Spring 2013

Computational methods for efficient exome sequencing-based genetic testing

Adam Peter DeLuca
University of Iowa

COMPUTATIONAL METHODS FOR EFFICIENT EXOME SEQUENCING-BASED GENETIC TESTING

by

Adam Peter DeLuca

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisor: Associate Professor Terry A. Braun

ABSTRACT

Exome sequencing, the process of sequencing the set of all known exons simultaneously using next-generation sequencing technology, has dramatically changed the landscape of genetic research and genetic testing. The incredible volume of data produced by these experiments creates challenges in: 1) annotating the effects of observed variants, 2) filtering to remove noise, 3) identifying plausible disease-causing variants, and 4) validating experimental results. Here we will present a series of bioinformatic tools and techniques intended to address these challenges with exome sequencing and associated validation experiments.

First, we will present the Automated Sequence Analysis Pipeline (ASAP), a tool for the efficient and automated management, detection and annotation of Sanger sequencing data. This tool enables large-scale Sanger sequencing based genetic testing and variant validation efforts.

This pipeline has also been extended to annotate variants discovered by exome sequencing. The ASAP NGS annotation system predicts the effects of a variant observed in genomic sequence on the amino acid sequence, and annotates these changes in the standard nomenclature expected in a clinical setting.

Exome sequencing experiments produce a great number of variants that do not cause a patient's disease. One of the biggest challenges in exome sequencing experiments is sorting through these false positives to discover the true disease-causing variants. We have developed several techniques to aid in the reduction of these errors. The techniques described include: 1) the construction of a catalog of systematic errors by reprocessing thousands of publically available exomes, 2) a tool for the filtering of variants based on family structure and disease assumptions, and 3) a tool for discovering regions of autozygosity from the exomes of several affected patients in consanguineous pedigrees.

Classes of variants that are undiscoverable using current analysis techniques gives rise to false negatives in exome sequencing experiments. We will present a tool, the Retrotransposon Insertion Detector for Exomes (RIDE) that uses the characteristic anomalies present in sequence alignments to detect the insertion of repetitive elements.

The process of identifying the cause of a patient's disease using exome sequencing data has been equated to finding a needle in a stack of needles. Only through the proper annotation of variants and the reduction of the error rates associated with exome sequencing experiments can this task be achieved in an efficient manner.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

COMPUTATIONAL METHODS FOR EFFICIENT EXOME SEQUENCING-BASED GENETIC TESTING

by

Adam Peter DeLuca

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisor: Associate Professor Terry A. Braun

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Adam Peter DeLuca

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Biomedical Engineering at the May 2013 graduation.

Thesis Committee: _____
Terry A. Braun, Thesis Supervisor

_____
Thomas L. Casavant

_____
Todd E. Scheetz

_____
Val C. Sheffield

_____
Edwin M. Stone

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

EXOME SEQUENCING AND ANALYSIS

1.1. Introduction

Exome sequencing is the high-throughput sequencing of every exon in the human genome. This technique provides the unprecedented ability to perform a standardized experimental procedure for patients suffering from many diseases, and then informatically discover the cause of a patient's disease. However, this ability is balanced by the challenge of properly calling variants, annotating the affects of the change, filtering to remove noise, identifying plausible variants and experimental validation of results.

The process of genetic testing and exome sequencing will be discussed over the next four chapters. This chapter will introduce the concepts of exome sequencing and genetic testing. Chapter 2 contains a discussion of genetic testing prior to the advent of exome sequencing and presents the Automated Sequence Analysis Pipeline (ASAP), a tool for the efficient and automated management, detection and annotation of Sanger sequencing-based genetic testing. Sanger sequencing remains relevant despite the advent of next-generation sequencing because of the need for validation on a per-variant basis. Chapter 3 contains a discussion of annotation of exome sequencing variants, and presents an addition to ASAP to allow for the annotation of exome-sequencing derived variants. Chapter 4 is a discussion of false-positives in exome sequencing experiments, and includes several techniques to aid in the reduction of these errors. The techniques described are: 1) the construction of a database of systematic errors by reprocessing thousands of publically available exomes, 2) a tool for the filtering of variants based on family structure and disease assumptions, and 3) a tool for discovering regions of autozygosity from the exomes of several affected patients in consanguineous pedigrees. Finally, Chapter 5 is a discussion of how classes of variants that are undiscoverable using

current analysis techniques gives rise to false negatives in exome sequencing experiments. Specifically, a tool will be presented to detect the insertions of retrotransposons in exome sequence data.

1.1.1. Efficient Genetic Testing Using Sanger Sequencing

Prior to next-generation sequencing, Sanger sequencing was the primary technology for genetic testing. Genetic testing results are used by physicians and genetic councilors to provide prognostic information, family planning advice, and treatment recommendations to patients. In this context, it is critical that Sanger sequencing-based genetic testing results are highly accurate. The cost and volume of data produced by Sanger sequencing lends to manual inspection of sequence traces by trained experts, a process that is difficult to scale to large volumes of sequencing. To this end we have implemented the Automated Sequence Analysis Pipeline (ASAP)[1], a system that provides automated calls on Sanger sequencing reads in addition to tools to aid in the manual review of sequence and the resolution calls when they differ between several sequence readers. This will be described in more detail in Chapter 2.

1.1.2. Variant Annotation

Variant annotation is the process of predicting the impact of a genomic change on the amino-acid sequence of a protein, and to give sufficient context to a variant to allow the identification of a disease-causing variant among tens of thousands of changes identified in a next-generation sequencing experiment. In Chapter 3 I will present a tool, the ASAP NGS annotation system[2], that can calculate the effect of a variant based on gene structure information. Accurate variant annotation is critical to the success of a next-generation sequencing experiment because the experiments produce too much data for manual inspection of results to be practical. Therefore a failed annotation represents a false negative in the experiment.

1.1.3. Reduction of False Positives in Exome Sequencing

Exome sequencing experiments produce many variants that can potentially cause a patient's disease. The vast majority of these variants are either benign polymorphisms, variants unrelated to the disease under study, or artifacts of sequence alignment and variant calling. Chapter 4 contains a discussion of ways to reduce false-positive rates in these experiments using a variety of methods.

To identify artifacts of sequence alignment and variant calling, a large set of publically available exomes has been processed using the exact techniques used in local exome sequencing. This produces a catalog of variants containing both common polymorphisms in addition to common artifactual variants. Because neither of these classes is of interest in an exome sequencing study attempting to identify the cause of a patient's rare disease, removing common variants in this catalog from a patient's exome data reduces the false positive rate. This reduction is greater that the reduction that can be achieved by removing variants using publically available variant catalogs derived from the same input data, because the local catalog captures common errors specific to a particular method of sequence alignment and variant calling.

1.1.4. Reduction of False Negatives in Exome Sequencing

A false negative in an exome sequencing experiment is also costly, not because of the wasted cost of the exome experiment itself, but because of the cost of the necessary validation experiments. False negatives arise in exome experiments from 1) incomplete capture design, 2) regions of insufficient coverage, 3) overly strict filtering when attempting to reduce false positives, and 4) variant types that bioinformatic tools are not designed to detect. In Tucker et. al. 2011, we discovered that an insertion of an ALU element into an exon of the MAK gene was a major cause of Retinitis Pigmentosa in the Jewish population[3]. This mutation was discovered because of experimental validation attempts of artificial mismatches. To close this analytical hole and reduce false positives I

have developed a Retrotransposon Insertion Detector for Exomes (RIDE). RIDE uses the distribution of discordant read pairs surrounding the site of an insertion, and the soft clipping of sequence at the breakpoint to detect these insertions. In simulations, RIDE achieved a sensitivity of 89.1% with a false discovery rate of 16.1% at detecting ALU insertions near exons with at least 20x coverage. Validation efforts are underway to evaluate the tool on real world datasets.

1.2. Background

1.2.1. Sequencing Technology

New sequencing technologies (often called "next-generation sequencing" in the literature) have been developed that have revolutionized how DNA sequencing studies and clinical genetic testing is performed[4,5]. This section is not intended as a detailed description of the technology behind each of these sequencing technologies, but rather is an introduction to provide appropriate context for the bioinformatic methods and tools described later.

1.2.1.1. Automated Capillary Sequencing – ABI 3730xl

Prior to the introduction of these technologies, genetic testing was performed by sequencing a single exon at a time using a capillary electrophoresis sequencer like that shown as A in Figure 1. The Applied Biosystems (ABI) (now Life Technologies) 3730xl sequencer has 96 individual capillaries that can each be used to perform a Sanger sequencing reaction by using four different florescent labels for each nucleotide. A PCR reaction is used to generate fragments at one base-pair increments each ending in a fluorescent-labeled base. The size of these fragments is measured using gel electrophoresis inside the capillaries. Labeled bases are identified as they pass the imaging window pictured on the left. Each capillary yields 300-500bp of sequence for a total yield of 48kB of sequence per run.

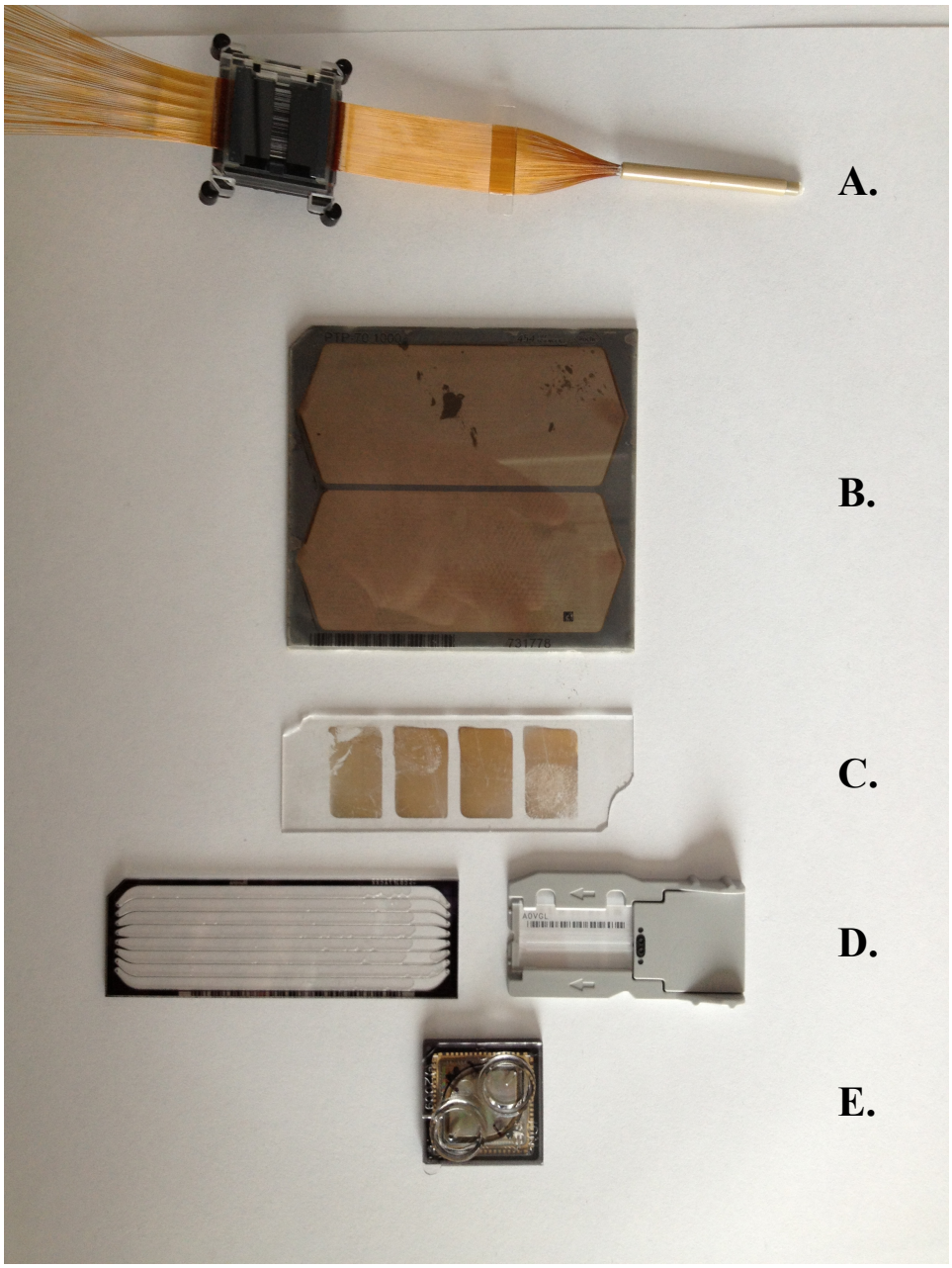

Figure 1. A comparison of sequencing consumables from various sequencing machines.
(A). The terminal end of a 96 capillary sequencing array from an ABI 3730xl Sanger sequencing machine.
(B). A sequencing plate from a 454 GS FLX sequencer, the earliest of the next-generation sequencing technologies.
(C). A sequencing plate from an ABI SOLiD 4 sequencer. Note the four 'quads' to support the simultaneous sequencing of several samples.
(D). Two sequencing flow cells from Illumina sequencers. An eight lane flow cell from a HiSeq 2000 sequencer is shown left, and the single lane flow cell from a MiSeq is shown on the right.
(E). A '318' sequencing chip from a Life Technologies Ion PGM.

1.2.1.2. 454 Pyrosequencing

The first of the next-generation sequencing machines to enter the market was the Roche 454 GS FLX sequencer[4]. This machine represented a radical departure from previous sequencing machines by doing away with individual reaction vessels. It employs an oil and water emulsion PCR to generate many sequencing-ready molecules from the same sample in a single reaction vessel. DNA molecules are bound to beads in the emulsion PCR, and these beads are physically isolated by settling into nano-scale wells in the surface of the plate depicted as B in Figure 1. Each well holds a single bead, so as fluorescently-labeled bases are incorporated, an imaging system can be used to capture the fluorescence from a single source molecule. Current versions of the 454 sequencer allow the imaging of up to a million beads and a read length of approximately 700bp, yielding approximately 700MB of sequence per run.

1.2.1.3. SOLiD Sequencing

Life Technologies's SOLiD sequencing uses a ligation-based chemistry to perform massively-parallel sequencing. Like 454 sequencing, SOLiD sequencing uses an emulsion PCR to generate fragments for sequencing. During each cycle of sequencing, an eight-base oligonucleotide interrogating two bases is incorporated onto the growing sequence. After the whole length of the sequence has been achieved, the process is repeated four more times so that each position has been interrogated by two overlapping octamers. Octamers representing the sixteen possible interrogated dinucleotides are tagged with four different fluorophores. Since fluorophores are reused, the nucleotide sequence must be reconstructed using the entire sequence of fluorophores. Bead-bound fragments are physically isolated for imaging on the surface of the slide shown as C of Figure 1 by randomly depositing the beads in a carefully controlled low-concentration solution on a flow cell, the disposable glass plate where the sequencing reaction takes place. Beads that randomly deposit too close to each other to be separable by the imaging

software are discarded. Current versions of the SOLiD sequencer produce 700 million paired end sequences of 50 base pairs forward and 25 base pairs reverse, yielding approximately 50GB per flow cell per run.

1.2.1.4. Illumina Sequencing

Illumina sequencing was the second next-generation sequencing system to enter the market and is currently the most popular. Illumina sequencing combines the steps of isolating fragments during PCR for colony generation with the physical isolation of the colonies for imaging by using a bridge-PCR technique. Figure 1 shows two different Illumina flow cells in line D, at left is the flow cell from a HiSeq 2000, at right the flow cell from a MiSeq. The MiSeq consists of a single lane per flow cell, the HiSeq has eight lanes per flow cell. In bridge-PCR the flow cell has sequencing primers bound to the surface of the lane. Template molecules are bound to the primers and a PCR reaction is performed where after each round the fragment bends to reach another primer on the surface of the plate. This forms a colony derived from a single template bound to the surface of the flow cell. Sequencing is performed by incorporating a single base at a time and imaging the resulting fluorescently tagged colonies. In contrast to 454 and Ion Torrent, homopolymeric stretches of DNA are sequenced one base at a time leading to a lower insertion/deletion error rate. The current HiSeq 2500 Illumina sequencer produces approximately 1.5 billion pairs of 100bp reads for a yield of approximately 300 GB per flow cell.

1.2.1.5. Ion Torrent Sequencing

Ion Torrent sequencing differs from the other technologies listed here in the way raw sequencing data is captured. In the 454, SOLiD and Illumina systems the raw sequence data is captured by a camera imaging fluorescence emitted by labeled molecules. In contrast, the Ion Torrent system works by using a semiconductor array of millions of pH meters to detect the release of H^+ ions as bases are incorporated. Fragment

preparation uses an emulsion PCR technique very similar to that employed by the 454 sequencing machine. An advantage of this system is that the entire detector assembly is part of the consumable chip pictured as E in Figure 1, meaning upgrades to the system are possible by providing new versions of the chips. The first version of the chip, the 314, produced approximately 100,000 reads of a length of 200 base pairs for a yield of 20MB. The 318 chip pictured in Figure 1 produces around 2.5 million reads of 400 base pairs for a yield of approximately 1GB.

1.2.2. Targeted Exon Sequencing

Even with the capabilities of next-generation sequencers, sequencing the entire genome of a patient is still cost-prohibitive because of: 1) the cost of sequencing, 2) the cost of analyzing the data, 3) the cost of storing the data, and 4) gross deficiencies in our ability to interpret non-coding variations. This led to the advent of technologies that allow the enrichment of regions of interest above the background level of the rest of the genome. These experiments initially targeted the exons of a few genes of interest, or a small genomic interval and quickly grew to capture every exon of every known gene, the exome. It is important to note that there is not a single definition of the exome because differences in gene annotations exist between databases and new genes and exons are being discovered and added to new versions of the capture definitions.

1.2.2.1. Array-Based Exon Capture

The first targeted exon captures were performed using technology very similar to a microarray[6]. Microarrays consist of a set of DNA probes bound to the surface of a slide. Molecules of interest are fluorescently labeled and then allowed to hybridize to the probes bound to the surface of the array. The array is then imaged, and the intensity of the fluorescence is proportional to the quantity of targeted molecules bound to a given probe. This technology has been used to perform genotyping, measure expression, locate binding sites of transcription factors, and many others.

