterative process to build contigs.

RUFUS.overlap performs assembly using a 4 step process, using two versions OverlapHASH and OverlapRegion. Both versions work in essentially the same way. First, all reads are loaded into memory. Due to the vast reduction in reads this is now possible with very little memory. A read is taken, and compared to all other reads. For each comparison the reads are slid across each other, in both the forward and reverse complement orientation, calculating the number of bases that match between the reads in every configuration, if two bases do not match whose quality score is over 10, the overlap is not considered. This prevents heterozygous sequences from being collapsed and masking variation. The highest scoring match is selected and the reads are collapsed.

In the first two steps, the reads are assembled with OverlapHASH. OverlapHASH reads in each raw read, converting any base with a quality less than 5 to an N, and subsequently trimming all leading and trailing N's. Reads are overlapped described above, using a hashing method to target overlaps and increase

speed. On this first pass reads are only collapsed if the similarity of sequence is over 98%, a minimum overlap of 50 bases and a hash size of 24. The hash size of 24 does not limit the assembly, as any overlap of at least 50 bases with 98% identity would have at least 24 contiguous identical bases between them, this simply reduces the number of useless calculations. As a second pass the reads are again read into OverlapHASH, preserving the depth of the contigs at each base. For contigs that were overlapped in the last pass (any contig with at least one base with depth greater than 1), hanging ends with a depth of 1 are trimmed off, this eliminates reads with high base quality singleton errors from blocking extension. Overlap is performed again, further extending the contigs, this time requiring 98% match, minimum overlap of 30 bases. Now only read reads that have found at least one overlap match are reported. At this stage, to correct for internal singleton error bases of high quality that will prevent two contigs whose belong together from overlapping, all base qualities are replaced with the sequence depth. Now, OverlapRegion is run twice to collapse any additional contigs. For both passes a minimum match of 95% is used and minimum overlap of 30 bp. OverlapRegion is similar to OverlapHASH though it does not use a hash method to seed assembly, thus ensuring that any possible overlap is considered.

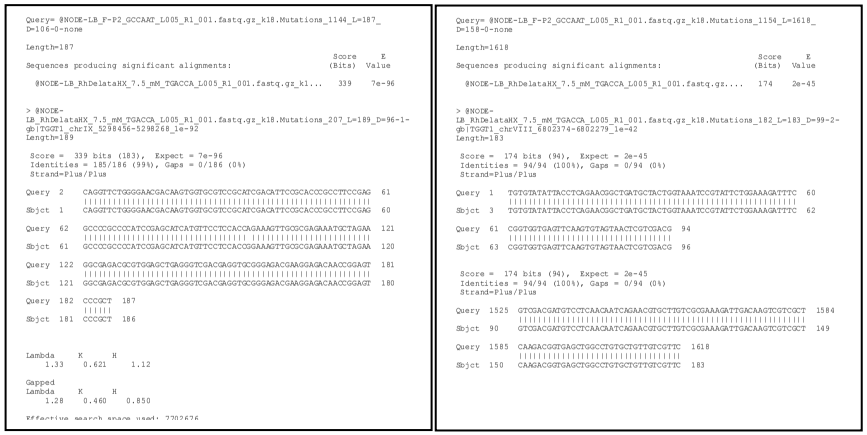
This method works extremely well for assembling RUFUS data into contigs. It would be far too slow for whole genome assembly, but due to the fact that the majority of data has been removed, this method completes in a reasonable amount of time. Additionally all repetitive regions of the genome have been removed, leaving only the variant sequences. This method will also preserve heterozygous

sequences with as little as one base difference between them, an advantage over many other assemblers designed for whole genome data.

3.4.6 Mutation discovery

Mutation discovery can be achieved in multiple ways and will differ depending on the organism and specific research goals. Here we will cover the two chief methods we employ; BLASTN based Reference Free and mapping based Reference Assisted.

Reference Free: Blastn is used to align assembled contigs produced by RUFUS.overlap between the two samples. This will produce a blastn report that describes the relationship between variant contigs in the samples. In the case of a SNV or indell (with unique sequence) this will show two overlapping contigs with the variant in the center (Figure 18A). In the case of structural events, or insertions where the inserted sequence is not unique, breakpoints will be identified, if the inserted/deleted sequence is unique within the genome the entire region will be assembled (Figure 18 B and C). This method is capable of identify any variation where there is at least one base pair different between the two samples including SNVs, insertion/deletions, translocations, inversions, copy number, etc. This method should be used if no reference is available. If a reference is available this method can be used to detect variants in regions that may be missing from the reference as well to improve the detection of structural events and indels.



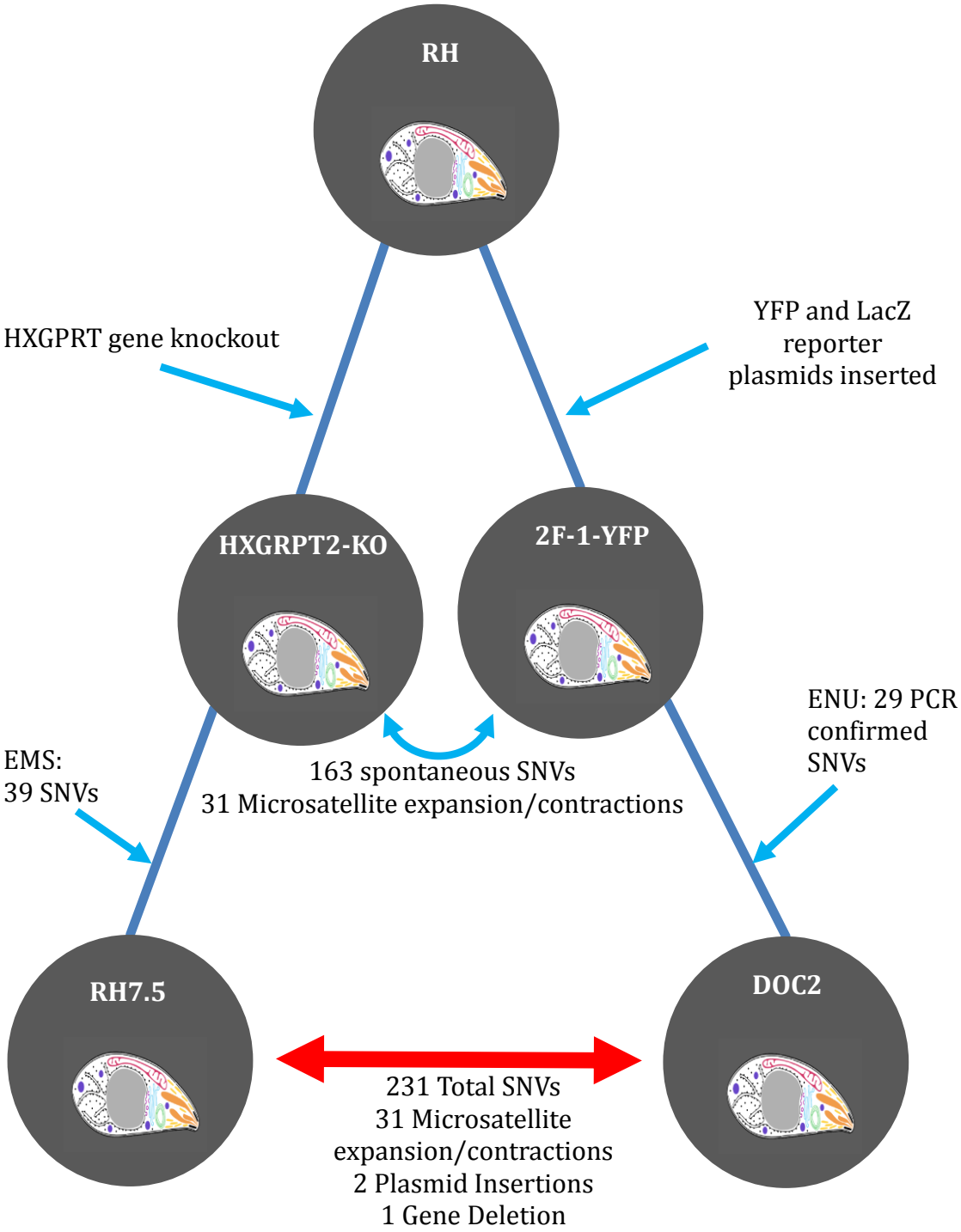
```
Query= @NODE-LB_F-P2_GCCAAT_L005_R1_001.fastq.gz_k18.Mutations_1144_L=187_
D=106-0-none

Length=187

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  @NODE-LB_RhDelataHX_7.5_mM_TGACCA_L005_R1_001.fastq.gz_k1...       339     7e-96


> @NODE-
LB_RhDelataHX_7.5_mM_TGACCA_L005_R1_001.fastq.gz_k18.Mutations_207_L=189_D=96-1-
gb|TGGT1_chrIX_5298456-5298268_1e-92
Length=189

 Score =  339 bits (183),  Expect = 7e-96
 Identities = 185/186 (99%), Gaps = 0/186 (0%)
 Strand=Plus/Plus

Query  2    CAGGTTCTGGGGAACGACAAGTGGTGCGTCCGCATCGACATTCCGCACCCGCCTTCCGAG  61
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CAGGTTCTGGGGAACGACAAGTGGTGCGTCCGCATCGACATTCCGCACCCGCCTTCCGAG  60

Query  62   GCCCCGCCCCATCCGAGCATCATGTTCCTCCACCAGAAAGTTGCGCGAGAAATGCTAGAA  121
            ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  61   GCCCCGCCCCATCCGAGCATCATGTTCCTCCACCGGAAAGTTGCGCGAGAAATGCTAGAA  120

Query  122  GGCGAGACGCGTGGAGCTGAGGGTCGACGAGGTGCGGGAGACGAAGGAGACAACCGGAGT  181
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGCGAGACGCGTGGAGCTGAGGGTCGACGAGGTGCGGGAGACGAAGGAGACAACCGGAGT  180

Query  182  CCCGCT  187
            ||||||
Sbjct  181  CCCGCT  186


Lambda      K       H
    1.33    0.621     1.12

Gapped
Lambda      K       H
    1.28    0.460     0.850

Effective search space used: 2202676
```

```
Query= @NODE-LB_F-P2_GCCAAT_L005_R1_001.fastq.gz_k18.Mutations_1154_L=1618_
D=158-0-none

Length=1618

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  @NODE-LB_RhDelataHX_7.5_mM_TGACCA_L005_R1_001.fastq.gz....         174     2e-45


> @NODE-
LB_RhDelataHX_7.5_mM_TGACCA_L005_R1_001.fastq.gz_k18.Mutations_182_L=183_D=99-2-
gb|TGGT1_chrVII1_6802374-6802279_1e-42
Length=183

 Score =  174 bits (94),  Expect = 2e-45
 Identities = 94/94 (100%), Gaps = 0/94 (0%)
 Strand=Plus/Plus

Query  1    TGTGTATATTACCTAGAACGGCTGATGCTACTGGTAAATCCGTATTCTGGAAAGATTTC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3    TGTGTATATTACCTAGAACGGCTGATGCTACTGGTAAATCCGTATTCTGGAAAGATTTC  62

Query  61   CGGTGGTGAGTTCAAGTGTAGTAACTCGTCGACG  94
            |||||||||||||||||||||||||||||||||
Sbjct  63   CGGTGGTGAGTTCAAGTGTAGTAACTCGTCGACG  96


 Score =  174 bits (94),  Expect = 2e-45
 Identities = 94/94 (100%), Gaps = 0/94 (0%)
 Strand=Plus/Plus

Query  1525  GTCGACGATGTCCTCAACAATCAGAACGTGCTTGTCGCGAAAGATTGACAAGTCGTCGCT  1584
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  90    GTCGACGATGTCCTCAACAATCAGAACGTGCTTGTCGCGAAAGATTGACAAGTCGTCGCT  149

Query  1585  CAAGACGGTGAGCTGGCCTGTGCTGTTGTCGTTC  1618
             |||||||||||||||||||||||||||||||||
Sbjct  150   CAAGACGGTGAGCTGGCCTGTGCTGTTGTCGTTC  183
```