Microarrays can be used to enrich a DNA library for regions of interest by designing probes to tile across the regions of interest. Targeted fragments from the library bind to the probes and are thus anchored to the surface of the slide as non-targeted fragments are washed away. The targeted fragments are then eluted from the surface to yield a library enriched for the regions of interest. During this capture process the array is never imaged as it would be for a microarray experiment, the array is simply a substrate to physically separate the targeted and non-targeted fragments.

1.2.2.2. Solution-Based Exon Capture

Array-based exon capture has been superseded by captures performed entirely in solution. In-solution captures start by designing a set of biotin-labeled RNA or DNA capture oligonucleotides often called baits[7]. A DNA library is allowed to hybridize to the baits. Streptavidin-coated magnetic beads are then added to the solution. The biotin-labeled baits and targeted fragment complexes bind to the streptavidin-coated beads. A magnet is used to pull the beads to the bottom of the tube as the excess solution containing the untargeted fragments is drawn off the top. The targeted fragments are then eluted off of the beads to yield an enriched library.

Solution-based capture have several advantages over array-based captures, most notably, the production of targeting oligonucleotides can take place on a large scale and the resulting bait pool can be divided to perform many captures. This has lowered the costs for commonly performed captures requiring a great number of baits, like an exome sequencing experiment. Solution-based captures have become the predominant form of targeted exome capture.

1.3. Exome Sequencing and Analysis Overview

This is an overview of the exome-sequencing process employed at the Coordinated Laboratory for Computational Genomics at the University of Iowa. For every stage of this pipeline there are dozens of publically available tools. Because of this

huge diversity of tools there are innumerable resulting pipelines to perform exome sequencing. This is intended to provide a point of reference for discussions in later chapters, not a survey of exome analysis techniques.

1.3.1. Sequence Generation

Exome libraries are generated using commercially-available exome sequencing kits from Agilent and Nimblegen. Libraries are sequenced using an Illumina HiSeq 2000 to obtain a depth of at least 20x over the approximately 90% of the targeted regions.

1.3.2. Sequence Alignment and Processing

Sequence is aligned to the genome using the Burrows-Wheeler Aligner[8]. Duplicate sequences, likely artifacts of the PCR used in the library preparation, are removed using the MarkDuplicates utility included in the Picard (http://picard.sourceforge.net) package. Indel realignment and base quality score recalibration are performed using GATK[9].

1.3.3. Calling Variants

Variants are called using the GATK Unified Genotyper[9]. This tool takes input in the BAM[10] file format produced by the sequence alignment tool and produces output in the VCF file format[11]. Variants are typically only called over targeted regions to save computational time, however due to a wide range of commercially available exome sequencing kits; the management of the target files becomes cumbersome. To prevent this variants are called over a target file derived from the union of the target spaces of all commercially available exome-sequencing kits. This eases operational complexities at the cost of computational efficiency.

1.3.4. Annotating Variants

Allele frequencies from the 1000 genomes project[12], the Exome Sequencing Project, dbSNP[13] and local sequencing efforts are annotated to variants. Allele

frequencies from a set of over 1200 exomes sequenced as part of the 1000 genomes project that were analyzed using identical methods to that of the local sequencing effort are also incorporated (see Chapter 4). Additional information, such as RNAseq-derived expression data in relevant tissue, and CHiPseq binding peaks from relevant transcription factors are often incorporated into the variant file. The effect of variant on the amino-acid sequence is predicted using the ASAP NGS annotation system presented in detail in Chapter 3.

1.3.5. Variant Filtering

Exome sequencing experiments yield tens of thousands of variants. A reduction in this number is necessary to make experimental follow-up possible. A detailed discussion of filtering strategies can be found in Chapter 4. Variant filters fall into four broad categories. These categories are: 1) filters intended to reduce false positives by removing low quality variants, 2) filters intended to remove variants that are too common to cause disease in a patient with a rare disease, 3) filters that remove variants that are not predicted to have an affect on the amino-acid sequence of a protein or variants that fall in splice sites, and 4) filters based on sequencing multiple individuals in the same family, removing variants that are inconsistent with disease segregation within the family.

Quality-based filtering is intended to reduce false positives by partially eliminating artifactual variant calls produced by sequence alignment and variant calling algorithms. Variants containing a GATK quality score of less than 25, or a quality by depth score of less than one are removed. In high sequencing-depth studies a minimum percentage of observations is set for a variant.

Allele frequency-based filters are intended to remove variants that are too common in the general population to cause a rare disease. Obviously the value of this filter depends greatly on the population prevalence of the disease and the size of the sampled population. For studies involving rare, genetically heterogeneous Mendelian

diseases a frequency cutoff of 0.6% is used for the populations in the exome sequencing project, and a cutoff of 1% is used for the 1000 genomes project. Variants representing more than 10% of alleles in a database of local sequencing efforts containing a few dozen patients are removed as well.

Positional filters remove variants unlikely to cause disease based on their predicted affects on amino-acid sequence and splice sites. This filtration process is highly dependent on the accuracy of variant annotation tools. Variants falling within 10 bases of a splice site, and variants falling within coding exons that are predicted to cause a change in amino acid sequence are retained.

Finally, family-based filtering is used to remove variants that are not consistent with disease segregation. These filtering rules are highly dependent on the family members that were available for sequencing and the assumed disease inheritance patterns. As such the filtering rules are developed on a case-by-case basis.

1.3.6. Experimental Confirmation and Validation

When used for genetic testing, an exome sequencing experiment is not complete when variants are annotated and filtered. Variants discovered by exome sequencing must be: 1) confirmed to exist in the proband, 2) confirmed to segregate with disease in the family, 3) found to not occur in ethnically-matched control individuals, and 4) variants must functionally impair normal processes, or when functional data is unavailable plausible disease-causing variants in the same gene must be identified in a statistically significant number of additional families. Sanger sequencing is used to confirm that a variant exists in a proband and that the variant properly segregates with disease in the family. Sequencing large numbers of controls and additional patients is accomplished using the Fluidigm Access Array. The Access Array is a microfluidics device that allows the simultaneous PCR amplification of 48 PCR reactions in 48 individuals, and applies barcodes and sequencing adapters that allow sequencing by an Illumina sequencer.

CHAPTER II

EFFICIENT GENETIC TESTING

2.1. Introduction

While next-generation sequencing is rapidly displacing Sanger sequencing for genetic testing, Sanger sequencing is still widely used to test small genes and as a confirmation of next-generation sequencing-based tests. Therefore it is important to efficiently manage Sanger-sequencing based genetic testing. Discussed here is the Automated Sequence Analysis Pipeline (ASAP), a system designed to speed the process of Sanger sequencing-based clinical genetic testing, through automation of file handling, variant calling and variant identification.

This chapter is intended as an overview of the genetic testing and validation exome sequencing results using Sanger sequencing and ASAP and to describe their roles relative to exome sequencing. A more detailed description of the ASAP testing process and interfaces can be found in the Master's thesis: "ASAP – An Automated Sequence Analysis Pipeline for Clinical Genetic Testing[1]."

2.2. Background

2.2.1. Mutation Density Probability Distribution

Sanger sequencing based genetic testing requires the individual amplification and sequencing of each exon of a set of genes. Because testing can stop once causative mutations have been found, the order in which tests are performed can greatly change the overall cost of the genetic test. The Mutation Density Probability Distribution (MDPD) is a sequencing strategy where the amplimers that make up the genetic test are ordered by the frequency that they harbor a disease-causing mutation in the population[14]. This is conceptually similar to a majority class predictor in the machine-learning field. By screening the amplimers that frequently cause disease first, the causative mutations in

patients will often be discovered before all the amplimers in the test need to be sequenced. MDPD can therefore reduce the cost of Sanger sequencing-based genetic testing.

In a dominant disease, testing continues until a single causative variant is identified. In a recessive disease, once a single causative heterozygous variant is identified, screening continues on a gene-specific MDPD until a second causative mutation has been identified.

Because of the serial nature of the MDPD strategy, the time needed to sequence and interpret the results of each amplimer quickly compounds to create impractically long test turn-around times. Therefore, reducing the time needed to interpret the results of each amplimer is critical to creating a practical genetic test. Automation in the quality control, file handling, variant identification, and variant annotation will therefore reduce the costs of genetic testing.

2.2.2. Standardizing Mutation Names

Causative mutations and polymorphisms need to be reported to physicians using a consistent and standardized nomenclature. The Human Genome Variation Society (HGVS) has established a standardized nomenclature for describing the effects of genomic changes on genes, both at the nucleotide level, and at the amino-acid level[15].

According to the HGVS nomenclature standard, the variant NM_000180.2:c.[154G>T]+[=], would be a heterozygous variant of a "G" changed to a "T" found 154 translated bases 3' of the translation start site in the GUCY2D gene (RefSeq number NM_000180, revision 2). In addition to the nucleotide-level HGVS name, variants can also be annotated at the amino-acid level. On the amino-acid level, the above example would be written as: NM_000180.2:c.[Ala52Ser]+[=].

2.2.3. Software and File Formats

2.2.3.1. BLAT

The BLAST-Like Alignment Tool (BLAT) is a sequence alignment program[16]. BLAT is useful for aligning long, highly homologous DNA sequence to the genome. While the performance of the algorithm is too poor for use with next-generation sequence data like exome sequencing, the tool is still very useful at aligning Sanger-sequencing read-derived contigs to the genome.

2.2.3.2. Phred, Phrap, and Consed

Phred is a base-caller for automated capillary Sanger sequencers[17,18]. In addition to producing base calls, Phred assigns quality scores to each assessed base. Phrap is an assembly tool for Sanger-sequencing reads. Consed is a tool for visualizing the Phrap-assembled sequence contigs[19].

2.2.3.3. PolyPhred

PolyPhred is a mutation detection software package that allows the discovery of mutations in Sanger sequencing data[20,21]. PolyPhred is capable of detecting both single nucleotide variants and small insertions and deletions. Because Sanger sequencing simultaneously assesses both alleles of a gene, PolyPhred lacks the ability to identify the content of the insertion or deletion, but can reliably detect its presence.

2.2.3.4. UCSC Annotation Database

The UCSC annotation database is a widely used collection of genome annotation that is available via both the UCSC genome browser, and as a MySQL database[22]. Included in this database are the `'refFlat'` and `'refSeqAli'` tables that contain gene structure annotation including genomic coordinates and exon structure[23].

2.3. Approach

Large-scale genetic testing using Sanger sequencing requires the coordination of many people and software systems. This level of collaboration requires not only the automation of sequence analysis, but also workflow and file management tools.

Because Sanger sequencing-based genetic testing is considered the gold standard for reporting result to patients, and because Sanger sequencing is often used to confirm the results of next-generation sequencing based genetic testing, the experiment requires both high sensitivity and specificity. To maximize sensitivity and specificity we have devised an approach that combines manual and automated calling. Each amplimer is screened automatically using a Phred/Phrap/PolyPhred derived analysis pipeline in addition to being called manually by an expert technician. Both sets of calls are entered into a database and a second expert resolves conflicting calls. During this process, these experts can also identify problematic sequences for further scrutiny or re-sequencing. ASAP provides not only the automated calling pipeline, but also the workflow and file management tools needed to manually read Sanger sequences efficiently. Overall, this process assures that results reported to patients are based on high-quality and consistently annotated sequence[2].

2.4. Methods

2.4.1. Automated Variant Calling

Following automated capillary sequencing, base calling is performed using Phred[17,18]. Forward and reverse reads from a patient are combined with forward and reverse reads from a control individual. This set of reads is assembled using Phrap. Variants are called using PolyPhred[20,21].

2.4.2. Alignment

The consensus sequence from the assembled contig is aligned to the genome using BLAT[16]. The location of variants called by PolyPhred translated into genomic coordinates using this alignment in addition to the information provided by the Phrap-based assembly.

2.4.3. Annotating Variants

Variants annotated in genomic coordinates are translated into nucleotide-space coordinates based on annotation present in the `'refSeqAli'` and `'refFlat'` tables of the UCSC genome annotation database[22,23]. These variants are then annotated using the HGVS variation nomenclature standard[15].

2.4.4. Manual Variant Calling

In order to increase the diagnostic sensitivity and specificity of Sanger sequencing-based genetic tests, manual variant calling is performed in parallel with the automated variant calling. This interface can be seen in Figure 2. Assembled contigs are transferred to the local machine where they can be viewed using the Consed[19] or Sequencher(Gene Codes Corporation) trace viewers. To avoid bias, technicians do not have access to the automatically generated calls. An interface is presented to the user that assists in the entry of variants by providing tools to translate variants into HGVS nomenclature and a list of variants previously observed in the region.

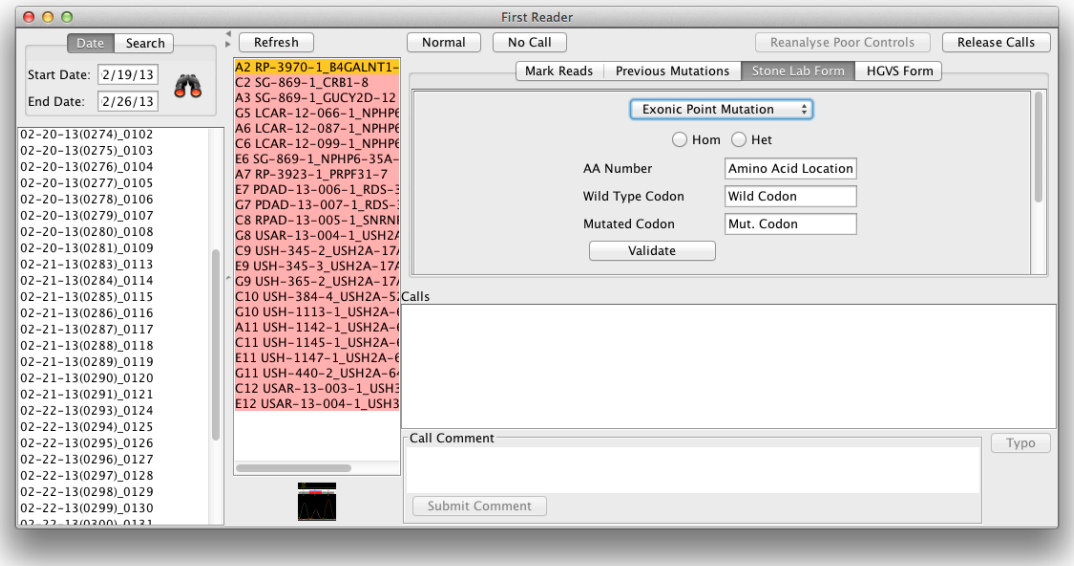

Figure 2. ASAP interface that allows manual calling of variants. This interface does not expose variant calls made by the PolyPhred-based pipeline. The interface provides the ability to launch trace viewer application, and enter variants using either form-based entry or list of previously observed variants in the region.

2.4.5. Resolving Conflicting Calls

Genotype calls in conflict between automated and manual reading are resolved by a second technician manually reading the sequence. This second reader is presented with both the annotated calls from both the automated and manual callers via an interface shown in Figure 3. Sequence viewers can be launched to resolve these conflicting calls. Once conflicting calls are resolved, and calls in agreement are confirmed, the reader can send sequencing results to the laboratory information management system.

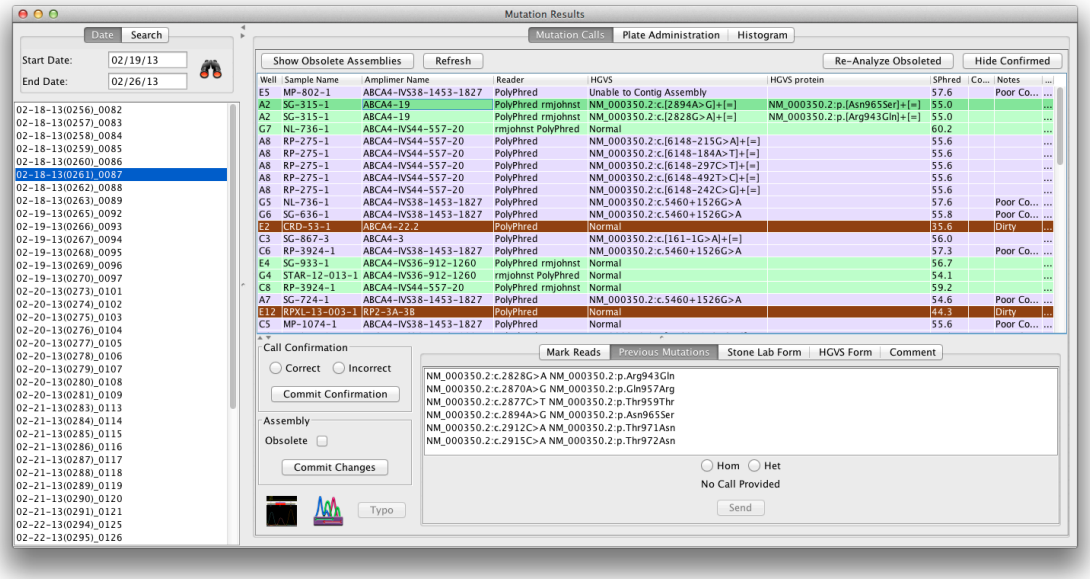

Figure 3. ASAP interface that allows users to resolve conflicting calls produced by manual and automated variant calling. Variants shown in green show genotype calls that were concordant between the software calls and technician calls. Purple shows where a single caller produced a call. The second reader can mark sequence quality, launch trace viewers, and enter new variation calls.

2.5. Results

As of February 2013, more than six hundred thousand Sanger sequencing reads have been processed through ASAP. Figure 4 shows that the cumulative sequencing output of the Carver Non-profit Genetic Testing Laboratory (CNGTL) at the University of Iowa. From these reads, automated and manual readers have called over 220,000 variants. Over this set, PolyPhred achieved an accuracy of 91.65%, manual sequence readers achieved an accuracy of 98.78%. This difference is significant (chi-squared: $p<2.2x10^{-16}$).