Figure 18: Blast Variant Detection. Blast can be used to identify a wide variety of variants. For each panel, alignments are shown for illustration only and are not required to detect variants. Four examples of variant detection are shown; SNV, Insertion/Deletion, Copy Number and Structural Event. For each, the first screen shot shows the traditional alignment using MOSAIK for both the reference and subject sample, the second panel shows an IGV screen shot showing the reads identified by RUFUS as variant, and the third panel shows a schematic of the contig association. Complete blast results are shown for both SNV and Insertion/Deletion. The extra IGV plot and green contig for Copy Number detection, labeled break point sequence, shows the additional reads identified as mutations novel to the duplicated sequence that span rolling circle junction created by the duplicated sequence.

Reference assisted: If a reference sequence is available, SNVs and other small variants that will map well can be identified by mapping the assembled contigs to the reference to detect regions with variation within the known genome. This has two advantages over traditional mapping methods; the assembled contigs are up to two times longer than the original reads for a SNV, increasing mapping accuracy and quality in ambiguous regions of the genome. If desired the read pair mates can be added into the assembly to increase the size of the contigs further. Secondly, filtering limits analysis to reads that contain variation, reducing false positives that occur due to incorrect mapping of reads in difficult regions of the genome. The reference-assisted method will not work well for detecting larger mutations as the contigs will differ greatly from the reference. If detection of larger variations is desired the contigs should be analyzed with the reference free method outlined above.

Figure 19: Relationship between DOC2 and RH7.5. DOC2 and RH7.5 are both derived from the RH strain. Both mutants were generated by chemical mutagenesis of their respective parent strains. Expected mutations, and the relative time of the events, are depicted by light blue arrows. The final expected mutations between RH7.5 and DOC2 are shown as a labeled red arrow.



3.5 RUFUS results

Here we show results using RUFUS compared to the mapping based approach described in Chapter 2. To test RUFUS on a biological set with a variety of variants, we compared the well characterized *Toxoplasma gondii* mutant strain F-P2 to the mutant sample EMS7.5 from a closely related laboratory strain HXGPRT-KO (Figure 19). The samples were sequenced in separate lanes on the same Illumina HiSeq run using 100 bp reads (sample detail in Table 3) as described in Farrell et al 2014 (Appendix B). A detailed list of mutations is shown in Appendix D. Mapping based methods called a total of 263 SNVs between the two strains, of these 40 have been confirmed with PCR. It is important to note that mapping based methods call an additional 3952 variants in regions of the genome that appear to represent reference errors and incorrect mappings and have thus been filtered out using methods described previously. None of the filtered SNVs tested have been confirmed by PCR indicating that they are false positives. In addition we expect 3 structural mutations between these strains; a known deletion of HXGRPT2 on chrVIII from 6,800,848 to 6,802,278, as well as 2 previously un-characterized reporter plasmid insertions in the F-P2 strain that mapping and assembly have thus far failed to identify.

3.5.1 K-mer size selection

When using RUFUS care should be taken to select an appropriate k-mer size for the genome sequenced. A very large k-mer will increase the percentage of the genome that is unique; however it will reduce the effective sequence coverage and increase the memory required. With no prior knowledge of the genome sequenced, optimal k-mer size can be estimated using the k-mer frequency histogram. Figure 20 shows a comparison of k-mer frequency histograms for the DOC2 sample with k-mers from 12 to 30. The average k-mer coverage for each copy number peak corresponds to the center of each copy number peak and decreases as k-mer size is increased. The ratio of unique to non-unique k-mers can be calculated based on the area under the single copy peak as compared to the area under all subsequent copy

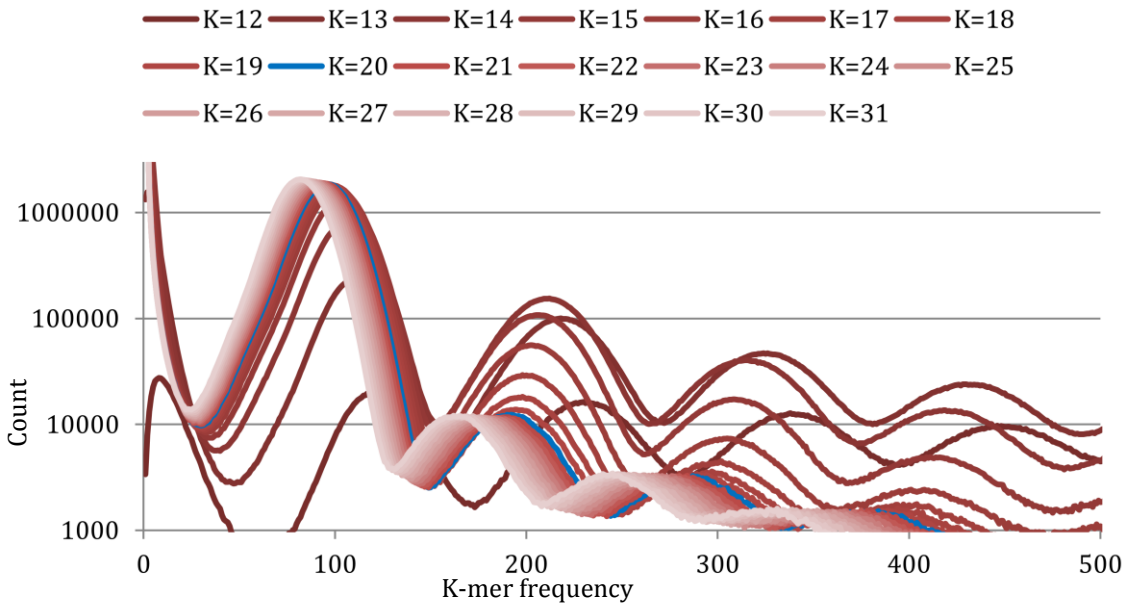


Figure 20: Comparison of k-mer size. Comparison of multiple k-mer frequency histograms for k-mer sizes from 12 to 30. As k-mer size is increased, the graph shows the trend towads greater k-mer uniqness with a corosponding decrease in coverage for each copy number peak. K=20, highlighted in blue, was selected as the optimal choice and was used for this research.

number peaks. As the k-mer size is increased this ratio will trend towards more unique k-mers. The k-mer size should be set to maximize the uniqueness of k-mers in the genome while not unnecessarily reducing the effective coverage. In this example, for k-mers larger than 20 the average coverage continues to decrease while the ratio of unique to non-unique k-mers does not continue to improve. This indicates that a k-mer size of 20 is likely the largest k-mer that should be used for this genome; larger k-mers will lower the effective coverage and will not improve the unique fraction of k-mers. A smaller k-mer size will increase the effective coverage, but regions which may be unique at a larger k-mer will now appear non-specific, and variations in these regions will now appear as copy number variations as opposed to mutations. These K-mer characteristics will remain constant across samples from the same genome and does not need to be run on every sample from the same species with the same read length.

3.5.2 Run Statistics

RUFUS was run to compare the F-P2 sample to the EMS7.5 sample using the following parameters; k-mer size of 20, prior probability of a mutation as $3.0*10^{-6}$ (or 200 SNVs/65,000,000 bp), and a prior for copy number $1*10^{-5}$. Complete run statistics are listed in Figure 21 and Table 3. In total, RUFUS completed in less than 4 hours using a maximum of 16 processors and with a maximum memory requirement of 446 Mb*. One important note when running RUFUS, as discussed section 3.2.2, the vast majority of k-mers exist as singletons, likely created by

sequencing random base substitution. These k-mers do not improve variant detection and can be ignored with no effect on specificity or sensitivity. Using a lower k-mer cutoff of 3 (eliminating counts of 1 and 2) reduces the memory needed to run RUFUS on this data to 446 Mb from 20.7 Gb with no effect on final results. We strongly recommend using a cutoff of 3 when running RUFUS on most data sets. The lower k-mer cutoff could also be used to reduce the effect of cross contamination between samples. P-value histograms from Bayesian k-mer count comparison in RUFUS.build are plotted in Figure 21c and show excellent separation between mutation events and uninformative k-mers. RUFUS.filter identified 286,780 F-P2 reads, and 26,741 EMS7.5 reads as candidates that may contain variation, reducing the number of reads that require analysis by 99.5% for F-P2 and over 99.9% for EMS7.5. The more modest reduction in reads in F-P2 is due to the higher level of human contamination in that sample, resulting in a larger proportion of human reads which by random chance were sampled at high depth and appear to be genomic. Of the F-P2 filtered reads, 52.4% align to the human genome. In contrast, 1.1% of the RH filtered reads align to the human genome. Contaminating DNA sequence will not affect mutation detection; if one uses a mapping based analysis method contamination is easily filtered out, and in the case of reference free analysis described above, the other sample will not contain a complementary sequence and will there for not produce a complementary contig. If significant contamination is expected from an organism with a known genome, contamination can be filtered out by aligning the filtered reads to the genome of the contaminating organism. Removing contamination reads will speed up the downstream analysis

by removing uninformative reads, though this will likely not affect the final results and is not required. Contigs were assembled for F-P2 and EMS7.5 using RUFUS.overlap and blastn was used to compare the libraries.

Table 3: Sample Details and RUFUS run statistics. Detailed sequencing statistics for nF-P2 and RH7.5. Mapping based substitution rate is calculated off MOSAIK alignments. Times are listed as mm:ss.00. Memory usage is listed to the left of run times for selected memory intensive steps. *RUFUS.Build-Min3 indicates RUFUS.build run with minimum k-mer count of 3, as opposed to the full Table. This does not affect the final results but greatly reduces the memory required.

Sample Details		nFP2		EMS7.5
Total reads sequenced		83,487,194		71,395,882
%Toxo		96.67		97.96
%Human		1.3		0.06
%unaligned		2.02		1.97
Mapping based substitution rate		0.003213		0.003455
RUFUS Run statistics		Time Elapsed		Time Elapsed
Jellyfish count	2.66G	03:56.0	2.66G	03:32.5
Jellyfish histo		0:24.79		0:29.51
RUFUS.model		01:43.4		00:30.9
Jellyfish dump		04:27.0		03:53.2
RUFUS.Build – full	20.7G	47:45.5	20.7G	48:45.5
*RUFUS.Build - Min3	0.446G	14:36.8	0.407G	18:16.0
RUFUS.Filter mate1	0.020G	68:18.0	0.020G	57:41.7
RUFUS.filter mate2	0.020G	61:06.0	0.020G	52:09.7
Filtered Mutant Reads		286,780		26,741
Filtered CopyMut Reads		71,023		15,113
%reduction		99.5% removed		99.9% removed
RUFUS.overlap		28:11.9		00:41.5
Contigs		573		255
RUFUS.overlap.CopyNumber		03:21.5		00:18.5
Contigs		22		16

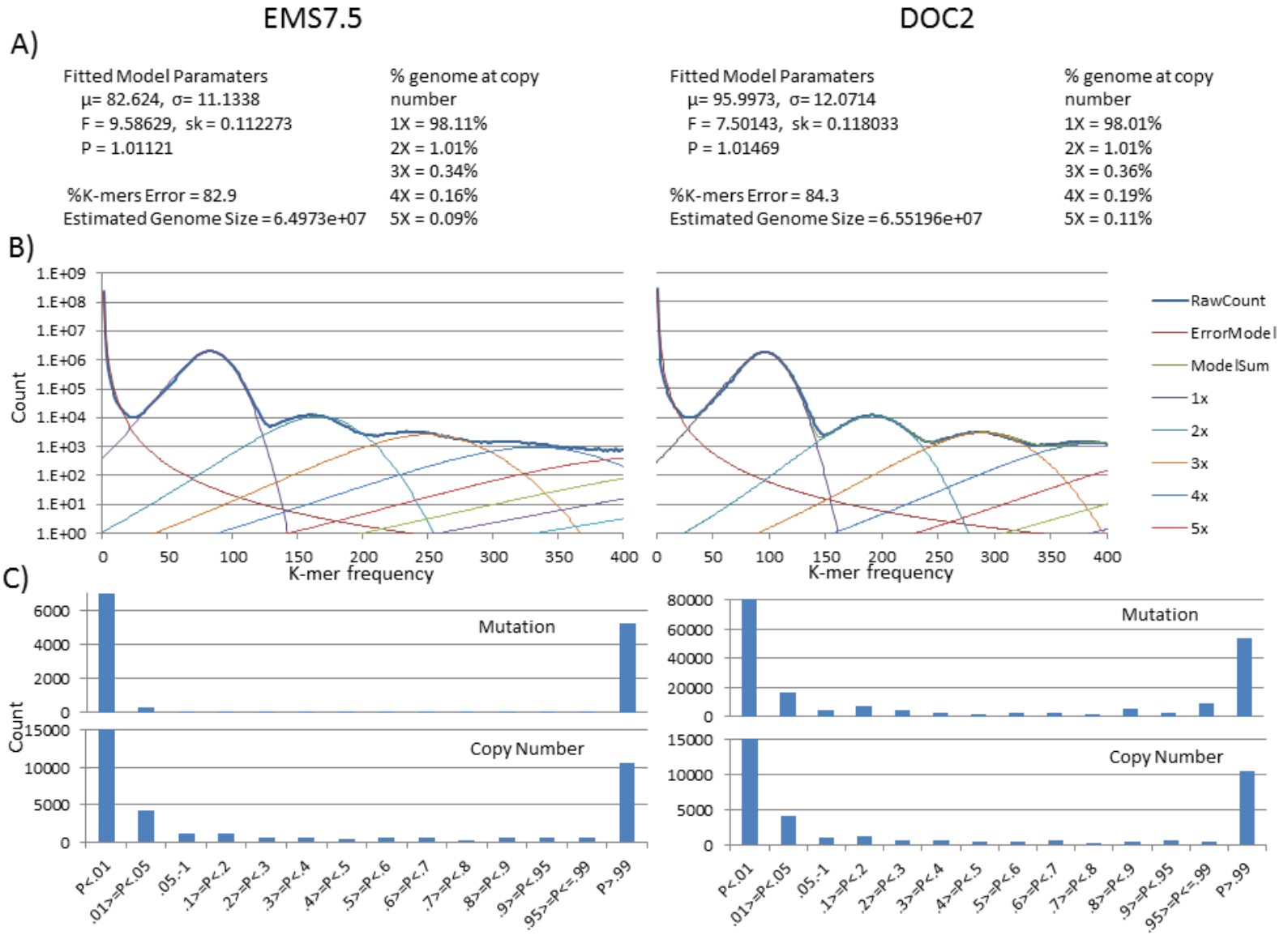
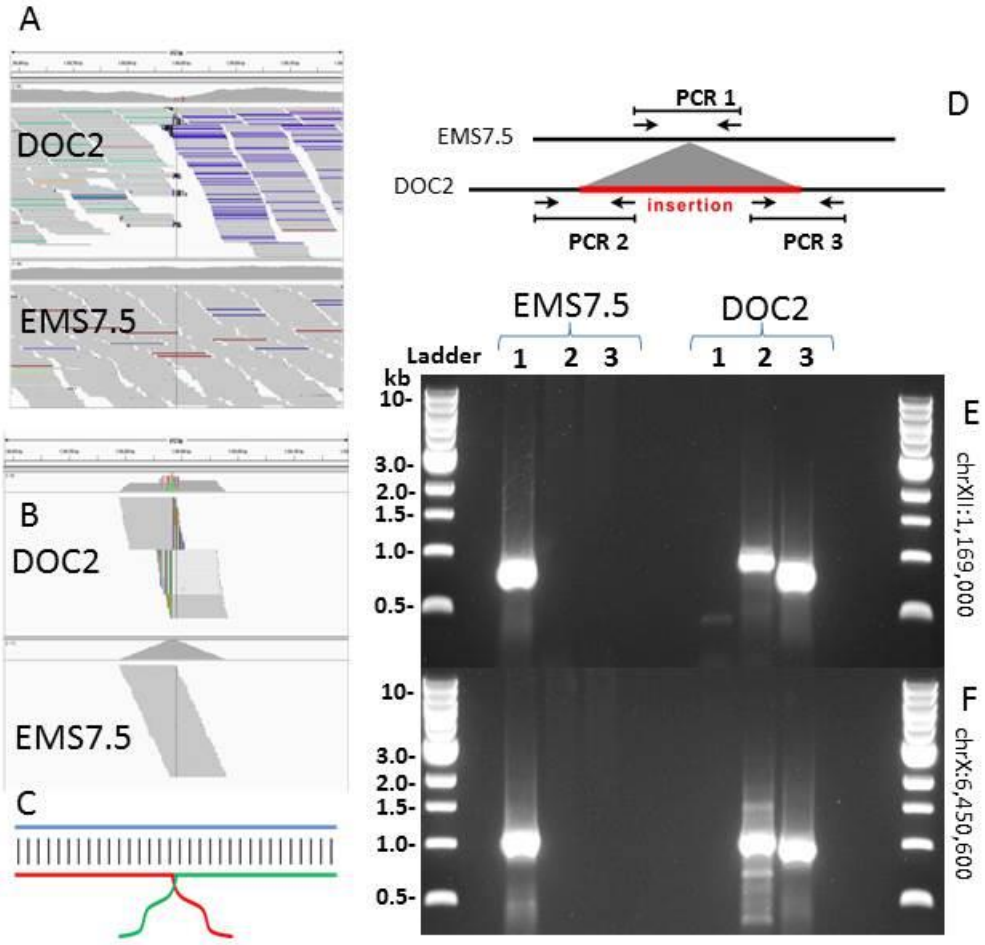


Figure 21: **Complete Run Report for EMS7.5 vs DOC2 RUFUS Run.** A, output by RUFUS.model for each sample showing the fit parameters for the final model, as well as genome predictions including the calculated genome size and proportion of genome that exists in each copy number state from 1 to 5. B, the fit model for each sample using the parameters shown in A. C, distribution of probabilities for both mutation events and copy number events as reported by RUFUS.

3.5.3 Insertion/deletion detection

In contrast to mapping methods that must first align reads to the reference, creating bias towards the reference and hindering detection of indels, RUFUS identifies variant reads by their dissimilarity, increasing the likelihood of identifying variations the more drastically they differ between the samples. This is an advantage over traditional methods as it removes bias towards the reference and increases the likelihood of identifying variants, in particular variants which differ drastically between samples making RUFUS ideally suited for detecting insertion/deletion events and structural events. RUFUS identified all of the expected structural events between F-P2 and EMS7.5. Of particular interest are the two plasmid insertions that have eluded paired end mapping analysis methods due to the plasmids high similarity with the *T. gondii* genome; gra1LacZ contains 1,599 bp out of 5.5 kb, and tubYFP contains 3,380 bp of 9.9 kb that are a perfect match to the T. *gondii* reference at a k-mer size of 100 bp. This significant homology with the reference, as well as the poor quality of the draft *T. gondii* genome, caused paired end methods to detect a staggeringly high number of false positives (over 600) making identification of the true events impossible. However, RUFUS accurately identified the breakpoints where the insertions occurred without any false positive calls; these have been confirmed by PCR (Figure 22). For the HXGRPT2 deletion RUFUS assembles the entire 1.6 kb deleted section and correctly identifies the break point as a single 186 bp contig in EMS7.5 whose ends perfectly match the 1.6 kb contig in F-P2 section in the other sample (Figure 18).

Figure 22: Plasmid Insertion Detection and Confirmation. (A) IGV screen shot showing and example of one of the two plasmid insertion locations in the original MOSAIK alignment. (B) IGV screen shot showing alignment of the individual reads identified by RUFUS as variations. (C) Schematic of the structure identified by BLASTN. (D) Schematic of the primer design used to confirm the insertions. 3 primer pairs were designed at each location, a pair flanking the entire region which only amplifies in the absences of the insertion in EMS7.5 (PCR1), shown in E and F. PCR2 and PCR3 use the same forward and reverse used in PCR1 but each uses a new primer complementary to insert sequence. PCR2 and PCR3 only amplify in F-P2.