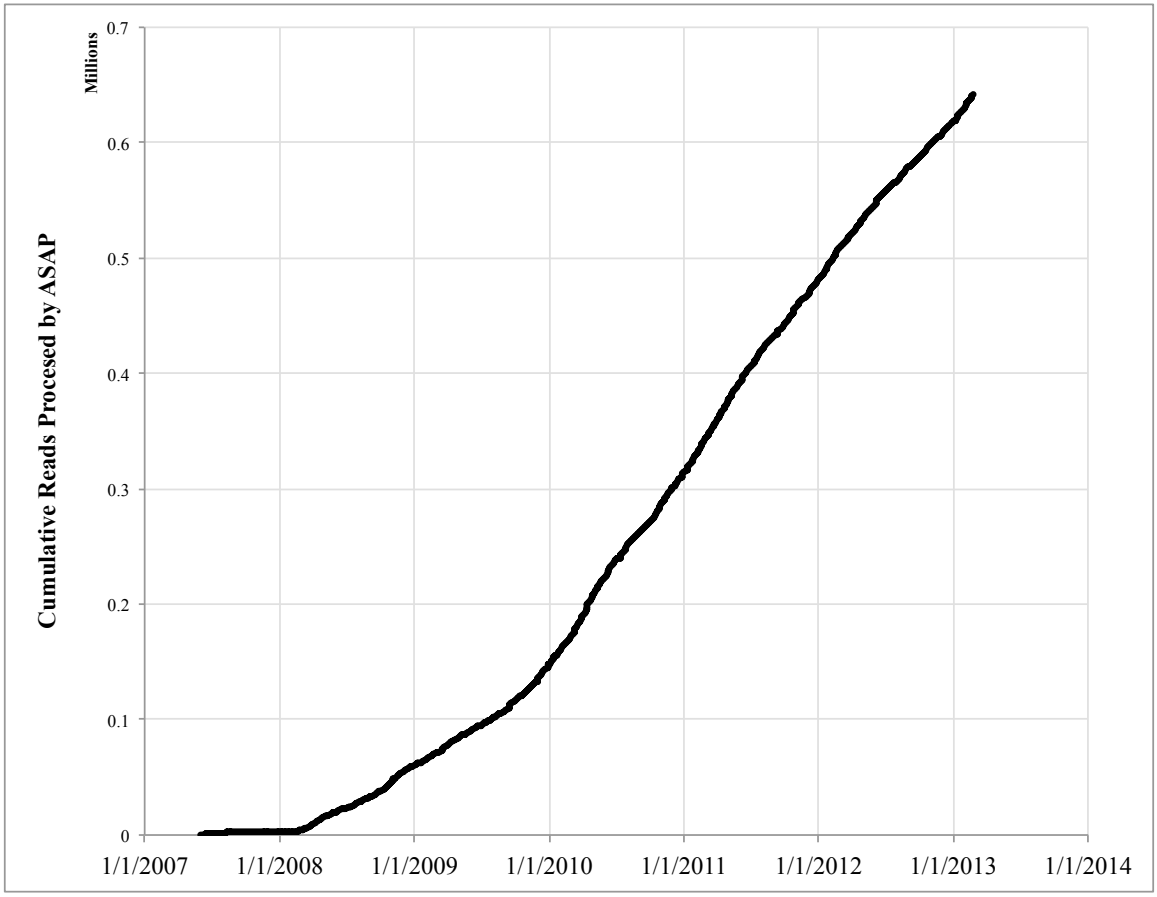

Figure 4. Cumulative count of Sanger sequencing reads processed by ASAP. Note that by 2013, more than 600,000 reads have been processed by ASAP at the Carver Non-profit Genetic Testing Laboratory.

2.6. Discussion

Even with the advent of next-generation sequencing, there is still a need for Sanger sequencing in clinical genetic testing. Diseases that can only be caused by variants in a small number of exons, a small footprint, are still more efficient to test using Sanger sequencing. The number of exons a genetic test must interrogate before it is more cost effective to perform the test using next-generation sequencing is decreasing as technology advances. However, the rapidly decreasing costs of next-generation sequencing are mainly due to increasing the number of reads produced by an experiment rather than reducing the cost of a single experiment. The only way for these advances to

reduce sequencing costs in small footprint genetic tests is to barcode and pool many samples in a single experiment. Cost effective pooled next-generation sequencing experiments of small footprint genetic testing currently requires simultaneous testing of hundreds of samples. This large number of required samples can quickly exhaust the pool of people requiring testing for a rare disease.

A way to overcome this limitation is to create a standardized test that can be used for multiple diseases. Such a test allows enough samples to be gathered in a clinically-useful time period to take advantage of the ever increasing read count produced by next-generation sequencers. An example of this approach is the OtoSCOPE® genetic test offered at the University of Iowa[24]. This test is a targeted capture of exonic regions of approximately 60 genes known to cause hereditary hearing loss and phenocopies such as Pendred Syndrome and Usher Syndrome. Genetic testing for one disease covered by this panel, Autosomal Dominant Non-Syndromic Hearing Loss (ADNSHL), can be accomplished efficiently by using a system called AudioGene to prioritize genes for screening using Sanger sequencing using phenotypic information[25-27]. However, the operational efficiencies gained by performing the same test on many patients, and cost efficiencies of using next-generation sequencing make OtoSCOPE® a more cost effective genetic test than Sanger sequencing exons prioritized by AudioGene[24,28].

In the same manor as the targeted capture employed by the OtoSCOPE® genetic test provides cost effective testing for hereditary deafness-related diseases, whole exome and whole genome sequencing can be employed to perform genetic testing for a number of diseases. While this requires a great deal more sequencing than a targeted capture test, with the costs of next-generation sequencing rapidly falling, these large scale tests will continue to grow in popularity for testing diseases that can be caused by variants in many exons. But until exome or whole genome sequencing is a part of most medical records, it will remain cost effective to perform small footprint genetic tests using Sanger sequencing.

2.7. Conclusion

ASAP provides the tools necessary to perform genetic testing efficiently using Sanger sequencing. This includes automated base-calling, assembly, alignment and variation calling and variant annotation in addition to tools to manage files and results. Software like ASAP is still crucial to the operation of a large-scale genetic testing lab even as next-generation sequencing based genetic testing such as exome sequencing becomes more popular because results still need to be validated. In addition, Sanger sequencing-based genetic testing is currently more cost effective than exome sequencing for diseases that are only caused by variants in a small number of exons. Additionally, as will be discussed in Chapter 4, the same software used to annotate variants using the HGVS standard nomenclature can be applied to annotate variants discovered by next-generation sequencing.

CHAPTER III

VARIANT ANNOTATION

3.1. Introduction

Variant annotation is the process of predicting the amino acid changes that would result from genomic variants in addition to attaching annotations of population prevalence of the variant and other information that help in making a determination of the effect of a variant. Presented here is a discussion of publically available variant annotation systems including an annotation system based on the Automated Sequence Analysis Pipeline (ASAP). The goal of the ASAP NGS annotation system is to provide consistent annotations with those produced by Sanger sequencing confirmation efforts by using a common collection of software.

The ASAP NGS annotation system predicts changes to amino acid sequence based on gene structure information from the RefSeq gene set in the UCSC annotation database[22] and the reference genome. These amino acid sequences are annotated according to the Human Genome Variation Society (HGVS) nomenclature standard[15]. Additionally the positions of variants relative to splice sites are reported.

3.2. Background

3.2.1. Variant Annotation Programs

There are many existing next-generation sequencing variant annotation programs. Presented here is not a comprehensive list, but a survey of programs used in the literature to highlight common features of variant annotation programs and identify some of the shortcomings. Institutional knowledge of variant annotation programs and integration with existing sequencing pipelines appears to be a major reason for selecting an annotation tool despite the particular merits of a tool.

Two common features that help in the integration of theses tools with sequencing pipelines are accepting input in VCF file format, and annotations of changes in transcript space using HGVS nomenclature. VCF files[11] have become the *de facto* standard for next-generation sequencing variant calling programs due in large part to its use in large-scale projects like the 1000 genomes project.

3.2.1.1. ANNOVAR

ANNOVAR[29] is a variant annotation program written in Perl. ANNOVAR uses General Feature Format (GFF) based gene structure annotation. In addition to predicting effects of base substitutions and small insertions and deletions on amino-acid sequence, ANNOVAR will attach annotations of conserved regions, allele frequencies of known variants, and pathogenicity prediction scores. Known variants are available from dbSNP, 1000 genomes, and the Exome Sequencing Project. ANNOVAR can annotate variants with pathogenicity prediction scores from SIFT[30-33], Polyphen2[34], LRT[35], MutationTaster[36], and MutationAssessor[37,38] that have been pre-calculated for every base position in the CCDS gene set by dbNSFP[39].

ANNOVAR uses a custom tab-delimited input file format that is similar to a BED file. Tools are provided to covert the VCF-formatted output of standard variation calling programs to the ANNOVAR input format. A similar custom output file format is also used. ANNOVAR-produced annotations can be added back to the original VCF format using VCFTools[11].

3.2.1.2. AnnTools

AnnTools[40] is a variant annotation program written in python that can annotate changes to amino acid sequence in addition to annotating variants that fall in putative promoter regions. Input and outputs are in VCF format. Annotations are made based on gene structure information contained in a custom MySQL annotation database based on data from the UCSC annotation database[22].

3.2.1.3. MU2A

MU2A[41]is a web-based variant annotation program written in Java. Gene structure information and other source annotation is obtained from a custom database. Input and output are given using custom file formats.

3.2.1.4. SeqAnt

SeqAnt[42] is a Perl-based sequence variant annotation program. Variants are annotated by SeqAnt relative to refSeq transcripts available from the UCSC genome browser database[22]. Input and output formats are custom file formats. Importantly, SeqAnt does not provide annotation of variations affecting amino acid sequence in nomenclature standardized by HGVS[15].

3.2.1.5. SnpEff

SnpEff[43] is a widely-used variant annotation program written in Java. SnpEff has achieved wide spread use through integration with commonly-used analysis tools such as GATK[9] and GALAXY[44,45]. SnpEff uses VCF files for input and output. Annotation databases can be built from a variety of sources including GFF files, UCSC database files, and GenBank files.

3.2.1.6. SVA

SVA is a variant annotation tool written in Java[46]. SVA uses the Ensembl database for gene structure information. Input files are given in the VCF format. In addition to variant annotation, SVA provides a graphical interface that allows users to dynamically adjust filters and visualize the context around variants.

3.2.1.7. VAAST

VAAST[47,48] is a suite of programs for variant annotation, filtering and prioritization. Included in the suite is the Variant Annotation Tool (VAT) that annotates

the functional impact of variants in transcript-space. The VAAST tool VAT should not be confused with another Variant Annotation Tool, VAT[49] of the same name.

3.2.1.8. VARIANT

VARIANT[50] is a variant annotation web site. The site takes VCF input files and outputs data in tabular files and VCF format. Variants are not annotated in standard HGVS nomenclature, limiting the utility of the tool.

3.2.1.9. VEP

The Variant Effect Predictor[51] is a tool available from Ensembl for annotating the effects of genomic changes on Ensembl transcripts. Inputs are allowed in several formats including VCF, and the output contains variant annotation using HGVS nomenclature.

3.2.2. Software and File Formats

3.2.2.1. GFF

The General Feature Format (GFF)[52] is a file format used to store gene structure information. GFF files are commonly used as input to variant annotation programs. The gene structure information is necessary to annotate amino acid changes. GFF files of gene structure can be obtained from the UCSC genome browser, Ensembl, the NCBI, and other sources.

3.2.2.2. PSL

PSL files are the default output of the BLAT[16] sequence alignment program. The format presents start positions in zero-based coordinates, and end positions in one-based coordinates.

3.2.2.3. VCF

The Variant Call Format (VCF)[11] has become the *de facto* standard for storing variants produced by next-generation sequencing variant callers[40]. The format is

extensible and importantly allows sparse representation of annotations that may not be present for all variants. Because VCF files have become the standard for file output from variant callers, it is a commonly used input file format for variant annotation programs. However, because of the sparse representation of annotation data, the VCF file format is rarely presented to clinical end users.

3.2.3. Sources of Gene Structure Data

3.3. Approach

3.3.1. Value of Consistent Annotation

The ASAP NGS annotation system uses the same software as the Sanger sequencing ASAP presented in Chapter 2 to annotate variants. This reuse of software has the advantage of allowing easy comparisons between variants discovered in next-generation sequencing projects and the validation of those variants by Sanger sequencing. The Sanger sequencing pipeline uses Phred, Phrap, and Polyphred[20] to call variants relative to sequenced controls, then uses BLAT[16] to align these contigs to the genome [1]. Next-generation sequencing experiments are aligned to the genome using short read aligners like BWA[8], Bowtie[53], or BFAST[54], variants are called using tools like the GATK Unified Genotyper[9]. ASAP provides different input and output formats to handle the very different experimental procedures and still provide consistent variant annotation.

3.3.2. Annotating Multiple Transcripts

In exome sequencing projects, variants are discovered in genomic sequence. Because several transcripts can overlap a given site, properly predicting the effects of these variants on amino acid sequences requires annotating variants against multiple transcripts of the same gene. The ASAP NGS annotation system handles multiple transcripts by annotating a variant against every possible transcript, then prioritizing variant effects to identify a primary interpretation. The primary interpretation is the

variant interpretation that has the greatest impact on the resulting amino acid sequence. For example an exonic variant causing a non-synonymous amino acid change is considered more deleterious than an intronic change in a different transcript. The spectrum of these interpretations is shown in Figure 5. Annotations based on other transcripts are reported as alternate interpretations in the output file.

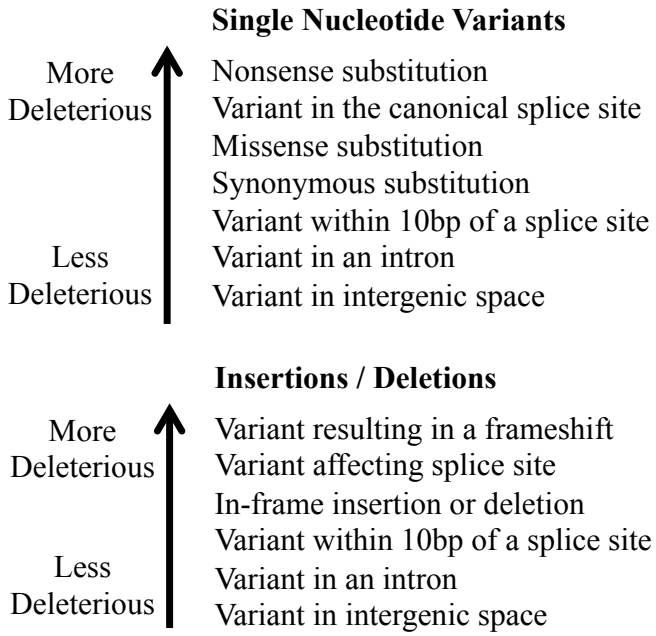

Figure 5. The spectrum of variant interpretations when annotating against multiple transcripts. The most deleterious interpretation is considered primary and others are retained as alternate interpretations in the resulting output file.

3.4. Methods

3.4.1. Input Files

VCF files have become the *de facto* community standard for the output of variant annotation programs[11]. The ASAP NGS annotation system allows the input of variant files in VCF format using a companion utility written in Perl. The companion utility uses

the Perl VCF file API provided by the VCFtools project[11]. The output of the companion utility is a file containing variants in an ASAP-compatible tabular format. This utility can handle VCF files containing multiple samples. In addition the utility recalls zygosity based on a simple rule of <25% allele frequency as homozygous-reference, 25-75% allele frequency as a heterozygote, and >75% as a homozygous variant. This simple rule can correctly identify homozygous variant alleles that are sometimes called heterozygous by the GATK Unified Genotyper.

The ASAP tabular format is an extensible tab-delimited file format where a single variant is given per row. The first column of each row is the only mandatory field. The field consists of the one-based genomic position the reference sequence at that site, and the variant sequence. As an example, "chr1:100000:C>T", The 100,000th base of chromosome 1, a C in the reference is mutated to a T. Insertions and deletions are described in a similar fashion. For example, "chr1:100000:C>-", is a deletion of the C at position 100,000 in the reference genome. Additional columns present in the input file will be retained in the output files. ASAP annotations are added as additional columns on the end of each row.

3.4.2. Obtaining Reference Data

Genomic reference sequence is obtained using a faidx-indexed FASTA file of the reference genome. Faidx indexed FASTA files are produced by samtools[10], and allow random access of sequence stored in the indexed FASTA file. This same indexed reference genome file is used by the GATK[9]. Picard is a Java package of command line tools and APIs for accessing and manipulating SAM/BAM files, and faidx-indexed FASTA files. The Picard API is used by the ASAP NGS annotation system to retrieve genomic sequence from the reference genome.

Gene structure information derived from the RefSeq[23] gene set is obtained from the UCSC annotation database[22]. Gene structure information from the `'refFlat'` table

is used in concert with the BLAT[16]-derived gene alignments of the reference sequences to the genome provided in PSL format via the `'refSeqAli'` table. ASAP can read these datasets via two different mechanisms, first the datasets can be queried from a MySQL database, or the flat files database download files can be directly loaded into memory. Using queries from the database allows quick response times, and allows the use of a standard resource, but the database becomes a bottleneck when simultaneously annotating many files. The flat file option requires users to specify both the `'genepred'` formatted `'refFlat'` file in addition to the PSL-formatted `'refSeqAli'` file. The flat file option is significantly faster than database queries, and scales to many annotation processes running simultaneously.

3.4.2.1. Mitochondrial Genes

Mitochondrial genes are not annotated as part of the regular RefSeq dataset. To properly annotate mitochondrial genes, annotations and alignments equivalent to the entries in the `'refFlat'` and `'refSeqAli'` tables of the UCSC annotation database must be generated. Mitochondrial gene structure was obtained from the UCSC annotation database `'ensGene'` table representing Ensembl transcripts. The sequence of each of these genes was obtained using the UCSC Table Browser[55]. The sequences were aligned to the genome using BLAT[16] to obtain PSL formatted output that corresponds to the entries in the `'refSeqAli'` table. Finally, these newly generated entries for the mitochondrial genes were added to the flat files used as reference data for the ASAP NGS annotation system.

3.4.3. Calculating Transcript-Space Coordinates

Input variants are in genomic coordinates. Predicting the effects of variants on amino acid sequences requires calculating the position of the variant in transcript-space coordinates. The required gene structure information is obtained from the UCSC annotation database[22]. The `'refFlat'` table contains information on the position of the

transcript in addition to the position of splice sites. The `'refSeqAli'` table contains the output of a BLAT alignment of the transcript reference sequence to the genome. It is necessary to have alignment information in addition to the gene structure annotation because there are cases where the reference transcript does not align perfectly to the genome due to insertions or deletions present in the reference sequence.