3.5.4 Copy number detection

By utilizing the k-mer frequency plot, and the corresponding model produced by RUFUS.model, RUFUS is capable of making accurate copy number predictions that take into account both the copy number of the sample as well as the variability of coverage within that sequence, possibly reducing the rate of false positives. As a positive control for copy number events, there are 5 regions from the inserted plasmids in DOC2 that are exact copies of the *T. gondii* genome (discussed in section 3.4.3). These regions present as copy number events between F-P2 and EMS7.5 and can be used as a bench mark for copy number detection. These regions range in size from 425 bp to 2,842 bp and total 5811 bp, giving a per base prior probability of roughly $1*10^{-5}$. Using this prior and a minimum contig length of 400 bp, RUFUS identified all of these regions accurately, with zero false positives. RUFUS additionally detected a novel 5 kb event on chromosome VIIa that we were not previously aware of. This event shows one feature that will be the hallmark of true copy number event; as with any structural event, a novel breakpoint should be created at the ends of the duplicate region linking them together. RUFUS identifies such a break point for this event that matches both ends of duplicated region, indicating a rolling circular structure (Figure 18). Interestingly, this event shows up in every sequencing run we have for F-P2, however it does not show up in any of the other strains, indicating that this duplication occurred during ENU treatment.

Removing the length restriction introduces 14 additional copy number calls, 10 of which do appear to be simple false positives caused by short local variations in

coverage. The other 4 however call SNVs that exist in DNA present in multiple copies in both samples, and are effectively heterozygous. These can be called by using blast to compare the unmatched mutation calls from each sample with the unmatched copy number calls from the other sample. True heterozygous variants will show as a match between an unmatched mutation contig and a copy number event, copy number calls due to random coverage fluctuations will not find a match with a mutant contig and will not cause a false positive. By extending RUFUS.model to diploid genomes, this will allow the detection of heterozygous events in diploid organism, this is currently being tested in human samples with promising results.

3.5.5 Variation detection in unmappable genomic regions

Possibly the greatest advantage RUFUS has over mapping based methods is that it is not limited a reference genome, allowing RUFUS to identify variants in completely novel DNA sequences, as well as highly repetitive regions. RUFUS finds 8 SNVs between these samples that mapping methods could not detect. Two of these calls are in sequence missing from the GT1 reference but present in other *T. gondii* reference genomes and have been confirmed by PCR. One variant is in a novel stretch of DNA that does not match the *T. gondii* genome or any sequence in the NCBI nucleotide database and has been confirmed by PCR indicating that it is not simply an assembly artifact. The final 5 variants are in low complexity regions, often differing by only 1 or 2 bp per 100 bp from dozens of other sites in the *T. gondii* genome. One of these was successfully confirmed by PCR; however the other

4 would not produce clean PCR reactions likely due to their high similarity to other regions. We do believe they are true variants but with current technologies these cannot be confirmed at this time.

3.5.6 SNV detection

In addition to drastic changes to DNA sequence, RUFUS detects single nucleotide variants (SNV) with greater precision than the mapping based methods employed in our previous work, calling 0 of the almost 4 thousand false positives that required filtering. RUFUS is ideally suited for the random variations induced by mutagenesis, successfully calling all of the variants induced by mutagenesis; the 29 ENU induced variants in DOC2 and the 39 EMS induced mutations in RH7.5. The remaining 194 SNV calls can be split into two groups, 163 that represent a true SNV, and 31 that represent expansions or contractions of microsatellite regions. Of the 163 true SNVs, RUFUS identified all of the variations using the reference added method outlined above. Using the completely Reference free method RUFUS called 161 SNVs, the remaining 2 depict an interesting rare situation where, at the given k-mer size, a mutation creates k-mers that matches a different region of the genome. In these instances a mutant contig is identified in one sample as a mutation and either missed in the second sample or detected as a copy number event and thus classified as a heterozygous call. Spontaneous mutations that have arisen during the strains separate passage in the lab are enriched for expansions and contractions of microsatellite regions. While RUFUS will detect SNVs within highly repetitive

regions, RUFUS will not detect expansions and contractions in microsatellites longer than the k-mer length as these regions will produce no unique k-mer. 31 of the 194 spontaneous SNVs are expansion/contractions in microsatellites longer than the k-mer size and, as expected, RUFUS did miss all of these. Some of these expansion/contraction events are recovered by running RUFUS with a larger k-mer size, however read length and k-mer size will limit the upper bound of microsatellite length polymorphisms.

3.5.7 Comparison with NIKS

NIKS is a similar k-mer based detection method, simultaneously developed at the Max Plank Institute in Germany[45]. NIKS was specifically designed to find homozygous mutations in mutagenized rice strains identified with forward genetic screens. It uses a similar k-mer comparison method as RUFUS to identify k-mers that represent novel sequences between two samples. NIKS detection is based on static cutoffs, where as RUFUS use a Bayesian detection method, thus NIKS will only detect homozygous/haploid mutations between two samples. NIKS is not capable of detecting any of the copy number variations or heterozygous variations detected by RUFUS.

To compare the sensitivity when detecting SNVs, NIKS was run with similar settings as RUFUS, k-mer size of 20, to compare the F-P2 mutant and EMS7.5. NIKS completed in 5.5 hours using 10 processors, more than double the time require for

RUFUS. Of the 226 SNV calls where RUFUS and Freebayes agree, NIKS correctly called 219 SNVs, identified a variant contig in both samples. There are two sources of false negatives shared by both of these methods; firstly if the k-mer detection method does not identify a difference between the samples for a given k-mer that represents a mutation, the method will not identify any reads that span the mutation. Secondly, the assembly method may fail to assemble a contig, which will also result in a false negative in that region. Of the 7 missed called, one of them NIKS did not identify any of the reads that span the mutation. The remaining 6 false negatives, NIKS did correctly identify the reads that spanned the mutation in both samples, however one of the contigs in either sample was lost in the Velvet assembly. For the microsatellite expansion/contraction events that RUFUS missed, NIKS too missed every one of these. With regards to the 8 novel SNVs detected by RUFUS NIKS identifies 6, missing 2. It is possible that a more sensitive assembler may improve detection by reducing assembly error. If we use RUFUS.overlap to assemble the reads identified by NIKS, 4 of the missed calls are recovered leading to a total score of 232 out of 234 RUFUS calls.

NIKS did correctly exclude all of the false positives that mapping based methods introduced. NIKS does produce a single unique call, however when we assemble with RUFUS.overlap this call is removed indicating that it likely is an assembly artifact associated with Velvet, further research will be required to confirm this. This indicates that NIKS demonstrates the same increased specificity as RUFUS with similar, yet slightly lower, sensitivity.

3.6 RUFUS conclusions

The genetics community is desperately in need of tools such as RUFUS and NIKS. To date, sequence analysis has not evolved significantly past the days of microarray analysis; currently researchers must have prior knowledge of the sequence they are investigating despite the fact that next generation sequencing technologies are capable of sequencing any DNA, regardless of its genome context or sequence. To truly take full advantage of the freedom and power these new sequencing technologies offer, we desperately need new tools that are both efficient and not limited to the known genome.

RUFUS offers completely reference free variant detection with the ability to detect all of the variation that mapping based methods are currently capable of. RUFUS analyzes the k-mer count distribution between two samples to create an accurate model of read coverage across the genome. This model can then be leveraged to detect SNVs, Insertions and deletions, and copy number variants between a pair of Illumina sequenced samples. RUFUS does not require the massive computing power, or multiple sequencing libraries, which are required by whole genome assembly methods[23,25]. This makes RUFUS particularly useful for researchers working on organisms for which there is no reference or a very poor reference. Additionally by removing the reference from mutation detection, reference bias is eliminated improving detection in 3 ways; Variations that occur in

DNA that is either missing from the reference or is unmappable can now be discovered, improved the detection of INDELLs as well as improving the detection of rare variations due to the lack of bias, and a massive improvement in specificity as mapping error has been completely eliminated which contributes the majority of false positive calls in mapping based method.

Chapter 4:

Concluding remarks and future applications with RUFUS

Second generation sequencing technologies, including the Illumina, will indiscriminately sequence any DNA sequence. This appears to be a trivial statement; however it makes these technologies immensely powerful. However, the current state of the art analysis methods, reference guided alignment or mapping, comply ignore this ability and limit the power of these new sequencers. Prior to second generation sequencing, Micro Array analysis was the state of the art method for cheap and efficient genetic analysis, and is still widely used today. In this method oligos are bound to an array, generally derived from assembled genomes, and the presence of that oligo can be identified in a sample through hybridization. This allows you to quickly identify known sequences in a given sample. This limits detection to previously known genomes and mutations. Next generation sequencing methods were supposed to free us from these limitations by allowing cheap and massively high throughput sequencing, allowing us to detect all mutations regardless of previous discovery. However, mapping methods have limited this promise, barely advancing us beyond microarrays. Methods that are independent of a reference can take full advantage of second generation sequencings abilities. However, until now the only methods for this were based on whole genome

assembly, which is far too costly and produces assemblies that are noisy and not suited to mutation detection.

RUFUS offers the ability to extend the scope of next generation sequencing studies beyond what was previously possible. Its unique ability to detect all variant types without a reference will enable whole genome studies in any organism with little or no prior work. Further, by removing the reference and the associated reference bias, RUFUS is capable of detecting a wider array of mutations, including variations in repetitive sequence and increase sensitivity towards insertions and deletions. Up to now we have only considered RUFUS as a tool for mutational profiling, however we believe that RUFUS is applicable to a much broader range of applications and its unique abilities will make it invaluable in future research. We will conclude by covering some of the current applications being explored as well as preliminary data where available.

4.1 Projects currently in development

4.1.1 Human Trio Analysis

With the 1000 Genomes project identifying approximately 99% of common SNPs in the human population, focus has shifted to understanding rare germline mutations and their role in human evolution and disease. Current mapping based methods are designed to detect common variation, at a rate of 10^{-3} or 3 million

events per individual. Typical false discovery rates (FDR) range from 10^{-5} to 10^{-4} per nucleotide, i.e. 30-300 thousand erroneous calls per individual. When using these methods to identify all variation in a human sample, on the order of 3 million variations, this FDR is reasonable and has little effect on the final results. However, research suggests that germline mutations occur a rate of 10^{-8} [48] and thus will be completely drowned out by the current FDR. The errors and misalignments caused by reference guided alignment approaches are largely caused by differences between the reference and the samples (for instance, hidden CNV states in the family, missing contigs, etc.). Removing the reference from the process will eliminate such errors and make detection of putative de novo events possible.

We used RUFUS to identify novel alleles in the child that were not inherited from the parents. The parents samples were pooled, and used as a reference sample in RUFUS to identify novel variants in the child. These reads were then assembled using RUFUS.overlap and aligned to the human genome to identify variations. For this work we used one of the available Illumina Platinum Genome trio data sets; child NA12882, father NA12878 and mother NA12877[49]. We compared the RUFUS calls to the calls generated by GATK that Illumina provides with the data set, as well as ran our own analysis using FREEBAYES, Figure 23. After identifying polymorphisms with genotypes unique to the child, GATK calls 215,270 (q>=100) germline events, consistent with a 3.18% polymorphism FDR, but orders of magnitude more than expected from novel mutation. FREEBAYES called far fewer events, 119,600. The concordance rate between these sets was extremely low with less than 10% of the calls present in any of the other two sets, indicating that the

vast majority of these reads may be false positives. RUFUS calls fewer events, 3,189, with over 93% calls present in one of the other two call sets, and 72% called in both the GATK and FREEBAYES call sets. This suggests that our reference-free method is able to substantially cut down on false positive error, and achieve far higher specificity to rare events. 3,189 are far more events than the approximately 100 that would be expected based on natural variation. We believe this is due to the fact that this data set was amplified for sequencing using cell lines, which may have introduced additional variation into the data that is not present in the original samples.

168 calls were completely unique to RUFUS, Figure 24. These calls fall into 2 categories; correct calls in complex variants shows in Figure 26, or calls in regions where traditional mapping cannot confidently place reads Figure 26. There are an additional 110 variant contigs which align to alternate contigs of the human genome according to NCBI blast, regions which are currently ignored in most alignment based methods. Finally, there are 92 calls that do not align to any known DNA sequence using NCBI blast, demonstrating the ability of our reference-free method to call variants in regions not represented in the current human reference which are completely ignored by other methods.

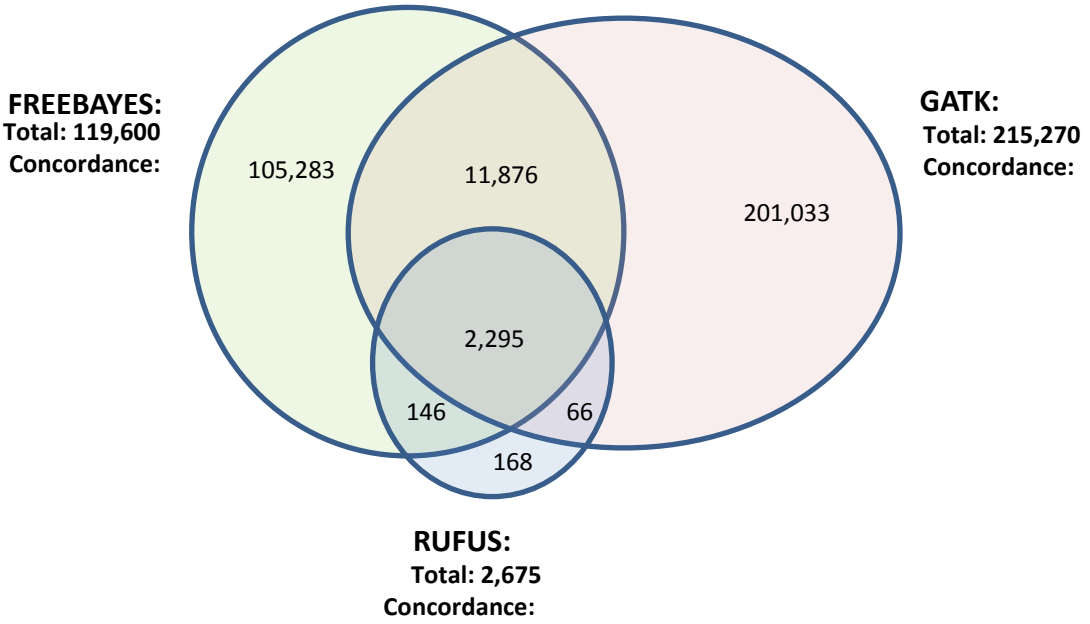


Figure 23: Analysis of SNV calls. Comparison of calls between RUUFS, FREEBAYES, and GATK for novel calls in the child sample NA12882 vs the parents NA12877 and NA12878. Numbers indicate SNP call counts as indicated

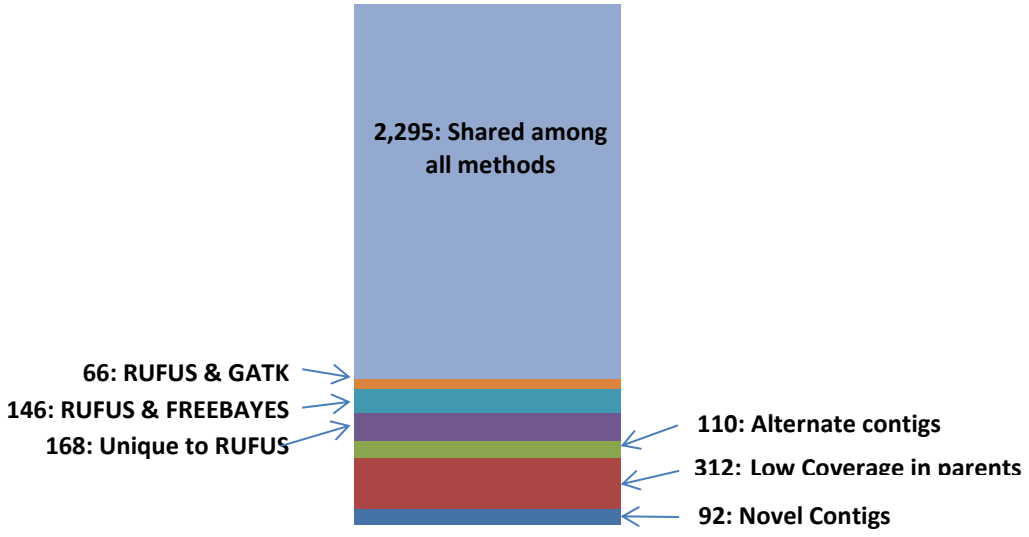
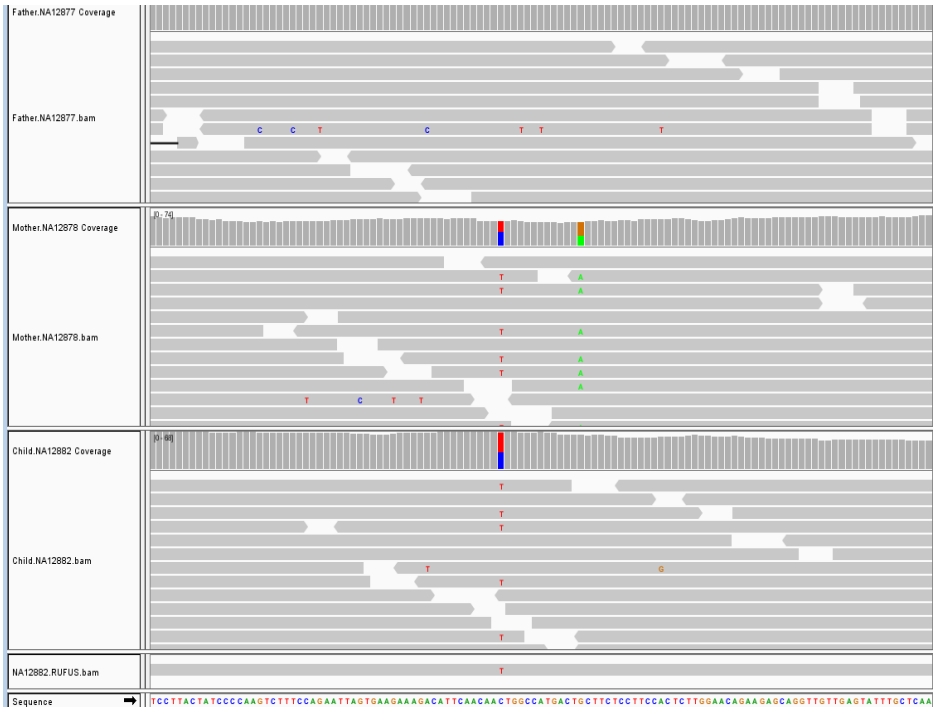


Figure 24: Full Breakdown of RUFUS calls. RUFUS called a total of 3,189 variations unique to NA12882. These calls can be separated into 7 groups outlined here.

Figure 25: Examples of missed complex events. Many of the unique RUFUS are in complex events or linked events where mapping based methods are incorrectly classifying the events, excluding them from the final set of variants unique to the child. A shows an instance where the child has the same T allele as the father, yet in the father that T is always followed by an A, missing from the child making this a new haplotype. B shows an instance where the child is heterozygote for a variant present in both parents, however the child has lost a T upstream snp present in the fathers reference allele.