To calculate the position of a variant given to the annotation system in genomic coordinate space, the first step is to find the transcripts overlapping that position. The position is then located within the BLAT[16] alignment of the transcript provided in the `'refSeqAli'` table. This gives the position as an offset from the transcription start site of the transcript. This offset is then used to calculate the position of the variant relative to the intron/exon structure of the gene provided in the `'refFlat'` table. Cases arise where the position variant itself, the translation start site, or a relevant splice site falls within an alignment gap between the transcript sequence and the genome. In these cases, the transcript-space coordinates of the variant cannot be unambiguously calculated, and the ASAP NGS annotation system therefore throws an error specifying a reference sequence alignment gap. These errors tend to occur in genes with multiple transcripts.

3.4.4. Generating HGVS Nomenclature

Effects of changes on amino acid sequence are annotated using the HGVS Nomenclature standard[15]. Each variant is annotated against every transcript it overlaps producing multiple interpretations. The most deleterious mutation interpretation is retained as the primary interpretation, and other interpretations are kept in a comma-separated field as alternate interpretations.

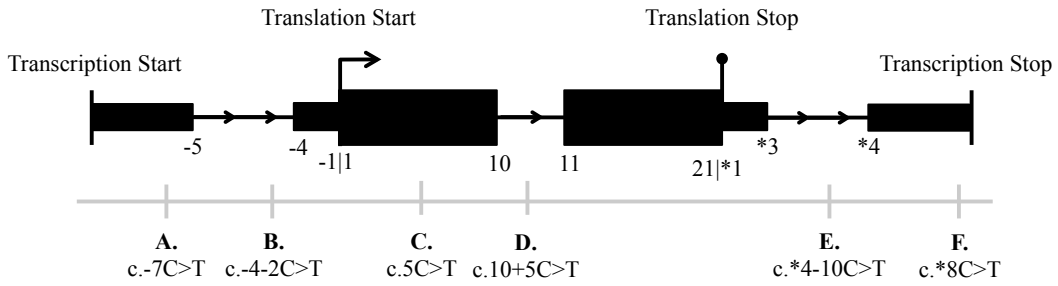

Figure 6. A diagram showing the notional exon structure of a gene to illustrate how HGVS standard nomenclature is generated. The bottom of this diagram is the 5' end of the transcript, and translation and transcription starts and stops are marked. Case A shows coordinates in the 5' UTR, cases B, D and E show coordinates in introns, case C shows a coding exon variation, and case F is in the 3' UTR. Intronic coordinates are annotated relative to the closest exonic base. Note the lack of a zero coordinate at the translation start site.

HGVS nomenclature is based on the position of the variant relative to the translation start site, translation stop site, and splice sites depending on the exact position of the variant. These coordinates are illustrated in Figure 6. The positions of coding, exonic changes are given as the number of exonic bases 3' of the translation start site as shown in case C. Exonic changes in the 5' untranslated region have a negative position as shown in case A. The first base of the start codon has a position of 1, and the exonic base immediately preceding it – the last base of the 5' untranslated region has a coordinate of -1. The positions of intronic changes are given relative to the position of the closest exonic base as is shown in cases B, D, and E. Finally, positions in the 3' UTR are given by the distance from the stop codon as is shown in cases E and F.

3.4.4.1. Mitochondrial Variants

The mitochondrial genome uses a different genetic code than the nuclear genome. Because variants in mitochondrial genes have been shown to cause disease, it is important to be able to annotate mitochondrial mutations properly. The ASAP NGS

annotation system incorporates the alternate codon table required for variants on chrM, and will automatically use this table when appropriate.

3.4.5. Annotations Generated

The ASAP NGS annotation system provides annotations of predicted transcript effects of a variant in HGVS standard nomenclature, both at the nucleotide level and at the amino acid. Alternate interpretations are provided at both the nucleotide and amino acid levels when a variant falls in multiple transcripts. Additionally context information is provided such as gene name, exon/intron number, the distance from the closest splice site, and a description of the variant's overlap with the splice site. Finally, for single nucleotide variants, a BLOSUM62[56] score is provided.

3.5. Results

3.5.1. Current Use of ASAP NGS Annotation System

The ASAP NGS annotation system has been successfully applied to exome and targeted exon experiments in a number of publications[3,24,57-59]. The system was also used as part of the University of Iowa's prize-winning submission to the CLARITY challenge, an international competition on the interpretation of clinical genomic sequencing. In addition the system is currently in use in several CLIA-approved genetic testing laboratories at the University of Iowa.

3.5.2. Limitations of the ASAP NGS Annotation System

There are several limitations of the ASAP NGS annotation system as currently implemented. First, variants that fall within alignment gaps between the reference transcripts and the genome cannot be annotated because the position of the variant in transcript-space coordinate cannot be unambiguously calculated. Similarly, if the translation start site or relevant splice sites fall within these gaps, the variants cannot be

mapped. Additionally the ASAP NGS annotation system cannot annotate multi-base substitutions or a simultaneous insertion and deletion (indel).

3.5.3. Performance of ASAP NGS Annotation System

The ASAP NGS annotation system using the flat file references annotates variants at a rate in excess of 850 variants per second using a single core of an Intel Xeon E3-1270 processor and using at peak of approximately 1GB of memory. This rate allows a whole exome VCF file of ~150,000 variants to be annotated in approximately 3 minutes. Because this process uses no shared database connections, the performance scales roughly linearly as many samples are annotated.

Using the database connection option, variants are annotated at a rate of approximately 50 variants per second on the same test system. Because a shared database is used the performance of this system does not scale well when many samples are simultaneously annotated.

3.6. Discussion and Conclusion

Annotation of variants is one of the last steps in the complicated process of analyzing human sequence data from next-generation sequencing experiments. Importantly, it is variant annotation that allows experts to interpret variants and identify causative variants among tens of thousands of exomic variants, derived from alignment data based on millions of sequencing reads. Conceptually, variant annotation is not a difficult problem, given an input variant calculate the position of the variant within a known gene structure. Because variant annotation is such an essential step in next-generation sequencing experiments, ideally, a common standard and tool for variant annotation would emerge. Such a common tool would produce consistent results between laboratories, and allow greater interoperability between downstream analysis tools.

However, differing needs and preferences of end users has driven a proliferation of variant annotation tools. Until a clearly superior method for inferring pathogenicity emerges, it is likely that there will continue to be a range of tools for variant annotation.

CHAPTER IV

REDUCTION OF FALSE POSITIVES IN EXOME SEQUENCING

4.1. Introduction

The high false-positive rate of exome-sequencing experiments is a driving factor in the costs of exome sequencing experiments by increasing the need for validation experiments. Discussed here are methods to reduce the false positive rate through more effective filtering. First, typical filters that are used to reduce false positives based on quality, predicted amino acid impact, and prevalence in common variation databases will be discussed. Second, a database built from observed variations from locally analyzed exomes will be shown to be a useful filter even after removing common polymorphisms based on thousands of previously analyzed exomes. Further, this filter will be improved by calling publically available datasets with the exact analysis techniques used for a disease study. Third, a tool will be presented that can reduce the false positive rate by sequencing multiple individuals in a family and removing variants inconsistent with segregation of the disease allele. A tool that allows this filtering to be performed on arbitrary family structures will be presented. Finally, a technique will be presented to identify and filter based on regions of autozygosity in multiple affected individuals in a consanguineous pedigree. These filtering techniques are then applied to a large, publically available exome dataset of consanguineous ciliopathy pedigrees.

4.1.1. Cost of a false positive

The cost of a false positive in an exome sequencing experiment is the cost of the required validation experiment needed to assess if the variant allele was truly disease causing. Experiment design drives the validation requirements. Experiments designed to detect polymorphisms or common disease-associated alleles require less validation than experiments that seek to implicate a rare variant as a highly penetrant disease causing mutation.

4.1.2. Sources of false positives

A false positive in an exome sequencing experiment is a variant that requires additional observation to determine if the variant contributes to the patient's disease. Several sources introduce false-positives into exome sequencing experiments; sequencing errors leading to low quality variants, polymorphisms that are too common to cause disease, regions including introns and intergenic regions that are less likely to harbor disease-causing variants, and variants or regions that are inconsistent with segregation of disease in a family. Various filtering techniques can be applied to each of these sources to combat false positives. Care must be taken, however, to avoid overly strict filtering that will lead to false negatives.

4.1.2.1. Low Quality Variants

False positives can result from artifacts in library preparation, sequencing, genomic alignment or in variant calling. These artifacts often lead to low quality base scores from the alignment algorithms and/or low quality variant scores from the variant calling algorithms. Filtering is accomplished by setting appropriate quality score thresholds. Additionally, systematic errors can be detected by processing many samples using identical methods to those employed in local sequencing efforts.

4.1.2.2. Common Polymorphisms

Highly penetrant mutations that cause genetic disease are rare in a normal population. Removing variants that are too common to cause disease based on assumptions made about the disease prevalence and the penetrance of the mutation is a powerful filter for reducing false positives. Databases containing variant calls and/or raw data from tens of thousands of individuals are publically available.

4.1.2.3. Regions Unlikely to Harbor Disease-Causing Mutations

Positional filters are used in reducing the false-positive rate by only considering variants predicted to affect amino acid sequence or RNA splicing. These filters are highly dependent on accurate annotation, and are discussed in more depth in Chapter 3. These filters will often remove intronic variants that could be involved in disease by activating a cryptic splice site, and variants in affecting the function of promoters, thus leading to false negatives. These positional filters will improve as better tools to predict non-exonic variants are developed.

4.1.2.4. Variants Inconsistent with Disease Segregation

There are several ways that pedigree-based information can be used to improve exome sequencing studies. The first way is sequencing a single individual as a follow-up to a region that was mapped genetically using traditional means such as linkage analysis. Second is sequencing multiple individuals from the same pedigree and removing variants that are inconsistent with the assumed inheritance pattern.

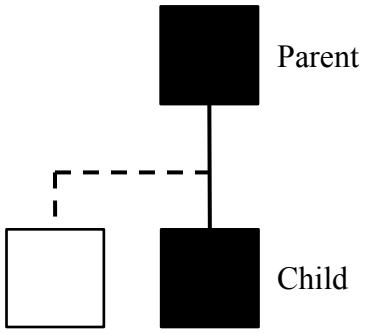

Figure 7. A dominant pedigree too small to perform traditional linkage analysis. With exome sequencing from only a parent and a child a 50% reduction in the number of variant is possible. Another 50% reduction can be achieved by sequencing an unaffected sibling.

Sequencing multiple individuals in the same pedigree allows the study of families that are too small to perform genetic linkage. This method has been employed to discover disease genes[60]. As an example, consider a dominant pedigree with an affected parent and child and no other available individuals as shown in Figure 7. In traditional linkage analysis this would yield a maximum LOD score of ~0.3, so no findings can be statistically significant. However sequencing the exomes of both individuals still provides great utility. The child would be expected to share half of the correctly genotyped variants for which the parent was heterozygous. A simple rule of "both samples must be heterozygous for the variation" would represent a 50% filter (assuming the disease allele is rare and is only one allele in one parent). An additional unaffected sibling would similarly provide an additional 50% filter. In this simple example, all else being equal, a parent-child combination is superior to two siblings because observing the parent ensures that you will observe at least one additional allele (two parental alleles in the child, and two alleles in the parent). With two siblings, you may unluckily observe only two alleles if the siblings share the same alleles.

In a real-world clinical setting the individuals available for sequencing will vary widely from family to family. Any procedure to handle filtering in these situations must be able to accommodate a wide range of family structures and disease inheritance assumptions to be able to maximally reduce false positives based on available samples.

4.1.2.5. Variants Outside of Regions of Autozygosity

Studying recessive disease in families with known consanguinity allows additional filtering beyond testing if the causative variant is consistent with segregation. In these families it is assumed that the same causative variant is inherited from both parents, and further that these alleles are inherited from a recent common ancestor of the parents. This homozygous variant from a common source is referred to as autozygous. Because of the common source of the variant and the small number of meiosis that took

place, the causative variant will be inherited in a large region of autozygosity. In affected individuals, this region will contain only autozygous variants and sequencing artifacts. The traditional approach to identify these regions is homozygosity mapping, where several polymorphic markers are genotyped to identify regions where affected family members are homozygous for the variant and unaffected family members are heterozygotes[61]. A similar approach can be followed using exome or whole genome sequencing by sequencing multiple family members, identifying variants, and using the identified variants to calculate regions of autozygosity. Variants that fall outside the identified regions of autozygosity are inconsistent with the hypothesis of a recessive disease in a consanguineous family and are therefore false positives that can be safely removed.

4.1.3. Variation Discovery vs. Disease Gene Discovery

Variation discovery and disease gene discovery efforts based on high throughput sequencing are similar in methods but have some important differences in goals. In a variation discovery effort like the 1000 Genomes Project, the goal is to catalog all common variations in a set of populations. In this context the cost of missing a mutation (a false negative) in a single individual is low, if the variant is truly common in the population, sequencing additional individuals will discover it. The cost of falsely reporting a variation (a false positive) is higher as it will lead to wasted validation efforts. Because of this the 1000 Genomes project employs a variation-calling scheme where several different analytical methods are used to call variants, and then a voting scheme is used to select the consensus call[12]. This has the effect of increasing specificity at the cost of sensitivity.

Disease gene discovery efforts differ from variation discovery efforts in the relative cost of the error modalities. A false negative in a disease gene discovery method means the causative mutation in the patient was missed, this means all effort and expense

spent on the experiment and subsequent validation are wasted. The cost of a false positive is lower as it requires only a single accurate validation sequencing experiment to reveal the false discovery. Given these costs it is logical to increase sensitivity at the expense of specificity.

As an example of how this difference in focus can lead to false positives, consider the case of chr1:144852476:T>C. This variant occurs heterozygously in 14 of the local exomes and in 1232/1237 analyzed exomes from the 1000 Genomes Project. Given the high proportion of heterozygous individuals, and that the region contains many mapping quality zero reads, the variant is likely an artifact caused by mapping reads from several regions onto a single locus. Variants are filtered from the 1000 genomes calls based on Hardy-Weinberg equilibrium. Because the 1000-genomes release calls do not contain chr1:144852476:T>C, the variant is not filtered out based on minor allele frequency, and thus becomes a false positive in a local exome sequencing experiment.

4.2. Background

4.2.1. Standard Filtering Practices

Exome sequencing projects produce large numbers of false positive variant predictions. These false positives fall into two broad categories, sequencing artifacts and variants unlikely to cause disease. Both of these categories are termed false positives here because neither class contributes to a patient's disease, and both classes require experimental follow-up. An individual variant arising from a sequencing artifact is actually less harmful to the overall experiment than a rare variant unrelated to the patient's disease. Both classes of variants require confirmation by Sanger sequencing in the patient, but the rare variant also requires validation in the family and control individuals.

Sequencing artifacts are erroneous calls produced by artifacts in the library preparation, sequencing, genomic alignment, or variant calling. Reducing these false

positives is a matter of assigning quality scores to the alignments and variant calls and setting appropriate quality score thresholds to filter out these errors. Because this filtering process is highly dependent on the sequencing technology, genomic alignment tool, and variant caller in use, a discussion of these quality-based filters is outside the scope of this work.

Variants that fall in intergenic space, in introns not near splice sites, and cause synonymous amino acid changes are less likely to cause Mendelian disease, and are therefore removed in many standard exome-sequencing pipelines. Because this positional filtering process is dependent on accurate annotation, these filters of variants unlikely to cause disease are discussed in Chapter 3. Additionally, variants that occur commonly in normal populations are unlikely to cause disease. The specific definition of "too common" depends both on the sample size of the normal population in addition to the prevalence of the disease.

4.2.2. Common Variant Databases

There are several sequencing efforts that make exomes publically available. Two in particular, the 1000 Genomes Project[12] and the Exome Sequencing Project release variation calls that are useful in identifying and removing common polymorphisms from exome sequencing projects. Variant calls from both of these projects are released in Variant Call Format (VCF)[11].

4.2.2.1. 1000 Genomes Project

The 1000 Genomes project is an effort to identify variants in the population. As of June 2012 it consists of exome and low-coverage genome sequencing from 1997 individuals belonging to 20 different populations. Of these 1997 exomes, 1237 have been sequenced on the Illumina exome platform. The project calls variants using a combination of three pipelines based on GATK[9,62], QCALL[63], and MaCH[64]. Raw sequence data from the 1000 genomes project are freely available from several sources.

4.2.2.2. Exome Sequencing Project

The Exome Sequencing Project identifies allele frequencies from unrelated individuals that have been sequenced by various disease-based studies. As of June 2012, this set consists of approximately 6500 exomes. Variants are called using the UMAKE pipeline from the University of Michigan. A subset of the raw sequence data used to derive these variant calls is available through dbGaP. It is important to consider that the samples that make up the ESP6500 dataset are derived from studies of a range of diseases. As the patients in the constitutive studies were selected for affection status of a range of mainly heart and lung disease, care must be taken when using the ESP6500 dataset as a control for related diseases. For unrelated diseases, the set may be considered un-enriched for the disease, but should not be considered disease-free. Individuals within the population are indeed carriers for known, rare, disease-causing mutations.

4.2.2.3. dbGaP

In addition to the variant calls released by the 1000 genomes project and the exome sequencing project, tens of thousands of additional exomes are available through dbGaP. Datasets in dbGaP are typically diseased samples and a smaller number of controls. One dataset available through dbGaP is the ciliopathy exome sequencing initiative, containing 1059 exomes from 353 families and additional sporadic individuals. Ciliopathies are a group of diseases affecting the primary cilium. The specific clinical diagnoses of the families in this dataset are not disclosed via dbGaP. The ciliopathy exome sequencing initiative data set contains consanguineous pedigrees with multiple individuals with many different pedigree structures and is used here for an example of the filtering techniques described in this chapter.

4.2.3. Existing Regions of Autozygosity Filtering Techniques

The most common method for filtering exome data to remove variants falling outside of regions of autozygosity is to run genotyping microarrays on the family [60,65-74]. Typically these studies involve exome sequencing in only the proband, and array-based genotyping in the entire family. While this strategy has been very successful at identifying disease genes in consanguineous lineages, with the costs of exome sequencing and whole genome sequencing dropping rapidly, the cost advantage of running a separate genotyping experiment instead of sequencing the additional family members is disappearing.

There are several existing bioinformatic tools for identifying regions of autozygosity based on exome sequencing of multiple individuals, namely IBD2 and the AgileVariantMapper. The difficulty in identifying regions of autozygosity based only on exome data is caused by a high false positive rate, the uneven distribution in the genome of exons containing polymorphic markers, and the presence of pseudogenes[75] and closely related gene families that produce artifactual heterozygous calls.