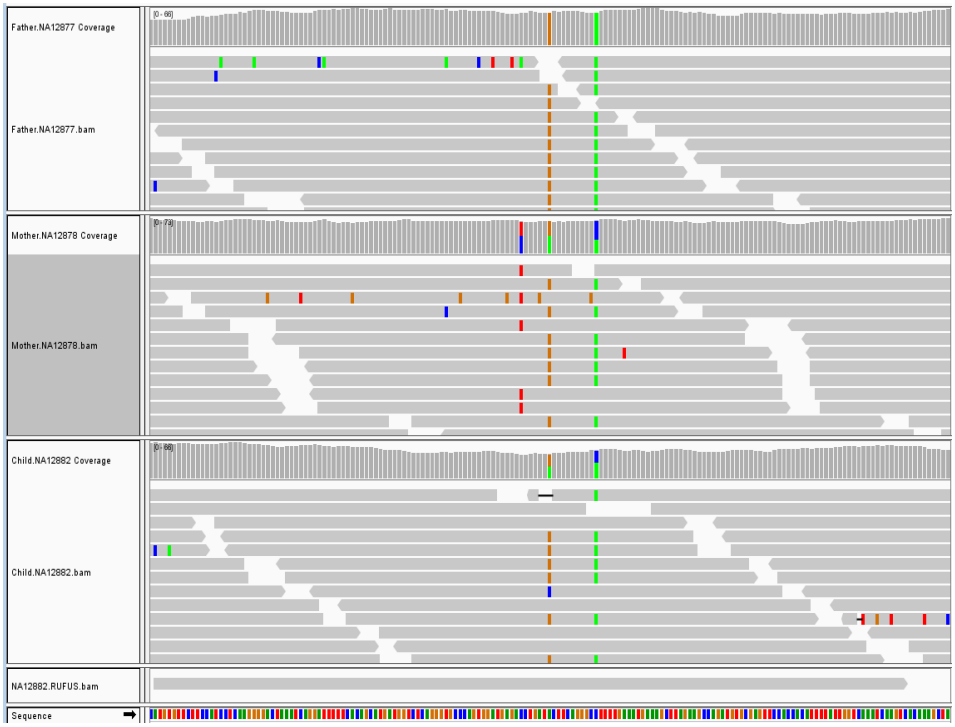
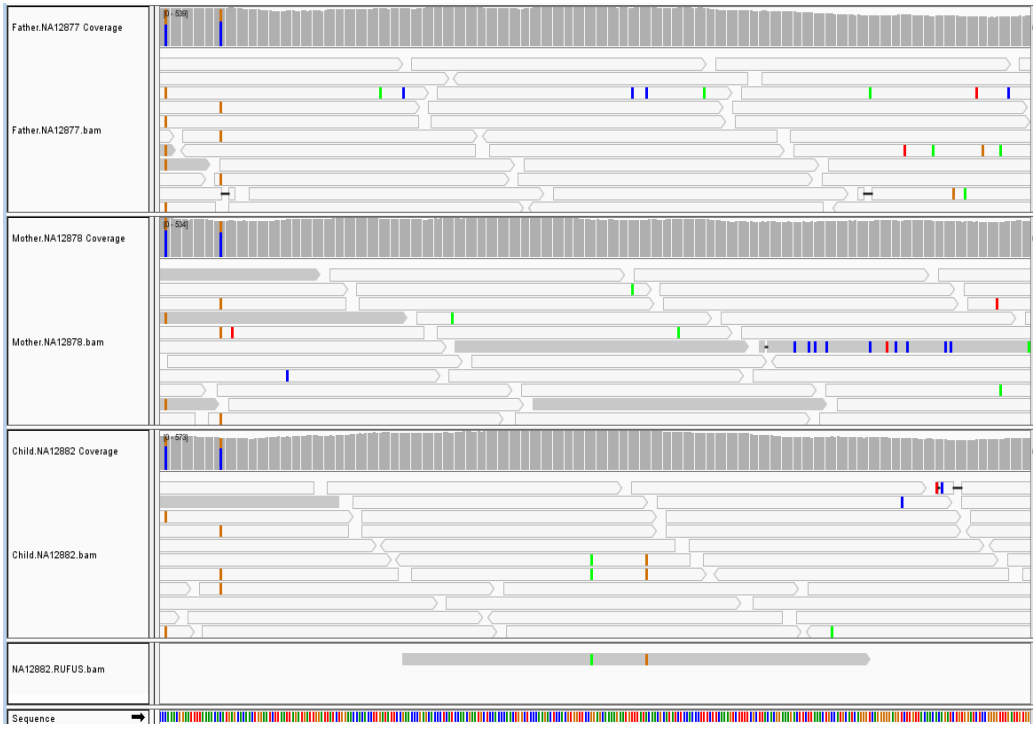
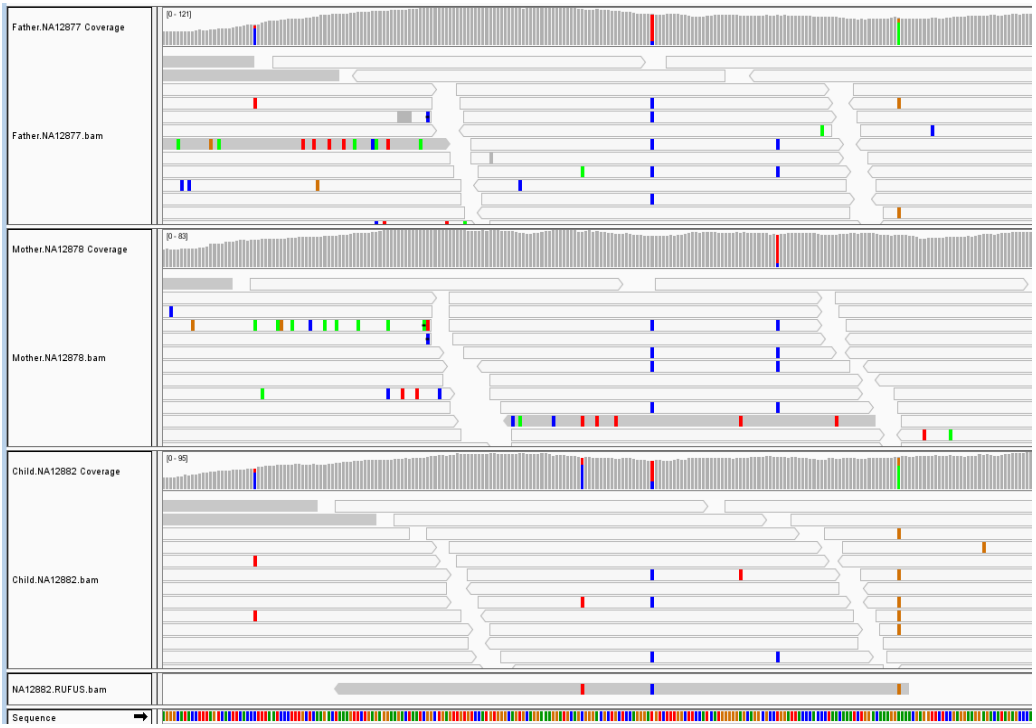


Figure 26: Improved Mapping Quality. RUFUS is additionally capable of identifying variants in regions where aligning short reads result in low mapping qualities. IGV color codes reads with mapping quality 0 as whit contigs, gray contigs indicate a high mapping quality. RUFUS pre assembles the reads allowing alignment of a longer contig with can improve mapping qualities and recover these areas. In both A and B, novel variants exist in the child, yet the mapping qualities are 0, so these regions are lost as reads with 0 mapping quality are ignore in variant calling. The longer RUFUS contigs (last line in each image) however aligns with a high mapping quality and correctly identifies the variation.





4.1.2 Bayesian RUFUS in diploid organisms (Human)

The above work analyzing a human trio was done with the simplest form of RUFUS that looks for novel k-mers based on static cutoffs, not the Bayesian version described in Chapter 3. In this simpler method, a cutoff is set for the subject sample; in the trio above we used 5. Any k-mer above 5 will be considered present in the genome of the subject (child), if the same k-mer is either absent from the reference sample (combined parents), or its count is below a specified cutoff (3), it is considered unique to the subject and identified as a mutation. Read filtering in then performed exactly as described in 3.4.4. This method identifies sequence mutations, such as SNPs, insertions, and deletions, extremely well with the similar sensitivity and specificity of the Bayesian RUFUS method. However in order to detect copy number events, and heterozygous mutations in human samples, we will need to apply the Bayesian RUFUS detection method. To do that we must be able to model the underlying copy number distributions for diploid samples.

Using the human reference to split up a human samples k-mer histogram shows that it follows the same pattern of overlapping copy number distributions as in *T. gondii* (Figure 28). A unique feature of a diploid genome however is the small bump that occurs before the single copy DNA peak. This bump is caused by haploid sequences such as heterozygous variants and the Y chromosome. By accounting for this extra distribution in RUFUS.model (Figure 27), we can add heterozygous detection in diploid samples to the Bayesian variant detection method.

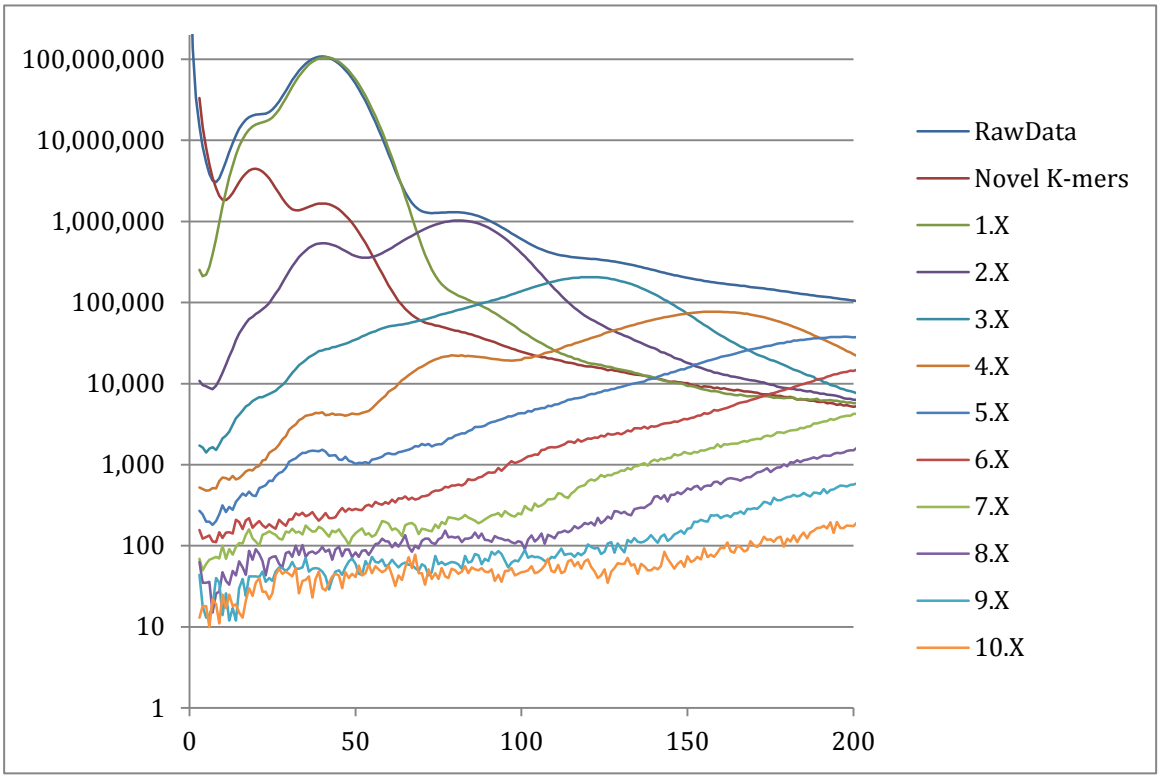


Figure 28: Reference Separated K-mer Histogram of NA12882: K-mer histogram of NA12882 with each k-mer binned based on its occurrence in the human reference genome hg19. Novel K-mers indicate k-mers absent from the genome