4.2.3.1. IBD2

IBD2 is an R package for discovering regions of identity-by-descent using a hidden Markov model[76,77]. Identity-by-descent analysis seeks to discover shared haplotypes between affected individuals, and can be used in a similar manner to autozygosity mapping to reduce an exome search space.

4.2.3.2. AgileVariantMapper

AgileVariantMapper[78] is a program written on the .NET framework, that allows the manual, visual identification of regions of autozygosity from exome data by displaying chromosome-level plots of zygosity. While such a manual process may be

useful when studying a small set of families, this, and other manual approaches do not scale well to sequencing many families.

4.2.4. Software and File Formats

4.2.4.1. BAM file format

The sequence alignment/map (SAM) file format is used to store sequence data and alignment information for short read sequence, including exome sequencing experiments[10]. A binary encoding of a file in the SAM format is referred to as a BAM file. The SAM/BAM file formats have become the *de facto* standard for storing aligned reads from next-generation sequencing projects.

4.2.4.2. BEDTools

BEDTools is a program useful in looking for overlapping genomic intervals in BED or VCF formats[79]. A BED file is a tab delimited file format used by the UCSC genome browser to define genomic intervals[22,55]. BEDTools is capable of performing many conceptually simple operations such as finding intersections between multiple files, merging overlapping features, subtraction of features in one file from another, and sorting. By combining multiple commands, powerful analyses can be performed.

4.2.4.3. Burrows-Wheeler Aligner

The Burrows-Wheeler Aligner (BWA) is a next-generation sequence alignment program that uses the Burrows-Wheeler transform to align paired-end short reads to the genome[8]. BWA takes input sequence stored in the FASTQ format and outputs mapped sequence in the BAM format.

4.2.4.4. Genome Analysis Toolkit

The Genome Analysis Toolkit (GATK) is a software package used for the processing, realignment and variant calling of exome data[9]. The GATK Unified Genotyper is used for calling variants in exome sequencing projects.

4.2.4.5. FASTQ File Format

The FASTQ file format is a standard file format containing base calls and quality scores from next-generation sequencers[80]. The exomes referred to in this chapter employ paired-end sequencing, in paired-end sequencing the forward and reverse reads are held in separate files.

4.2.4.6. Picard

Picard is a set of tools written in Java for manipulating SAM and BAM files. Included in this toolset is MarkDuplicates, a program that removes redundant reads from SAM and BAM files, thereby eliminating apparent PCR artifacts.

4.2.4.7. PLINK

PLINK is a widely used software package for the analysis of genome wide association study (GWAS) data[81]. Typically the input data for PLINK are genotypes derived from microarray genotyping arrays. In addition to software for association tests the package contains software to detect identity by descent, large regions of homozygosity and other effects expected in pedigrees.

4.2.4.8. Tabix and bgzip

Bgzip is a program for compressing genomic data stored in a number of common formats and allows indexing by Tabix[82]. Once a file is indexed, Tabix can efficiently perform random access operations. The Tabix software package also provides APIs in a variety of programming languages including Perl.

4.2.4.9. VCFTools and VCF File Format

The variant call format (VCF) is the standard file format used by variant calling programs like the GATK Unified Genotyper to report variants discovered in exome sequencing projects[11]. VCF is a sparse and extensible file format that can handle custom annotation of variants called in one or multiple samples. The file format is user extensible, and allows the addition of arbitrary annotation as a key-value pair in the INFO field, and allows arbitrary tags to be added to variants via the FILTER column. VCFTools is a software package that allows the creation, filtering, and manipulation of VCF files. The VCFTools package provides programmatic interfaces to the VCF file format written in Perl and several other languages.

4.3. Approach

4.3.1. Common Variant Filtering

Because of the differing goals of variation discovery and disease gene discovery efforts, variant calls produced by variant discovery projects are less than ideal for filtering out analytical artifacts from sequencing performed for disease gene discovery projects. To address this, a collection of publically available exomes from the 1000 genomes project and from dbGaP has been downloaded and processed using the same procedure as used for the exomes in the disease gene identification effort. This filtering is compared to filtering performed based on a Local Variation Database (LVD) of a few dozen individuals sequenced locally. Variations are called on samples individually to be most similar to the sequencing performed for disease gene identification, and in addition all samples are called in a single genotyping run so that the total number of samples assessed at a given site can be accurately determined.

4.3.2. Disease Segregation Consistency Filtering

Sequencing multiple family members can be a powerful technique for reducing the number of false positives in an exome sequencing experiment. Removing variants whose genotypes across the family are inconsistent with the segregation of the disease can reduce false positives that arise from errors in the sequencing, alignment, or variant calling, in addition to reducing real genomic variants that are not involved in the disease. Because of the large cost of validating variants, reducing the number of false positives by sequencing multiple family members is often more cost effective than validating additional variants. To realize these benefits in as many patients as possible it is necessary to be able to handle a wide variety of family structures and disease inheritance assumptions.

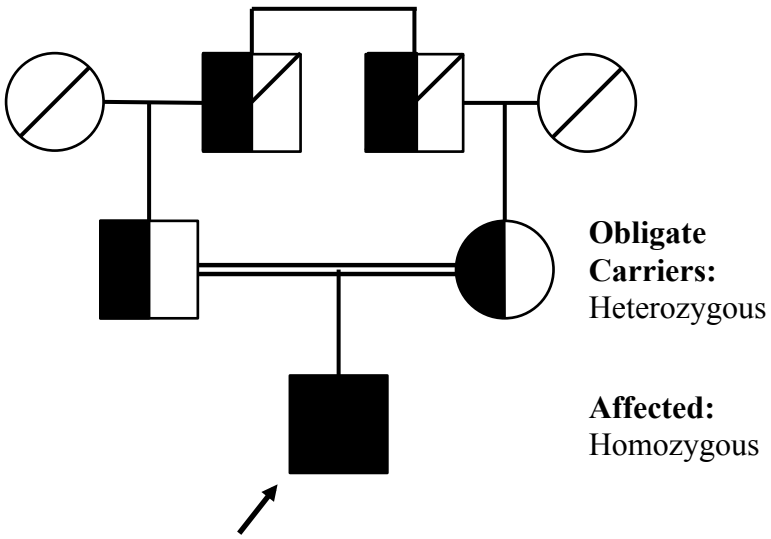

Figure 8. An example consanguineous pedigree with a recessive disease. In a consanguineous lineage the same disease allele is likely inherited through both parents from a common ancestor. Therefore the obligate carriers will be heterozygotes for the disease allele and the proband with be homozygous.

For example, consider the case of a recessive disease in a consanguineous pedigree, like that shown in Figure 8 where a trio is sequenced, an affected proband and both parents. In this case, the disease-causing variant is likely to have been inherited from a common ancestor of both parents. This would require that the affected proband be homozygous for the disease-causing allele and each parent to be heterozygous. Variants not fitting this pattern do not fit the disease assumption and can be discarded.

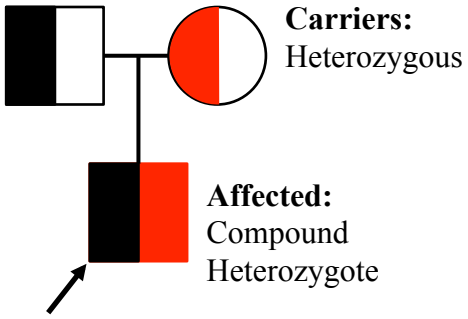

Figure 9. A trio sequenced in an outbred population can have disease caused by compound heterozygous changes. In this case each disease allele is heterozygous in exactly one parent and the affected proband (indicated by the arrow).

If another trio were sequenced in an outbred population like that shown in Figure 9, the filtering would be different. In this case, a compound heterozygous genotype is likely in the proband, thereby requiring two heterozygous variants in the proband that are each shared by one of the parents. In addition to the recessive disease inheritance assumption, the possibility that the disease allele in a single affected child of unaffected parents could have arisen from a *de novo* dominant mutation must be considered. In this case the proband would have a heterozygous variant that was not shared by either parent.

In the case of dominant disease, multiple affected and unaffected family members are often sequenced. In this case all affected individuals in the pedigree must share a heterozygous mutation that is not shared by any unaffected individual.

In addition to these traditional family structures, it can still be advantageous to sequence unaffected siblings and children of probands. For example a heterozygous mutation in the proband will be shared by ½ of their children. When looking for a compound heterozygous variant in a gene, phase of the variants can be established by a single child ½ of the time.

While these rules are easily understood to anyone with knowledge in genetics, devising a software tool that can cope with the combinatorics of an unlimited number of different family structures and any disease assumption would be difficult. In large research cohorts only selecting families with the predetermined family structures that the software is built to handle can easily overcome this difficulty. Another simple heuristic is to only look for either one or two shared alleles in all affected individuals, but this does not use all of the information represented by a family structure and inheritance. These are not attractive solutions in clinical testing where any conceivable family structure is possible. To address this I have developed a rules-based filtering approach that allows custom rules to be specified at runtime. This dynamic rules-based approach allows the definition of new sets of rules when new family structures and disease inheritance structures are encountered.

4.3.3. Regions of Autozygosity Filtering

Studying a recessive disease in a consanguineous family allows an additional stage of filtering beyond common variant filtering, positional filters and filtering for disease consistency in the family. In a consanguineous family a recessive disease is far more likely to result from the same disease allele being inherited from both sides of the family. This is in contrast to the case of an outbred population where compound

heterozygotes are common. Because the same variant is inherited from both sides of the family and there are a limited number of meioses separating closely related affected individuals, the disease-causing variant will be inherited in a large block of autozygous variants.

The challenge in implementing an algorithm to detect large regions of autozygosity in sequence data is creating an algorithm that is tolerant to sequencing artifacts that appear to be heterozygotes. Variants are filtered by quality to reduce the number of artifactual heterozygous calls. Regions of autozygosity are found using a seed-and-extend algorithm while allowing for a number of erroneous heterozygous calls within the region. Variants that fall within identified regions are annotated to allow for filtering.

4.4. Methods

4.4.1. 1000 Genomes Dataset

Paired-end Illumina exomes from 1053 patients were obtained from the ciliopathy exome sequencing project via dbGaP. This dataset represents a near-ideal evaluation set for the filtering methods presented here because multiple family members from consanguineous pedigrees were sequenced. Provided phenotype data contains affection status for all individuals and there are a large number of families. Sequence data was aligned to the genome using BWA and variant calling was performed on families individually using the GATK Unified Genotyper[8,9].

4.4.2. Common Variant Filtering

The 1237 paired-end Illumina exomes completed as part of the 1000 genomes project were reanalyzed using the exact procedures used locally to call variants in disease gene discovery exome sequencing projects (see Chapter 1). Available whole-genome data, and exome sequencing based on other platforms was not used for this analysis to most closely replicate a large set of samples processed locally. The filtering efficiency of

this dataset was compared to a local variation database consisting of a few dozen exomes. Filtering efficiency is shown for an exemplar clinical sequencing family, and for 344 families sequenced as part of the ciliopathy exome-sequencing project.

4.4.3. Disease Segregation Consistency Filtering

In order to remove variants inconsistent with disease segregation in the family under study, a filtering tool has been implemented. The tool uses the Perl module provided by the VCFTools package to read input VCF files. Custom filtering rules are specified on the command line using the nomenclature shown below. Variants satisfying these rules have a filter tag added to the VCF filter column. Tagged variants can be removed from the file using VCFTools, the filter tag can be used as annotation in the final file, or further analysis can be done using these tags.

4.4.3.1. Nomenclature For Specifying Filtering Rules

A nomenclature has been developed for specifying filtering rules for an arbitrary family structure and different disease models. The nomenclature, shown in Figure 10, can specify a set of allowable genotypes to be defined for several individuals. If all of the logical "and" conditions are met, the specified filter tag is applied. More complicated rules requiring a logical "or" can be specified by listing two rules with the same name.

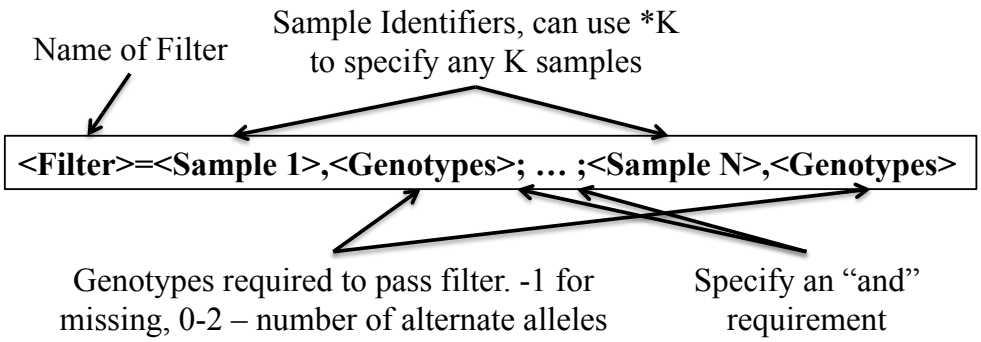

Figure 10. Nomenclature for defining custom pedigree-based filtering rules to identify specific allele combinations. “And” requirements are specified using multiple statements separated by semicolons. “Or” relationships are specified using multiple lines. In this way complex statements can be constructed.

4.4.4. Regions of Autozygosity Filtering

Regions of autozygosity filtering for consanguineous families with exome sequence from multiple affected individuals has been implemented using a combination of publically available tools. Variants are filtered by quality to reduce artifactual heterozygous variants using GATK, variants with a quality to depth ratio below 10 are excluded[9]. At this threshold many true variants will be removed, this is acceptable since the identification of the region means that the variants in the region will be re-examined. The variants in VCF format are then converted into the PED and MAP file formats using VCFTools[11]. PLINK is used to identify regions of autozygosity using the “Runs of homozygosity” program. This program was designed to detect regions shared by distantly related individuals using microarray genotype data[81]. Regions of at least one million bases containing at least 50 non-reference homozygous alleles and no more than ten heterozygous variants are considered to potentially harbor disease causing autozygous alleles. The original VCF file containing all variants is then annotated with these identified regions using a combination of bedtools, Tabix, and VCFTools[11,79,82].

4.4.5. Evaluation with Ciliopathy Exomes

FASTQ files were obtained from dbGaP for the Ciliopathies Exome Sequencing Initiative (dbGaP Study Accession: phs000288.v1.p1). These sequence files are aligned to the hg19 reference genome using BWA. Picard was used to remove duplicate sequences that likely arose from PCR duplicates. Local realignment and variant score recalibration were performed using GATK. Variants were called simultaneously in all members of a family with the GATK Unified Genotyper.

The ciliopathy data set contains exome sequence data from 353 families. Variants from each family were filtered to remove low quality variants, to remove variants too common to plausibly cause disease based on the 1000 genomes dataset and the Exome Sequencing Project (ESP). Additionally variants found to be too common in the 1000 genomes-based LVD were removed. Families consisting of multiple affected individuals were then filtered using the Regions of Autozygosity procedure described above.

The selection criteria of the study required each family to have known consanguinity, therefore regions of autozygosity filtering is applied to families consisting of multiple affected individuals. The common family structures in this dataset are shown in Figure 11. Pedigree-based filtering is performed in each family by retaining all variants where affected individuals in the family are homozygous alternate, obligate carriers (e.g. parents of affected individuals) are heterozygotes, and other unaffected family members are either homozygous reference, or heterozygous.

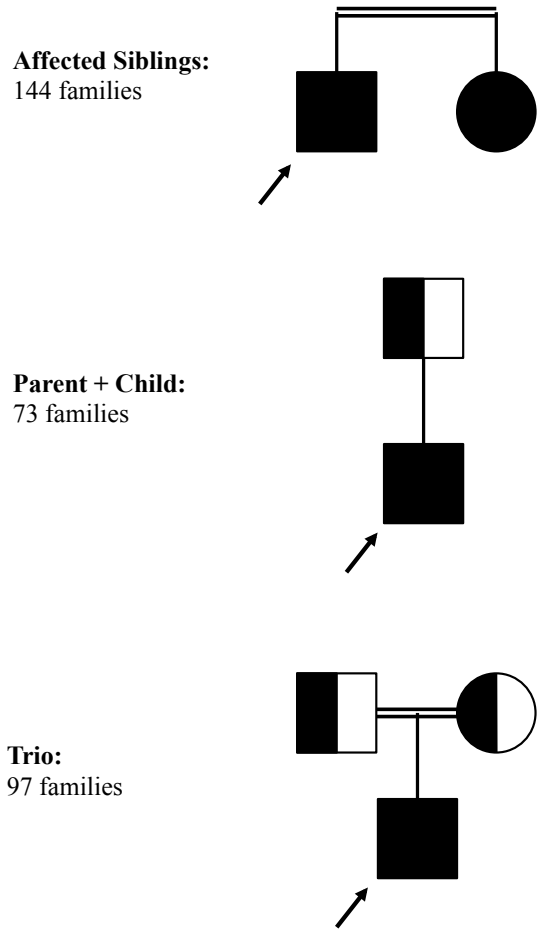

Figure 11. Common family structure present in the ciliopathy exome sequencing initiative dataset. Because families have known consanguinity, affected individuals are expected to be homozygous for a variant and obligate carriers are expected to be heterozygous.

4.4.5.1. Gene List Scoring Metric

Lists of genes harboring plausible mutations were generated using the filtering described above for each of the 353 families in the ciliopathy exome sequencing initiative dataset. In an attempt to prioritize genes that could cause disease in several of the families, a scoring metric, shown in Figure 12, was developed that attempts to reward genes that are present on lists (L) containing few candidate genes (N), and genes that occur on many lists.

$$S = \sum_{i \in L} \frac{1}{N_i}$$

Figure 12. Scoring metric (S) used in the ciliopathy exome analysis. This simple metric rewards genes (i) occurring on multiple lists (L), and genes present on lists with fewer genes(N).

4.5. Results

4.5.1. Common Variant Filtering

An exome dataset from a family segregating recessive disease was filtered using standard protocols as shown in Table 1. A 1000 genomes-based filtering was applied in place of the previously used local variation database consisting of 18 exomes. There were 10,414 variants present prior to LVD filtering. Removing variants with greater than 3 out of 36 variant alleles in the local sequencing caused 588 (5.6%) variants to be removed. This cutoff was chosen using a proportion test to find a cutoff representing a true population prevalence of less than 1% (p=0.000169). When the 1000 genome-based LVD is used for filtering then 6522 (62.6%) variants are removed. It is important to note that this filtering takes place after polymorphisms called by the 1000 genomes project had been removed, therefore this filtering likely removes analytical artifacts instead of polymorphisms.