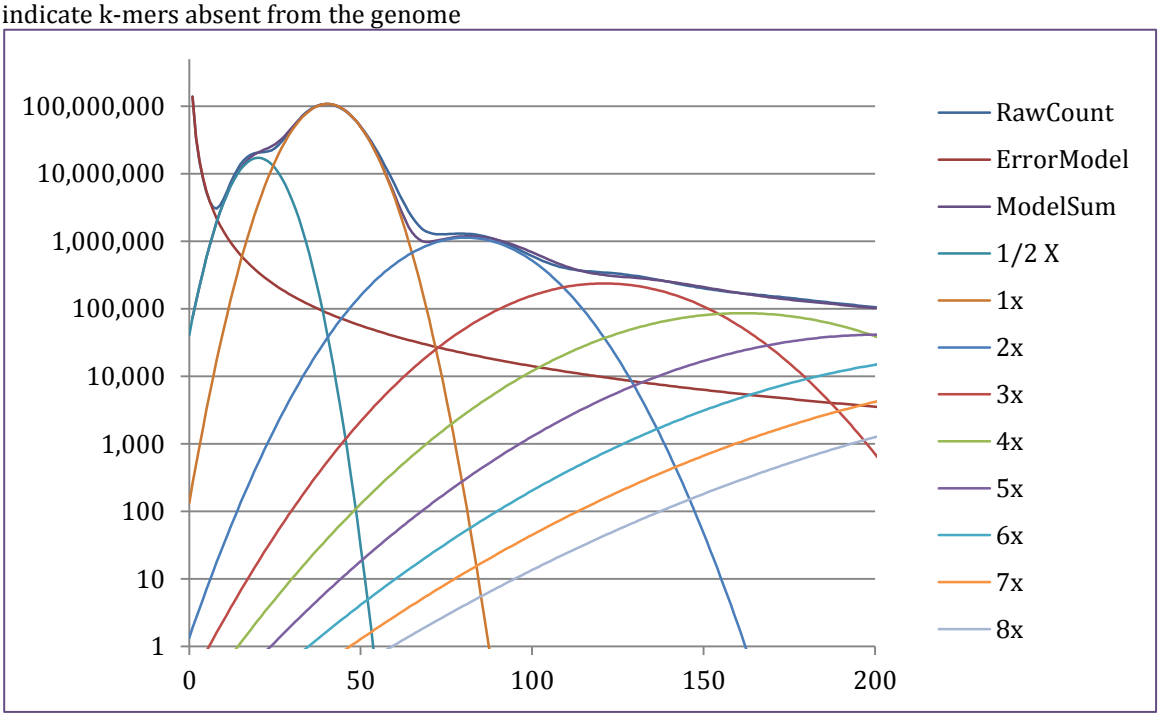


Figure 27: Heterozygous RUFUS.model. Model produced by RUFUS for human sample NA12882 modified to account for diploid samples.

4.1.3 RUFUS for population based analysis

All of the RUFUS examples to this point have been very specific and controlled experiments where two sets of reads are compared from individual closely related samples, i.e. parent and child. This is done to ensure that RUFUS only identifies variations that are true mutations between the samples and not due to population polymorphisms or differences in experimental procedures. This limits the application of RUFUS to a very limited range of experiments. There are many instances in human genetics where a closely related sample is not available due either to experimental limitations or the cost of segueing. Also, there are many instances where the researcher wants to know if a given variant is novel in a population, not simply in a close family trio.

It would be useful to use a static reference, such as human hg19. However, RUFUS should never be used to compare a sequenced sample to a standard reference for two reasons. Firstly, when comparing raw sequenced samples directly to each other, all systematic contamination (such as sequencing adapters) and error associated with sequencers is present in both samples, and thus removed by RUFUS. If you were to compare a sequence sample to the reference, all of these reads would be identified as novel and will drown out possible mutations. Secondly, we have already shown the major advantage of RUFUS is that it eliminates false discovery due to errors between the reference and the sample. A static reference does not account for alternate assemblies, which will not account for all of the natural structural variations in a population. Using the reference as one of the samples in

RUFUS would reintroduce those errors and negate any advantages gained by RUFUS.

Instead of using the human reference, we can instead use the entire 1000 genomes project as reference set, with almost no increase in time or memory required to analyze a sample. We are in the process of creating a reference hash Table for RUFUS that includes the original Illumina sequence for every single sample included in the 1000 genomes study. By hashing the raw data, this set will include every observed DNA sequence in these samples. This will account for all alternate contigs, all genomic structures, and all variations, regardless of whether or not these sequences can be assembled or aligned to the human. This will allow a researcher to take a new sample and simply compare it to this reference set to quickly identify any mutations that have are completely novel to this sample. Additionally this will allow detection of variations in regions of the genome, which up till now, have been completely ignored. This could allow detection of novel variations that explain numerous diseases that have eluded genetic research.

This 1000G reference data set will not be appreciably larger than the k-mer for a single human sample, and thus will not take longer to run than the standard two sample human analysis. This set will also not suffer from the n+1 problem associated with standard variant call sets. The n+1 problem is caused when single samples calls are added to a large data set, such as the 1000G. Bayesian SNP calling uses information from every sample when calculating variant probabilities. In order to add a newly sequence sample, you must re-call the entire data set from scratch. With a hash based method, including additional samples can be accomplished by

merging the counts with the current set, a greatly less computational expensive calculation.

4.1.4 Genomes with non-standard GC content.

Working with extremely AT rich genomes poses a challenge for all sequencing methods but RUFUS may offer a solution. Working with *Plasmodium berghei* samples we noted that RUFUS.model does not do a good job predicting the distribution past the single copy curve (Figure 29). We believe this is due to the AT rich nature of the genome. If we look at the AT distribution for *Toxoplasma gondii* (Figure 30A) the relative ratio of GC content remain constant though all copy number peaks. There is a slight associtaion between GC content and coverage, as would be expected due to the well known assocation between GC and the

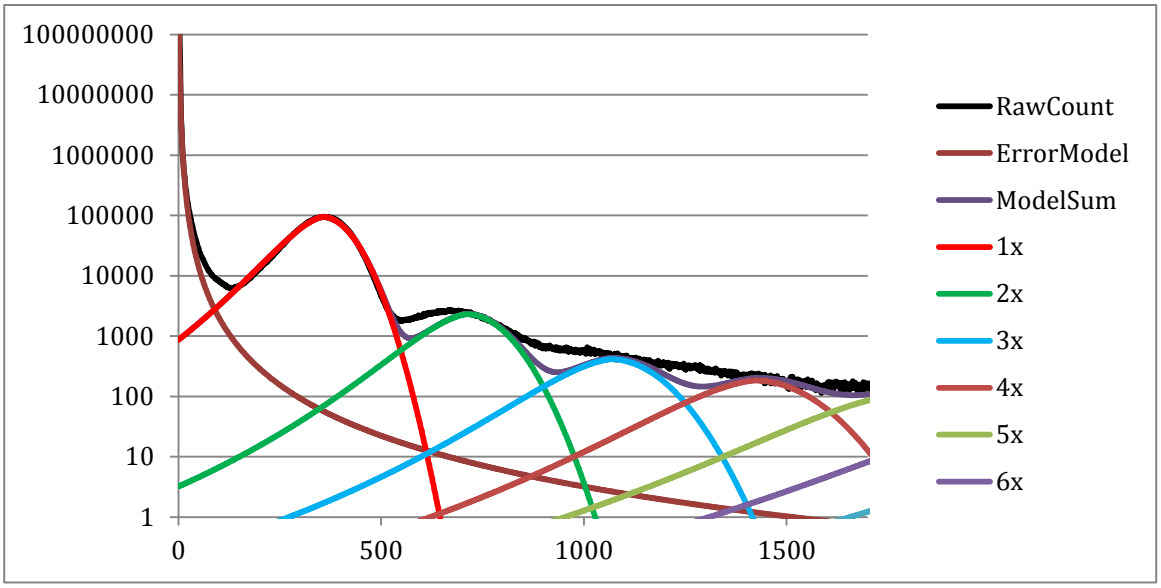


Figure 29:Model prediction for *Plasmodium berghei*. Predicted model by RUFUS.model for a *Plasmodium berghei* sample. RUFUS does not do a good job modeling the distributions beyond 1x.

amplification speed of DNA polymerase. Despite this, the variations in coverage for difference GC percentages are negated by the fact that the GC bias remains constant across the copy number peaks in T. gondii. However, in *Plasmodium berghei* the GC content is not consistent across the copy number peaks. There is a bias towards extremely AT rich k-mers in when the copy number shifts above unique sequence (Figure 30B). This causes the center of the double copy number peak to shift towards lower coverage, breaking the assumptions made in RUFUS.model that each copy number is simply a multiple of the single copy peak. We can account for this effect by modifying RUFUS to bin k-mers based on their GC content when the k-mer tables are created. We can then model each distribution separately, and account for this effect. This may allow us to improve our call confidence in genomes with such biases.