Table 1. Filtering of an example family with Retinitis Pigmentosa, a recessive disease under two scenarios, using a local variation catalog derived from several dozen locally sequenced exomes, and a local variation catalog derived from reprocessing over 1200 exomes originating from the 1000 Genomes Project.

	Local Exomes	1000 Genomes
All Variations	132,187	132,187
Variant Quality >25	121,736	121,736
Variant Quality/Depth < 1%	120,912	120,912
1000 Genomes Allele Frequency < 1%	13,692	13,692
EVS EA Allele Frequency < 0.6%	10,552	10,552
EVS AA Allele Frequency < 0.6%	10,414	10,414
Local Variation Catalog	9,826	3,892
Retain Exonic and Splice Variants	2,121	1,360
Remove Synonymous Variations	880	667
Remove Single Allele Genes	287	139
Consistent With Segregation	59	23

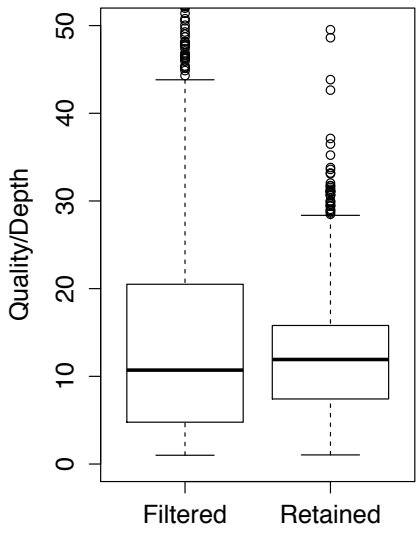

Figure 13. A boxplot showing the distribution of QD scores for variants filtered and retained by the 1000 genomes-based LVD filter.

Further, the filtering performed by the 1000 genomes-based LVD is not simply a filter to remove low quality variants that could have been filtered by other means. Figure 13 shows for this same family the distribution of the ratio of variant quality scores to sequencing depth (QD) for variants both filtered and those retained by the 1000 genomes-based LVD. The clearly overlapping ranges would make filtering by QD ineffective.

4.5.2. Disease Segregation Consistency Filtering

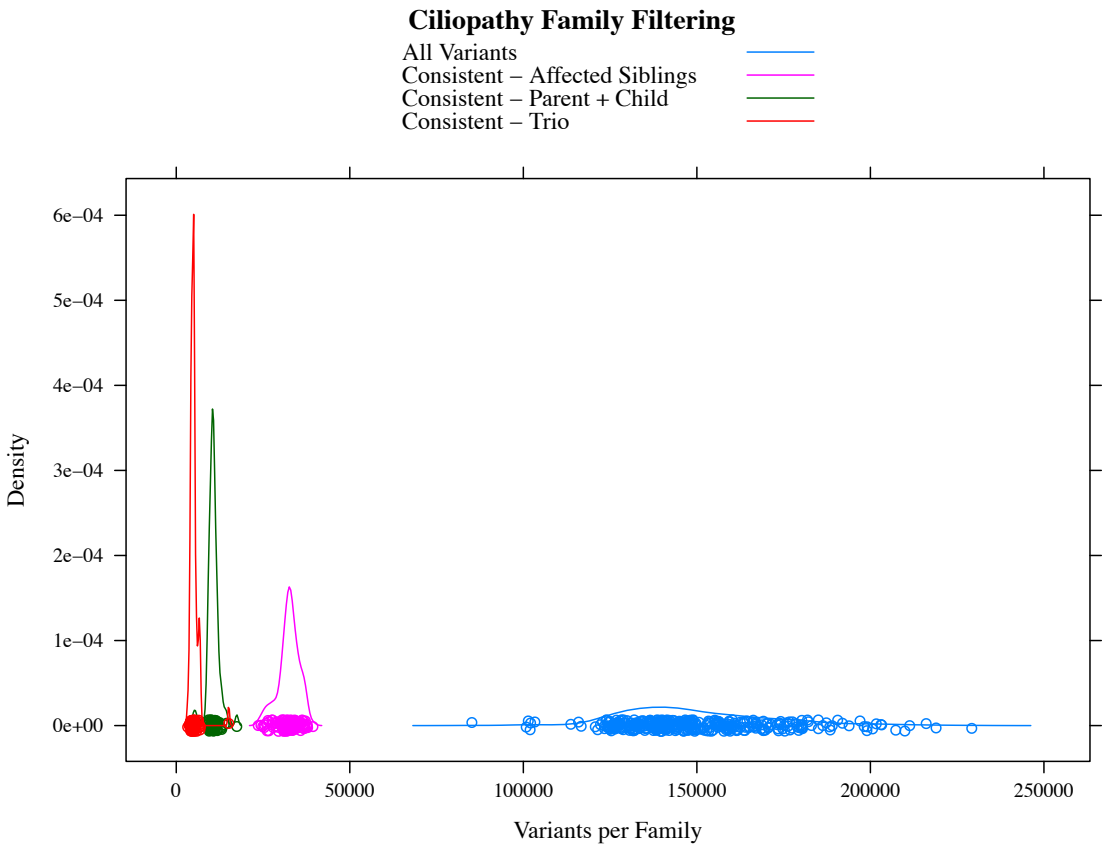

Figure 14. A kernel density estimate showing the distribution of variants per family in the ciliopathy exome set. Starting with the set of all called variants in blue, the variant count is reduced by filtering to remove variants that are inconsistent with segregation in the families. Moving from right to left on the horizontal axis, the impact of using additional family structure information to filter variants shows the substantial reduction of plausible mutations. For clarity, no quality, allele frequency, or positional filters have been applied to these data.

To evaluate the performance of filtering based on consistency with disease segregation in the family, exomes were processed representing the three most common pedigree structures from the ciliopathy exome sequencing initiative (see Figure 11). All members of a given family were genotyped in a single run of the Unified Genotyper, yielding the distribution shown in blue in Figure 14. Note that this filtering is more powerful in a family consisting of a proband and an unaffected parent (shown in green), than in a family consisting of two affected siblings (shown in magenta). This is due in large part to the additional power of an individual expected to be homozygous to remove sequencing artifacts and ethnic polymorphisms. The trio families in red show the most filtering power of the 3 family structures compared.

4.5.3. Regions of Autozygosity Filtering

To evaluate the effectiveness of regions of autozygosity filtering, exomes from families with multiple affected individuals from the ciliopathy exome sequencing initiative were filtered using the autozygosity-based filter and a filter for consistency with segregation of disease. The results of this experiment are summarized by the kernel density estimate shown in Figure 15. Each distribution in this plot represents the same set of families, but with different filtering rules applied. The regions of autozygosity filter (green) is expected to outperform the family filtering (magenta) because both filters require all family members to share homozygous variants. The family filtering requires just the variant under consideration to be homozygous in all family members, the regions of autozygosity filtering requires that a variant fall within a region containing many shared homozygous variants. Note that the magenta distributions from Figure 14 and Figure 15 are drawn from a nearly identical set of families using the same filtering criteria.

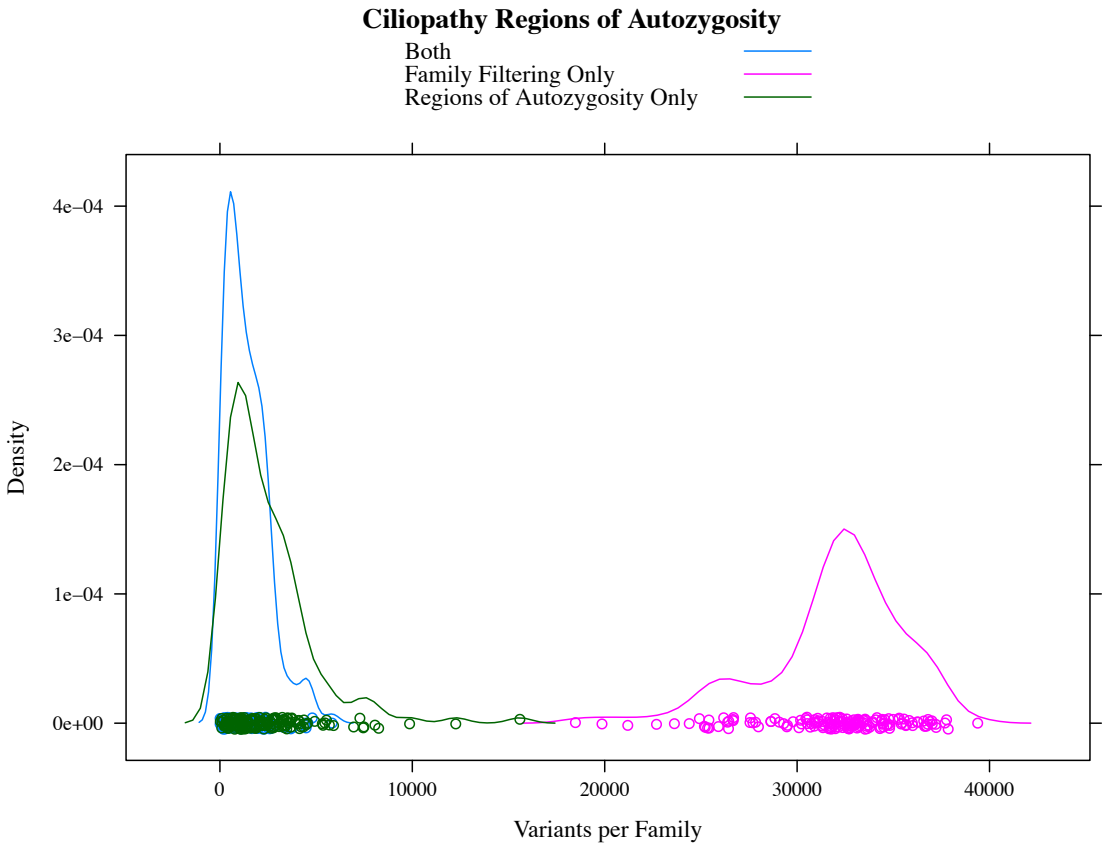

Figure 15. A kernel density estimation showing the distribution of variants per family in the ciliopathy exome set using regions of autozygosity filtering. This shows that filter on autozygosity is more effective than filtering on family structure alone. Only families with multiple affected individuals are used for this processing. Only incremental gains are expected using family filtering in addition to the regions of autozygosity analysis because most of the families represented here contain only affected individuals. For clarity, no quality, allele frequency, or positional filters have been applied to these data.

4.5.4. Ciliopathy Exome Sequencing Initiative Candidate Gene List

To demonstrate the utility of the filtering strategies shown here, the variants from families in the ciliopathy exome sequencing initiative dataset have been filtered to remove low quality variants, variants too common to plausibly cause disease, variants that fall outside of coding exons and splice sites, variants that are inconsistent with disease segregation within the families, and variants that fall outside of identified regions

of autozygosity. Genes containing these variants were prioritized using the metric given in Figure 12. The top 30 candidate genes are given in Table 2.

Table 2. Top 30 candidate genes produced using the family filtering and regions of autozygosity techniques scored using a metric that rewards genes with plausibly disease-causing variants in many families, and also genes that are found on lists of fewer genes.

Score	Gene Symbol	Score	Gene Symbol
2.5011	MUC4	1.3677	POLDIP2
2.3570	RSU1	1.3495	HRNR
2.1373	AHNAK2	1.3307	MICA
1.9569	HEATR1	1.3141	WDR66
1.9560	CREB3L1	1.2175	IFNG
1.7960	MSTO1	1.1849	HLA-DQA1
1.7614	LDHB	1.1634	LILRB3
1.6386	IGSF9	1.1323	KRTAP4-7
1.6205	ACIN1	1.1250	**TSEN54**
1.5716	FCGBP	1.1210	AMPD2
1.5438	RCOR2	1.1082	DARS
1.4705	HLA-DRB5	1.0869	MAGEC1
1.4073	FAM131C	1.0771	YEATS2
1.4027	KLRC3	1.0411	**PCDP1**
1.3771	RPL14	1.0368	RALY

4.6. Discussion

4.6.1. Computation Efficiency Considerations

To generate the data presented above, 2296 exomes were processed and the total size of the input dataset is in excess of 20TB. While sufficient CPU resources were available to process hundreds of these samples concurrently the aggregate input-output

(IO) bandwidth exceeded the capacity of available file systems. To address this it was necessary to develop an optimized job submission protocol that would reduce IO load by taking advantage of inter-process streaming. Previous versions of the submission protocol involved the following steps: 1) Align forward reads with BWA[8], 2) Align reverse reads with BWA, 3) Pair reads with BWA, 4) Convert SAM[10] file format to binary, 5) Sort BAM file. Steps 1 and 2 occurred concurrently on separate nodes taking full advantage of the multiprocessing built into BWA. Steps 3-4 followed sequentially and have only single-threaded implementations. At each stage, large output files were streamed to the file server cluster. The optimized version of the submission algorithm has steps 1-5 running concurrently on a single node, with Unix pipes used to stream results between processes. Buffers between the steps were added and carefully tuned to keep the longest running single-threaded portion of the application (Step 3) running as much as possible. Overall this optimization provided a 4x reduction in read load, a 6x reduction in writes, and an improvement in overall runtime.

It is anticipated that the ever increasing volume of publically available data will continue to stretch the limits of available computing resources that can be allocated to improving filtering based on the reanalysis of publically available data. Relying on cloud computing resources does not present a solution to this problem as the problem of scarcity of computation resources is simply transformed into a problem of the scarcity of financial resources. Possible alternatives to address this issue include: 1) Further computational optimizations, 2) Limit the total size of the reanalyzed data set when adding additional samples until an arbitrary threshold of diminishing returns is reached, 3) Standardization of exome sequencing workflows between research groups, thereby allowing the sharing of reprocessed datasets.

4.7. Conclusion

The false positive rates in exome sequencing-based genetic testing drive the cost of the overall experiments. This is because, in contrast to variant discovery projects, each result must be validated in the patient, confirmed to segregate with disease in the family, and be absent in ethnically matched controls. This confirmation and validation work requires many individual Sanger sequencing reactions, and often large scale experiments such as amplicon-based sequencing on a Fluidigm Access Array. In all but the simplest cases, the cost of these experimental validations exceeds the original costs of sequencing the patient's exome.

In this chapter, several methods aimed at reducing the false positive rate were presented. First, a local catalog of variants was constructed from a reanalysis of over 1200 exomes using identical methods to local sequencing efforts. This catalog was shown to remove more variants than a catalog constructed from a few dozen locally sequenced exomes. Second, a rules-based filtering strategy has been developed to remove variants that are inconsistent with disease segregation when several members of a family are sequenced. Finally, a strategy was described for removing variants that do not fall within regions of autozygosity – the autozygous regions being calculated from sequencing data from several affected individuals from a consanguineous lineage. When combined, these strategies can substantially reduce the false positive rate in exome sequencing experiments.

CHAPTER V

REDUCTION OF FALSE NEGATIVES IN EXOME SEQUENCING

5.1. Introduction

False negatives in exome sequencing experiments are variants that cause a patient's disease, but are not discovered by the exome sequencing experiment. False negatives in exome sequencing projects originate from several sources. First, regions of the genome receive inadequate coverage to call variants because of capture design or the efficiency of capture or sequencing. Second, filtering to reduce false positive based on quality and positional filters will cause the introduction of false negatives. This effect is particularly pronounced on positional filters. These filters remove variants in intergenic regions and deep in introns. While these regions harbor disease-causing variants less frequently than exons or splice sites, there are still many known examples of disease causing mutations in introns and regulatory regions[83]. Finally, some types of variants are not detectable with current sequencing and analysis techniques. These variant types include large insertions or deletions, variants occurring in repetitive regions, and certain types of genomic rearrangements.

A large insertion of repetitive sequence in next-generation sequence data is a particularly difficult variation to detect. Even though these variants are rare, it is still important to detect these events as they have been shown to cause disease [3,84]. Additionally, a large insertion in a coding region is a particularly interesting finding as the disease mechanism is immediately apparent without expensive validation experiments.

Proposed here is a tool, RIDE, that reduces the false negative rate in exome sequencing experiments by detecting insertions of repetitive sequence. RIDE can detect these mutations by taking advantage of characteristic anomalies in the sequence alignment near the site of the insertion. As an example, artifactual single nucleotide

variants (SNVs) led to the discovery of a new gene (MAK) implicated to cause retinal degeneration.

5.1.1. Cost of a False Negative

A false negative carries a large cost in an exome sequencing effort. While a false positive requires the expenditure to validate a single variant, a false negative will result in wasted validation of many false positives as well as the overarching failure to detect the disease-causing variant.

5.1.2. Incidental Detection of an ALU Insertion

Retinitis Pigmentosa (RP) is an inherited eye disease that displays a high degree of locus heterogeneity. While dozens of genes causing RP have been described, a large portion of the disease cases remain unexplained. In a recent study, we implicated MAK as an RP disease gene by exome sequencing[3]. The particular mutation discovered was a homozygous insertion of an ALU element into the ninth exon. Using induced pluripotent stem cells derived from fibroblasts the mutation was shown to disrupt splicing and led to the loss of a retina-specific exon.

The data used to discover the role of MAK in RP was informative and fortunate. The insertion of the repetitive element was not discovered by the bioinformatic variant calling pipeline, but rather the insertion caused errors in the genomic alignment that resulted in several artifactual single nucleotide variants (SNV) being identified near the insertion site. These artifactual SNVs were interpreted as a possible compound heterozygote by the variant calling algorithm. This led to attempted confirmation by Sanger sequencing that instead revealed the repetitive insertion. These artifactual SNVs were just above quality threshold, and therefore slight experimental variability may have caused this important variant to be missed.

This exposes the ability to detect large insertions of repetitive elements as a serious shortfall in current exome sequencing analyses, like that described in Chapter 1.

This is a particularly difficult genomic variant to detect because the insertion will disrupt mapping and capture. In addition, the repeat blocking steps used in library preparation will selectively remove an allele harboring repetitive elements. Another complication includes pairs of chimeric reads that will be induced through mutual priming.

5.2. Background

5.2.1. Transposable Elements

Retrotransposons such as Alus are incorporated into the host genome by a reverse transcriptase (RT). An RT makes a single stranded nick in the host genome at a recognition site, and then synthesizes DNA based on an RNA template. This requires a portion of the RNA molecule to hybridize with the single stranded DNA adjacent to the nick. The effect of this is that several bases from the host DNA are duplicated so they flank the insert in the same orientation. This target site duplication is typically 4-16 bases [85].

5.2.2. Evaluation of Existing Insertion / Deletion Detection Tools

Existing tools such as NovelSeq used to detect insertions use a feature referred to as One-End-Aligned (OEA) read pairs[86,87]. An OEA pair consists of a single read that maps to the reference genome and a read that could not be mapped. Attempts are then made to assemble or otherwise characterize the set of unaligned reads at a plausible insertion site to detect the insertion. The concept of an OEA read pair does not address the case where the inserted sequence is elsewhere in the genome either in a single copy, or in many copies (such as for retrotransposons and similar elements). If there is a single copy elsewhere in the genome the unmapped reads that should lie in the insertion will map to another region of the genome, and therefore will present similar to a translocation. If the insertion is of a repetitive element present in many copies in the genome the

unmapped reads will instead map to a diverse set of loci. One tool that allows for this miss-mapping is PAIR[88], however this algorithm fails to model the effects the duplicated region will have on the insertion of an Alu element. In addition, because repetitive elements are deliberately blocked in the capture the presence of an inserted repetitive element can either cause the capture to fail, or can induce the formation of chimeric reads by mutual priming of fragments. The sequence of these chimeric reads will have no relationship to the sequence of the insert other than to share a region of homology great enough to allow the mutual priming.

5.2.2.1. ClipCrop

ClipCrop[89] is a tool for detecting structural variations using soft-clipping information provided by genomic alignment tools such as BWA[8]. Soft clipping is discussed in detail in Section 5.3.2, but briefly, sequence alignment algorithms will partially align a sequence to the genome and the remaining portion of the read is trimmed. This trimming is encoded in the SAM[10] file, so it can be easily calculated. Reads overlapping the insertion breakpoint will partially align to the genome, and be trimmed past the breakpoint, thus yielding the position of the insertion.

While ClipCrop recognizes the multiple breakpoints present in deletions and tandem duplications, it fails to model the multiple breakpoints present in insertions of transposable elements caused by the insertion-mediated duplication.

5.2.2.2. CNVator

CNVator[90] is the best-known member of a class of tools designed to detect structural variants in next-generation sequencing data using a statistical analysis of read mapping density. The algorithm relies on the number of reads sequenced over a region being proportional to the copy number at that site. Differences in read depth between individuals can indicate deletions or duplications. The authors acknowledge that this approach is ill suited to discover copy number variants created by transposable elements.

5.2.2.3. NovelSeq

NovelSeq[86] is a bioinformatic tool designed to detect insertions of novel sequence. The tool works by preforming a standard paired-end genomic alignment of the reads generated by an exome or whole genome experiment, and separating the read pairs into three categories. First the pairs where both reads align properly to the genome. These pairs do not contain an insert and are discarded. The second category, the read pairs where one read aligns properly to the genome and the other does not map are referred to as OEA paired-end reads. These OEA pairs may span an insertion breakpoint. The final category, called orphan pairs, contains pairs where neither read aligns to the genome. These pairs might lye completely within the insert.

Orphan pairs are processed using the ABySS[91] *de novo* assembly tool to produce contigs that may represent inserts. Contamination from non-human species is removed using a BLAST search[92]. OEA pairs are clustered, and the non-mapped reads are assembled. The OEA assemblies and the Orphan assemblies are then combined to identify the inserts.

Like its name implies, the NovelSeq tool relies on the inserted sequence not being homologous to other regions of the genome. In the case of an insert of repetitive sequence, this is not the case. The discordant read pairs are conceptually similar to the OEA pairs used by NovelSeq, except the non-anchored read will map to another region of the genome instead of failing to map.

5.2.2.4. PAIR

PAIR[88] is an algorithm for detecting Alu insertion events using discordant reads. Inconsistent read pairs are identified by finding reads whose separating distance falls outside the typical range for the experiment. The reads from these pairs are then aligned to known Alu families to identify pairs with one read in the genome, and the other read containing Alu sequence. Sets of these Alu pairs are used to detect the insertion.

Unlike OEA-based tools like NovelSeq, PAIR allows for the detection of repetitive elements by tolerating discordant read pairs where both reads map to the genome. The PAIR model, however, does not allow for the presence of mutual-priming derived chimeric read pairs, or the insertion mediated duplication event that occurs at point of the Alu insertion.

5.2.2.5. VariationHunter-CR

VariationHunter-CR is an algorithm to detect Alu insertions using repeat-anchored mapping[93]. Repeat-anchored mapping in VariationHunter-CR works by constructing an artificial reference chromosome from the consensus sequences of known mobile element families. A genomic mapping is performed using mrsFAST[94], a sequence alignment tool that maps reads to all matching genomic locations. Discordant read pairs where one end read maps to the genome, and the other read maps to the artificial chromosome are evidence for an insertion of a repetitive element. Clustering of these read pairs is performed to identify insertion sites of repetitive elements. The region of the artificial chromosome to which the reads maps identifies the particular repetitive element inserted.

VariationHunter-CR relies on mobile element sequence being the content of discordant reads that fail to map, however the insertion event in an exome capture cause the creation of fragments that do not contain the insert due to mutual priming. The tool also fails to account for the affects of the insertion-mediated duplication present in the insertion of transposable elements.

5.2.3. Software and File Formats

5.2.3.1. BAM file format

The sequence alignment/map (SAM) file format is used to store sequence data and alignment information for short read sequence, including exome sequencing experiments[10]. The BAM file format is a compressed, binary SAM file.

5.2.3.2. Burrows-Wheeler Aligner

The Burrows-Wheeler Aligner (BWA) is a next-generation sequence alignment program that uses the Burrows-Wheeler transform to align paired-end short reads to the genome[8]. When aligning paired-end reads, each read is independently aligned to the genome. If one read maps and the other does not, a Smith-Waterman alignment is performed to rescue these unmapped reads. Because Smith-Waterman is a local alignment, reads that were initially unmapped will be trimmed in the output BAM file to only the subsequence aligned to the area surrounding the properly mapped mate.

5.2.3.3. GATK Framework

A proof-of-principal algorithm has been implemented inside the Genome Analysis Tool Kit (GATK) framework[62]. The GATK framework provides programmatic interfaces to BAM formatted alignment files, reference files, and a convenient mechanism to perform computations in parallel. Because other tools implemented inside the GATK framework are commonly used in the detection of SNVs and small insertions and deletions in exome data, it is convenient to incorporate the tool into existing sequencing workflows.

5.2.3.4. Integrative Genomics Viewer

The Integrative Genomics Viewer (IGV) is a visualization tool for inspecting reads from next-generation sequencing projects in addition to genome annotation tracks[95].

5.2.4. Figure Conventions Used In This Chapter

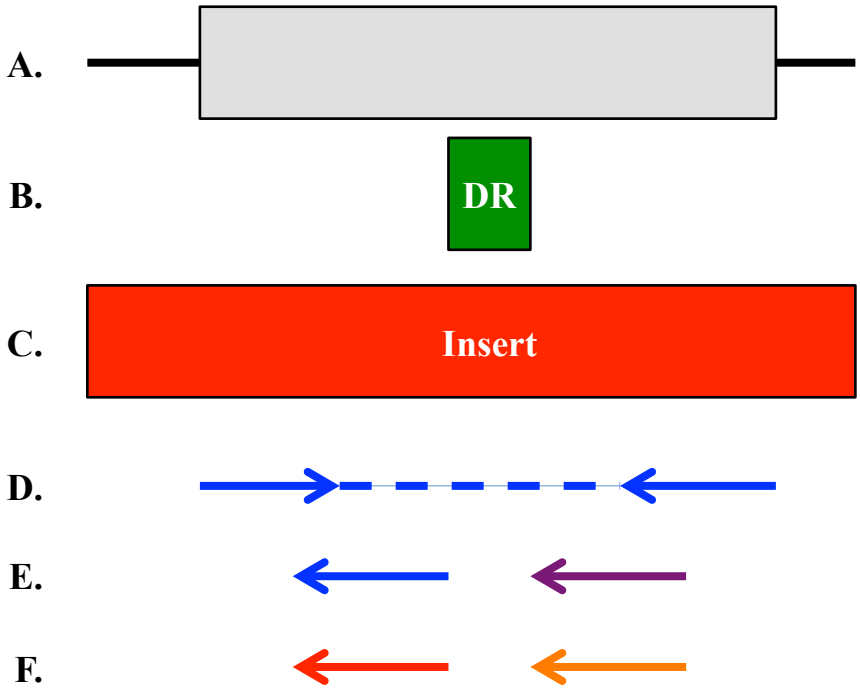

Figure 16. Conventions used in figures in this chapter to describe characteristic sequence anomalies.
(A). The symbol used to describe an exon.
(B). The symbol used to describe an insertion-mediated duplicated region.
(C). The symbol showing the repetitive element insert.
(D). A paired-end sequencing read pair. An arrow denotes each read with the sequencing insert shown with a dashed line.
(E). By convention, reads or portions of reads that map properly to the genome are shown in blue and purple.
(F). By convention, reads that would fall within the insert are shown in red and orange.

Figures in this chapter use common visual elements to describe characteristic alignment anomalies that occur near insertion breakpoints. Some of the insertion scenarios have sufficient complexity making it difficult to depict with figure, so for clarity these conventions are illustrated in Figure 16. Exons are denoted as a grey box on a black bar. A green box labeled “DR” indicates the insertion-mediated duplicated

regions. Inserted repetitive elements are illustrated using a red box labeled "Insert." Paired-end sequencing read pairs are shown as arrows with the origin at the start of the sequencing read and the tip at the end of the read. The fragment sequence between the reads, the sequencing insert, is shown with a dashed line. Blue and purple arrows indicate reads that map properly to the genome, red and orange arrows indicate reads that would fall within the repetitive element insert.

5.3. Approach

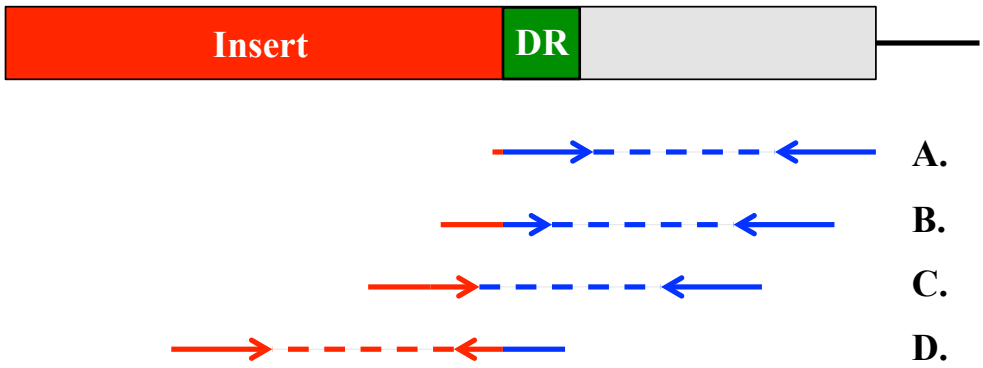

Figure 17. Position of reads relative to the insertion breakpoint causes a variety of characteristic anomalies in the sequencing data.
(A). The read pair has only a few bases of overlap and produces artifactual mismatches.
(B). The read pair initially fails to map and produces trimming upon local alignment.
(C). The read within the insert from this pair fails to map, and leads to a discordant read pair.
(D). Both reads fail to map, this will lead to a reduction of coverage.

At the sites of insertions of repetitive sequence, standard genomic alignment tools like BWA fail in predictable ways. I will take advantage of these characteristic alignment anomalies to detect these insertion events. I have identified five major types of anomalies at the insert sites: 1) reads similar to those in Figure 17, case C causes an overabundance

of discordant reads pairs where the discordant reads are mapped to many different regions of the genome, 2) 5' read trimming occurring at the insertion breakpoint in reads similar to those shown in Figure 17 case B, 3) similar sequence trimmed at insertion breakpoint, 4) the loss of mapping of fragments similar to those in Figure 17 case D causes decreased coverage relative to a sample lacking the insertion, and 5) low quality SNVs occurring within several base pairs of the insertion breakpoint in reads similar to those shown in Figure 17 case A. Because all of these features can be calculated on a single pass through of the aligned sequence, a tool based on this approach will be computationally efficient both in terms of CPU – O(n) where n is the length of the exome, and memory – O(m) where m is the size of the window where discordant pairs are considered (several hundred base pairs).

5.3.1. Low Quality Variants near Insertion Breakpoints

The characteristic anomalies that led to the discovery of the Alu insertion in the original MAK patient were several low quality variants near the site of the insertion. As depicted by Figure 18, reads that overlap a breakpoint of the insertion by only a few bases will still align to the genome. Mismatches and short insertions/deletions can be called at the 5' extreme end of this read. These variants will be of low quality because of the small proportion of reads that will only overlap by a few bases.

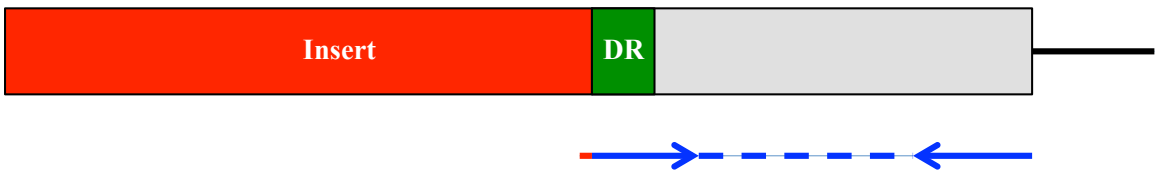

Figure 18. Low quality variants near the insertion breakpoint. When 1-2 bases of a read overlap the insertion breakpoint, mismatches can be called. This leads to low quality mismatches surrounding the insertion breakpoints.

5.3.2. 5’ Trimming at Insertion Breakpoints

The BWA algorithm is based on a seed-and-extend alignment strategy, by default the 32 5’ most bases are used as a seed, where fewer variations are allowed from the reference genome. If more than a few of these seed bases are in the inserted sequence, as depicted in Figure 19, the read will not be able to align to the reference genome. BWA, and other alignment algorithms, attempt to rescue these reads by performing a Smith-Waterman alignment of the read to the region downstream of its mapped pair. In the case where the unmapped read overlaps the insertion breakpoint this high-fidelity rescue alignment will align the 3’ portion of the read to the reference genome and trim the remaining 5’ portion of the read. Because Smith-Waterman alignments were used to perform this trimming, locations of frequent trimming can identify these sites with high fidelity.

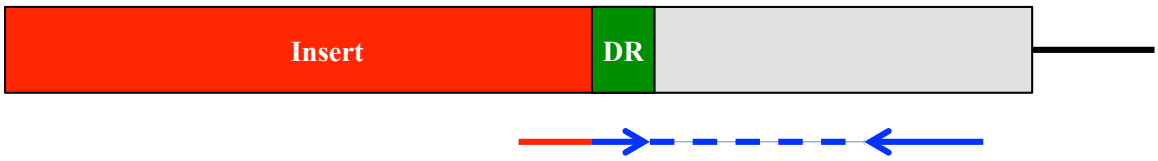

Figure 19. 5’ trimming at the insertion breakpoints. When a significant portion of the read overlaps a breakpoint of the insertion (red), a read will fail to align. If the paired read (blue) aligns properly nearby, the read can be rescued by a local alignment performed by BWA. This leads to trimming of the 5’ end of the originally unmapped read.

In the case of exome sequencing, this effect may not be symmetrical around the site of the insertion. When an insertion occurs at the edge of an exon there may not be targeting oligonucleotides to capture the fragments on one side of the duplicated region. Therefore, differences in the number of trimming events that occur on each side of the insert should be ignored.

The content of the trimmed sequence can also be informative. Since the sequence is removed by this trimming is sequence from the insert, it should therefore be similar between reads. This is an important characteristic in distinguishing true inserts from background regions because poor quality reads that were trimmed would be expected to contain random sequence. Additionally if the insertion is a retrotransposon, a Poly-A tail will be present.

5.3.3. Discordant Read Pairs as a Signature of Insertion

When one read falls fully inside the insertion and its pair outside as shown in Figure 20 a discordant read pair could be formed. A discordant read pair is where the ends of the fragment map to different loci. These reads can either be interpreted as either an artifact of the library preparation where small fragments with some sequence similarity are able to mutually prime and create a biologically meaningless fragment, or as evidence of a structural variation. Some references in the literature define a discordant read pair as a read pair with a longer than normal insert size[96], these can be evidence of deletions, but here the term is more broadly defined to mean a pair of reads mapping to two distant locations.

At the site of an insertion of repetitive sequence, there is an abundance of discordant read pairs, where one end of the fragment maps to the genomic area surrounding the insertion, and the other maps to another occurrence of the repetitive element in the genome. For an element such as an Alu that is present in very high copy, this will result in discordant read pairs to many chromosomes.

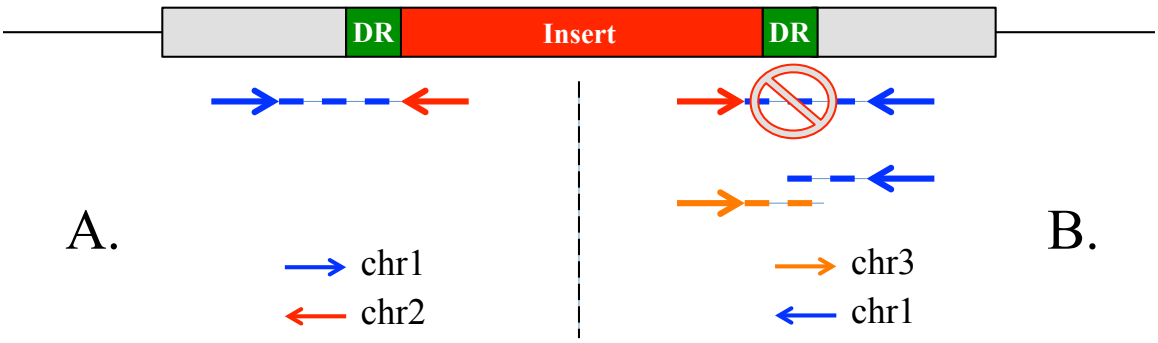

Figure 20. Discordant read pairs occur at the site of the retrotransposon insertion when one read falls within the insert, and the other falls in genomic sequence. The content of these discordant pairs can arise from two different sources.
(A). The insert (here into chr1) shares homology with a region elsewhere in the genome (for example chr2). In this case, while the paired ends of the fragment apparently map to different chromosomes, the fragment represents a true portion of the genome.
(B). A portion of the fragment contains an Alu insert, and the library preparation procedure blocks repetitive elements. This can lead to the artificial enrichment of fragments formed through mutual priming with another region of the genome, for instance chr3.