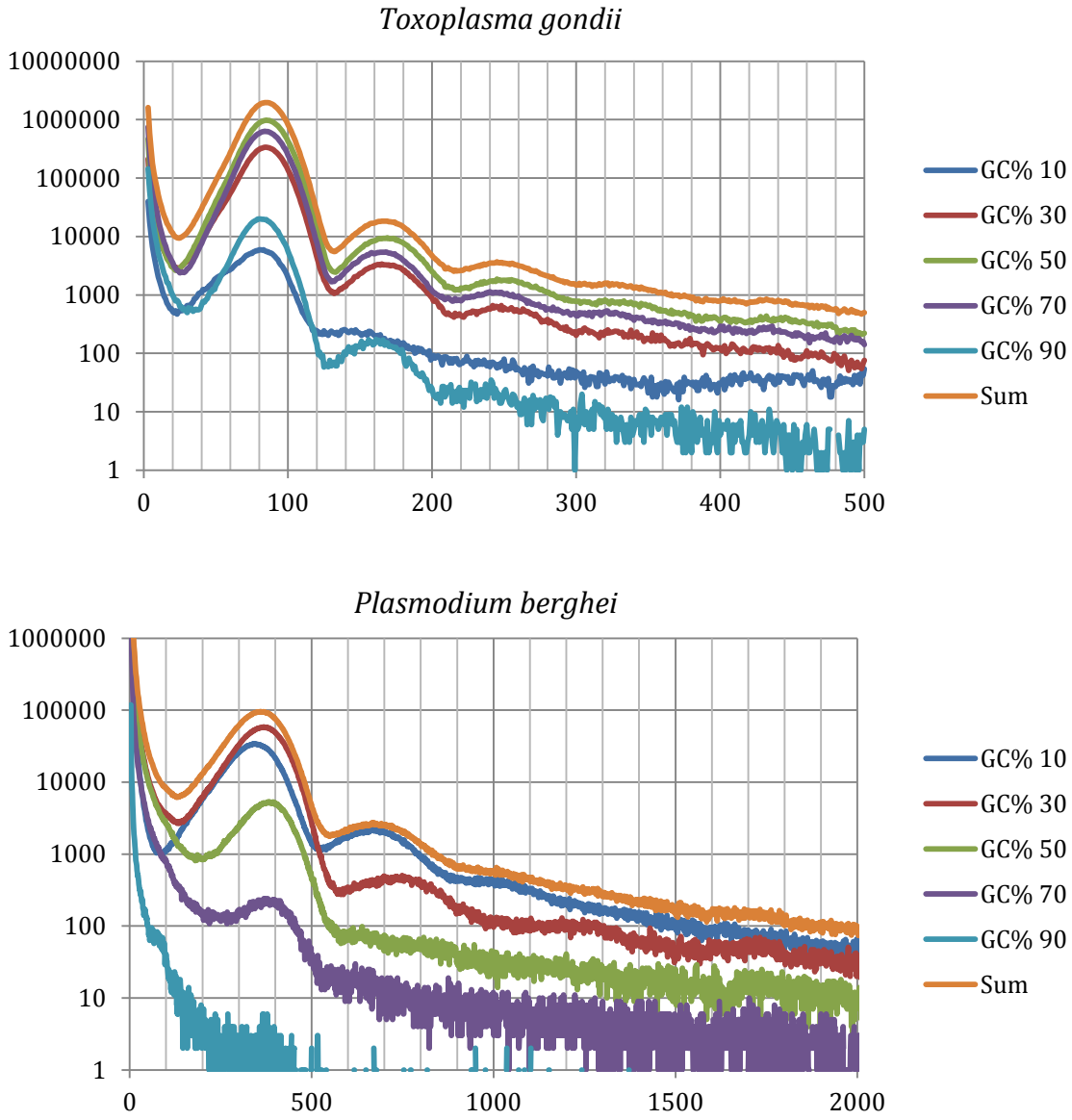


Figure 30: K-mer Histograms divided by GC content. A and B represent the k-mer histograms for two samples, with the k-mers binned based on their GC content. The GC contents listed are +- 10 for each number, thus 10 includes all GC percentages from 0 to 20, 30 includes all from 21 to 40 and so on.

Appendix A:

"A DOC2 Protein Identified by Mutational Profiling is Essential for Apicomplexan Parasite Exocytosis"

This is the first published work that used the mutational profiling pipeline described in Chapter 2. This project was used to both develop and test the pipeline.

Appendix B:

"Whole genome profiling of spontaneous and chemically induced mutations in *Toxoplasma gondii*"

This work outlines our experience in the lab working with 15 sequenced *Toxoplasma* strains. It covers both the experimental procedures used in the forward screens and sequencing as well as findings based on the results from the mutational profiling pipeline outlined in Chapter 2.

Appendix C:

Toxoplasma gondii BLASTN Hits for Contigs Assembled from Reads Unaligned in *Toxoplasma gondii* Reference Guided Alignment

Spread sheet showing the top BLASTN alignment hit against all *Toxoplasma* reference sequences for each of the assembled contigs from the unaligned reads over 1kb. Contigs were assembled from the unaligned reads in the *Toxoplasma gondii* sample nF-P2 to the GT1 reference sequence 7.0 and Human reference gh19 build 37.

Appendix D:

Complete list of variants identified in the F-P2 vs EMS7.5 comparison. RUFUS contig names are listed in the first two columns according to their names assigned by RUFUS.overlap. FREEBAYES SNPs are grouped by color based on their likely source, listed in the key on the top right of the first page. Lines with multiple FREEBAYES calls indicate regions where the RUFUS contigs overlap more than one event called by FREEBAYES.

1. Morgan TH. What are "factors" in mendelian explanations? *American Breeders Association Reports*. 1909(5):365-368.

2. Kohler RE. *Lords of the fly : Drosophila genetics and the experimental life.* Chicago: University of Chicago Press; 1994.

3. Morgan TH. Sex limited inheritance in drosophila. *Science*. 1910;32:120-122.

4. Morgan TH. Random segregation versus coupling in mendelian inheritance. *Science*. 1911;34:384.

5. Sturtevant AH. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of associations. *Journal of Experimental Zoology*. 1913;14:43-59.

6. Boothroyd JC. Genetic and biochemical analysis of development in toxoplasma gondii. *Philos Trans R Soc Lond B Biol Sci*. 1997(352):1347-1354.

7. Montoya JG, Liesenfeld O. Toxoplasmosis. *Lancet*. 2004;363(9425):1965-1976.

8. Liesenfeld O, Dunay IR, Erb KJ. Infection with toxoplasma gondii reduces established and developing Th2 responses induced by nippostrongylus brasiliensis infection. *Infect Immun*. 2004;72(7):3812-3822.

9. Jones JLea. Toxoplasma gondii infection in the united states: Seroprevalence and risk factors. *Epidemiol*. 2001;154:357-365.

10. Barsoum RS. Parasitic infections in organ transplantation. *Exp Clin Transplant*. 2004(2):258-267.

11. Luft BJ, Remington JS. Toxoplasmic encephalitis in AIDS. *Clin Infect Dis*. 1992;15:211-222.

12. Remington JS, McLeod R, Desmonts G. *Toxoplasmosis. in Infectious diseases of the fetus and newborn infant* . Philadelphia: Saunders; 1995.

13. Sahasrabudhe NS, Jadhav MV, Deshmukh SD, Holla VV. Pathology of toxoplasma myocarditis in acquired immunodeficiency syndrome. *Indian J Pathol Microbiol*. 2003;46(4):649-651.

14. Weiss, Louis M.,editor of compilation, Kim K,editor of compilation. *Toxoplasma gondii : The model apicomplexan - perspectives and methods.* Second edition.. ed. ; 2014.

15. Torrey EF, Yolken RH. *Toxoplasma gondii* and schizophrenia. *Emerg Infect Dis*. 2003;9(11).

16. Carruthers VB, Suzuki Y. Effects of toxoplasma gondii infection on the brain. *Schizophrenia Bulletin*. 2007;33(3):745-751.

17. Bosch-Driessen LHea. A prospective, randomized trial of pyrimethamine and azithromycin vs pyrimethamine and sulfadiazine for the treatment of ocular toxoplasmosis. *Am J Ophthalmol*. 2002(134):34-40.

18. Gubbels MJ, Lehmann M, Muthalagi M, et al. Forward genetic analysis of the apicomplexan cell division cycle in toxoplasma gondii. *PLoS Pathog*. 2008;4(2):e36.

19. Farrell A, Thirugnanam S, Lorestani A, et al. A DOC2 protein identified by mutational profiling is essential for apicomplexan parasite exocytosis. *Science*. 2012;335(6065):218-221.

20. Illumina Inc. HiSeq 2500 specifications. http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn. Updated 20142014.

21. Kent WJ, Haussler D. Assembly of the working draft of the human genome with GigAssembler. *Genome Res*. 2001;11(9):1541-1548.

22. Myers EW, Sutton GG, Delcher AL, et al. A whole-genome assembly of drosophila. *Science*. 2000;287(5461):2196-2204.

23. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res*. 2010;20(9):1165-1173.

24. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*. 2008;18(5):821-829.

25. Bradnam KR, Fass JN, Alexandrov A, et al. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*. 2013;2(1):10-217X-2-10.

26. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biol*. 2011;12(2):R18-2011-12-2-r18. Epub 2011 Feb 21.

27. Gnerre S, Maccallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A*. 2011;108(4):1513-1518.

28. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195-197.

29. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48(3):443-453.

30. Lee W, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*. 2014;9(3):e90581.

31. Smith DR, Quinlan AR, Peckham HE, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*. 2008;18(10):1638-1642.

32. Garrison E, Marth G. **Haplotype-based variant detection from short-read sequencing**. *Cornell University Archive*. 2012;arXiv:1207.3907.

33. Marth GT, Korf I, Yandell MD, et al. A general approach to single-nucleotide polymorphism discovery. *Nat Genet*. 1999;23(4):452-456.

34. Saeij JP, Boyle JP, Boothroyd JC. Differences among the three major strains of toxoplasma gondii and their specific interactions with the infected host. *Trends Parasitol*. 2005;21(10):476-481.

35. Gajria B, Bahl A, Brestelli J, et al. ToxoDB: An integrated toxoplasma gondii database resource. *Nucleic Acids Res*. 2008;36(Database issue):D553-6.

36. Khan A, Bohme U, Kelly KA, et al. Common inheritance of chromosome ia associated with clonal expansion of toxoplasma gondii. *Genome Res*. 2006;16(9):1119-1125.

37. Flibotte et al. Whole-genome profiling of mutagenesis in caenorhabditis elegans. *Emerging Infectious Diseases*. 2003;9(11).

38. Farrell A, Coleman BI, Benenati B, et al. Whole genome profiling of spontaneous and chemically induced mutations in toxoplasma gondii. *BMC Genomics*. 2014;15:354-2164-15-354.

39. Niedelman W, Gold DA, Rosowski EE, et al. The rhoptry proteins ROP18 and ROP5 mediate toxoplasma gondii evasion of the murine, but not the human, interferon-gamma response. *PLoS Pathog*. 2012;8(6):e1002784.

40. Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, Boyle JP. Polymorphic family of injected pseudokinases is paramount in toxoplasma virulence. *Proc Natl Acad Sci U S A*. 2011;108(23):9625-9630.

41. Shen Y, Sarin S, Liu Y, Hobert O, Pe'er I. Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. *PLoS One*. 2008;3(12):e4012.

42. Hillier et al. Whole-genome sequencing and variant discovery in C. elegans . *Nature Methods*. 2008;5(2).

43. Araya CL, Payen C, Dunham M, Fields SJ. Whole-genome sequencing of a laboratory evolved yeast strain. *BioMed Central Genomics*. 2010(11):88.

44. Crom et al. Tracking the roots of cellulase hyperproduction by the fungus trichoderma reesei using massively parallel DNA sequencing. *PNAS*. 2009;106(38):16151-16156.

45. Nordstrom KJ, Albani MC, James GV, et al. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat Biotechnol*. 2013;31(4):325-330.

46. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-770.

47. Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: A fast, lightweight genome assembler. *BMC Bioinformatics*. 2007;8:64.

48. Kong A, Frigge ML, Masson G, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*. 2012;488(7412):471-475.

49. Illumina Inc. Platinum genomes | illumina. http://www.illumina.com/platinumgenomes/. Updated 2014. Accessed January 1, 2014.

50. Morgan TH. <*br />A CRITIQUE OF THE THEORY OF EVOLUTION* Princeton: Princeton University Press; 1919.