The model would predict that the average distance of the local reads from the insertion point is a function of the insert length, and the orientation of the local reads should place the discordant pair in the insertion. The model should not expect to see discordant pairs from both sides of the insert as the insertion itself can be expected to disrupt the binding of complementary probes used in the exome library preparation. In whole-genome sequencing this bias would not be expected.

In addition to discordant read pairs this same effect can manifest as a poor mapping of the paired read. This happens because local alignments are performed when a read initially fails to map, or if the read maps in an apparently erroneous way (discord). These forced local alignments would have a high frequency of variations, and an abundance of trimming on both the 5' and 3' ends. It is important to note that this mechanism is only possible if the mate of the poorly aligned read is properly aligned to the local chromosome to anchor the read.

5.3.4. Directionality of Sequence Overlap Features

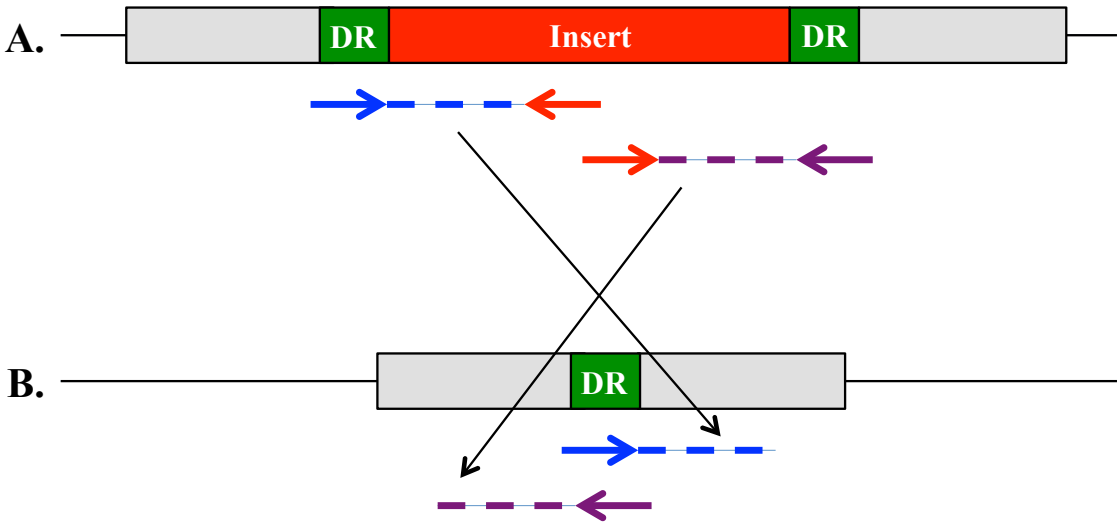

Figure 21. The duplicated genomic region (green) caused by the insertion of the retrotransposon (red) creates overlap in the discordant reads.
(A). This line depicts the region in the patient's genome. Note the insert flanked by insertion-mediated duplicated regions (DR)
(B). Line depicting the region when aligned to a reference genome that does not contain the insert. Note the position of the read pairs relative to the duplicated genomic region (green - DR). The red arrows are the mates that fall within the inserted sequence. When reads in this area are aligned to the genome, these insert reads will not align, leaving the discordant reads shown in blue and purple.

In addition, this discordant read pair effect is directional. As illustrated in Figure 21, because insertions by a reverse transcriptase are surrounded by a small region of duplicated genomic sequence, there will be a region where trimmed reads overlap and discordant read pair alignments overlap. In this figure, line A shows the expanded view of the insert in red, surrounded by the duplicated sequence in green. Line B shows this same region when aligned to the reference genome that does not contain the inserted sequence. Note how the blue and purple properly aligned reads overlap in the reference genome view.

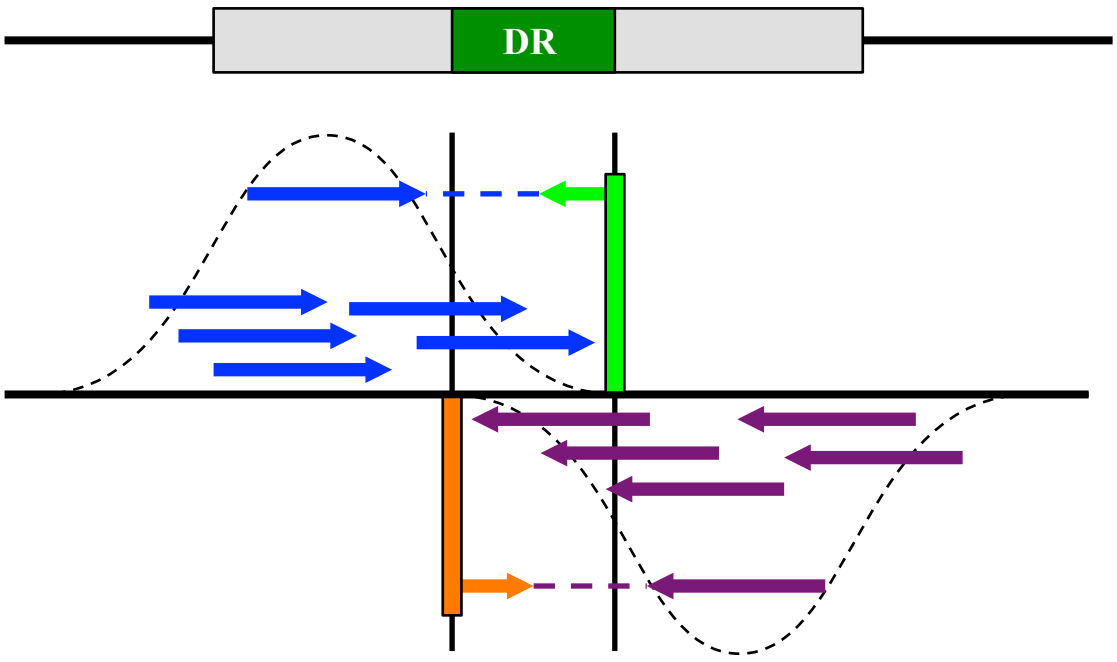

Figure 22. Trimming events and discordant read pairs will flank the duplicated region between the two vertical lines. Blue and purple arrows represent the discordant reads on the forward and reverse strands respectively. The green bar shows the location of forward strand trimming events, the orange bar is reverse. The green and orange arrows show the read pairs that lead to the 5' trimming events of the 5' and 3' ends of the insert respectively.

This same effect leads to directionality of the trimming events. Fragments that harbor trimming events 5' of the insert will have a trimming event 3' of the trimming event for the fragments 3' of the insert. This is depicted graphically in Figure 22 where the blue and green show the discordant read and trimming signature respectively for the forward strand. Figure 23 shows the duplicate region at the insert site in the MAK patient.

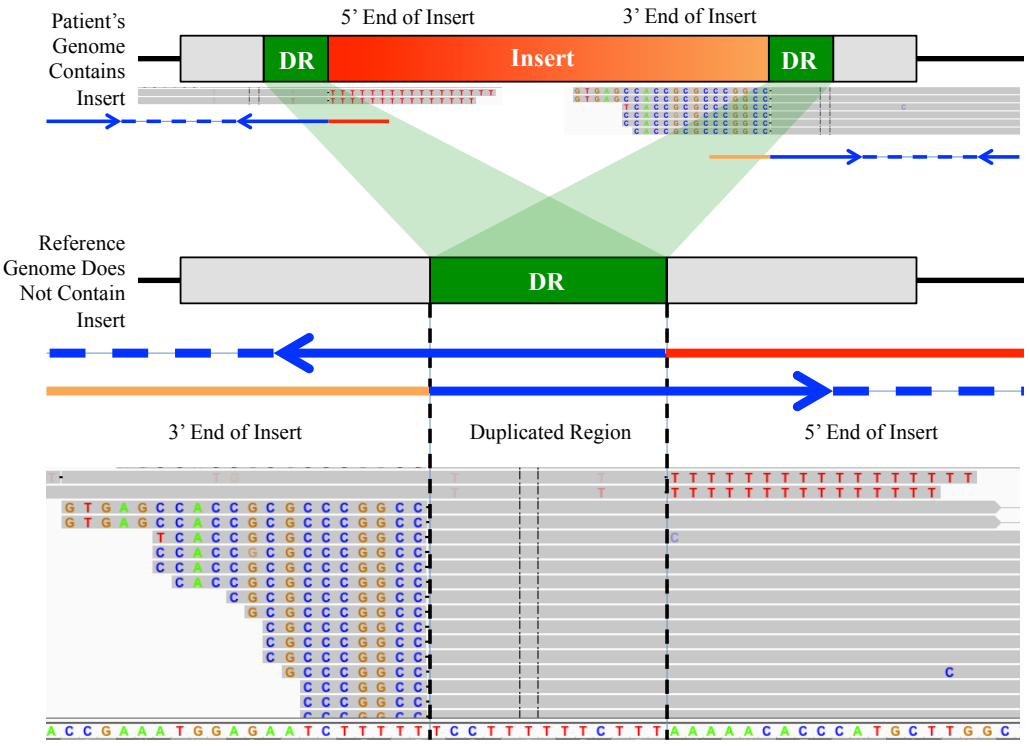

Figure 23. Screenshot from IGV showing the insert site of an Alu in MAK. The top of the figure shows the insert in the patient's genome, and the lower portion shows how the insert aligns to the reference genome. Bases that do not match the genome, and were trimmed from the alignments, are shown in the reads. The sequence from the 3' end of the insert is shown on left side of the duplicated region. On the right side of the duplicated region is the 5' end of the insert. The arrows depict paired end reads that give rise to trimmed sequence at the insertion breakpoints.

5.3.5. Decreased Coverage at Insertion Breakpoints

Coverage is reduced near insertions due to two effects. First, reads from fragments that overlap an insertion breakpoint similar to those shown schematically in red in Figure 24 can fail to map to the reference genome because the insert in which they fall is not contained in the reference genome. Second, the insertion (vertical green lines) can interrupt the region designed to hybridize with the capture oligonucleotides (purple line). This will lead to a reduced capture efficiency, and in the case of exons covered by a

single targeting oligonucleotide could lead to the complete loss of sequence from the exon.

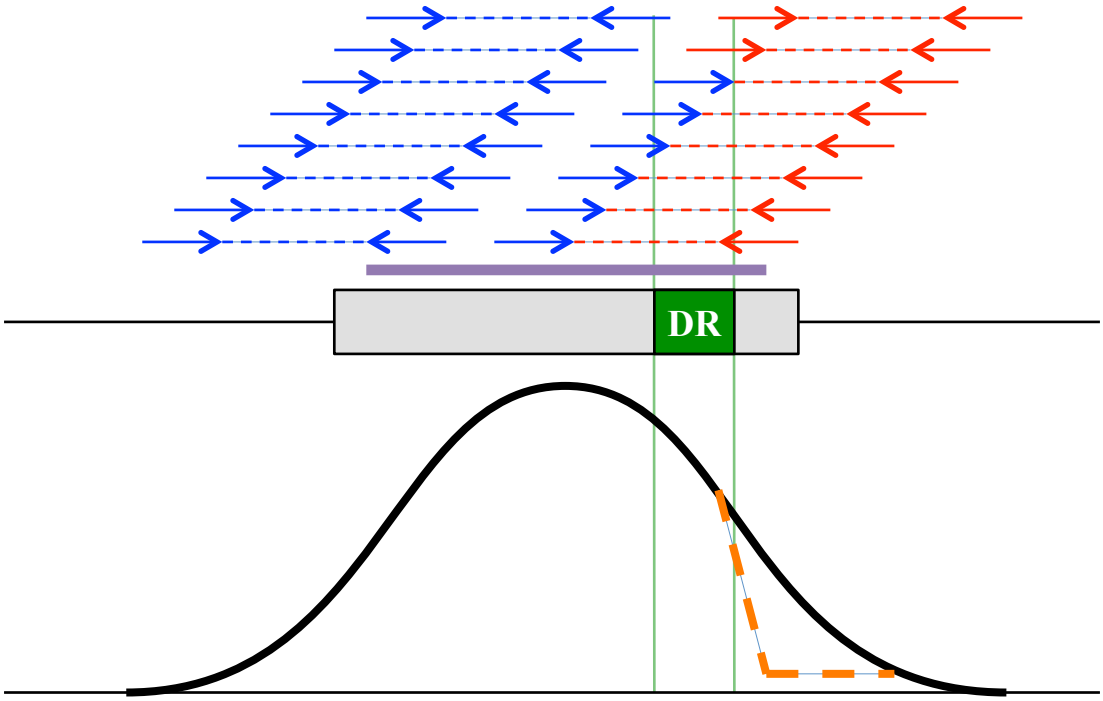

Figure 24. Upper: Reduced coverage at the insertion breakpoint in an exon. A portion of the reads that cross the insertion's duplicated region breakpoint (vertical green lines) will fail to map. Additionally the insertion can disrupt the hybridization of the capture oligonucleotide (purple line) further reducing coverage. Lower: A graph (black) of the expected coverage. These combined effects can cause a loss of coverage (orange dotted) over the insertion breakpoint. Trailing coverage after the breakpoint is due to miss mapping.

5.4. Methods

5.4.1. RIDE: Retrotransposon Insertion Detector for Exomes

A tool has been implemented that uses the discordant read pairs and 5' trimming anomalies to detect insertions. Figure 22 shows the assumed model of the sequence anomalies in the vicinity of the duplicated region (two vertical lines), consists of 5'

trimming events at the breakpoints (green and orange), surrounded by discordant read pairs oriented such that the discordant read would be in the insertion. As shown in Figure 25, reads aligning to the forward strand (blue) that overlap the region *–d* to *-b* are rewarded by factor *S*. Those that overlap *+b* to *+d* (these reads do not support an insertion at the current locus) are penalized by factor *P*. Negative strand discordant pairs (purple) are scored similarly according to their support of the hypothesis of an insertion at the current locus. Note the overlap *O* caused by the insertion-mediated duplication (DR), the typical length of *O* is 4-16 bases[85].

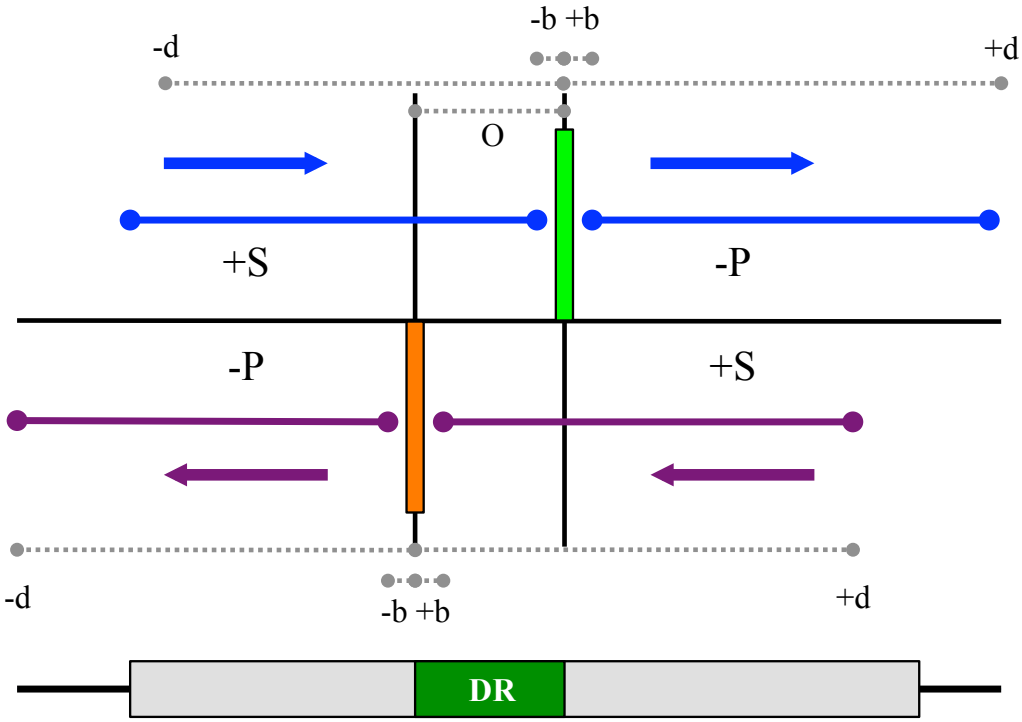

Figure 25. Assumed model of features around the sites of the insertion breakpoints that surround the duplicated sequence. Discordant reads within the S regions support an insertion and are rewarded, reads in the P regions are penalized.

This detection algorithm was implemented in the GATK framework[62] as a subclass of the LocusWalker class. A LocusWalker is a map-reduce based method of

accessing the data surrounding each genomic base. The trimming events at each base in each orientation are counted, and the presence of low-quality SNVs are assessed during the map step. During the reduce step regions containing both a forward and a reverse trimming event within the maximum overlap distance O and that exceed a threshold (default 20% of expected coverage) are identified. The reads in the surrounding region are then scored for the discordant read pair pattern. This evaluation of discordant read pairs is computationally expensive, but performing it only at the sites surrounding 5' trimming allows for practical application genome-wide.

5.4.2. Simulated Dataset

The approach outlined above was able to discover the homozygous Alu insertion in the MAK gene. However, since these insertions are rare events, there has only been a single case of discovering a disease-causing retrotransposon insertion via exome sequencing, a larger set of know events will be needed to validate the performance of the insertion detection tool. This can be partially simulated by removing annotated retrotransposons that happen to be near captured exons from the reference genome. Reads from a normal individual can then be aligned to this modified reference genome in an attempt to find the Alu elements. There are some limitations to this simulation. First, only homozygous insertions can be simulated. Second, the duplicated region cannot be simulated because of sequence divergence between the two copies present in the genome. Finally, it cannot simulate the affects of capture efficiency when a true insertion disrupts the hybridization of targeting oligonucleotides. Because of these limitations in the simulation, a version of RIDE was modified that that only uses discordant read pairs and trimming at the insertion breakpoints as features.

5.5. Results

5.5.1. Simulation Results

I carried out a simulation by removing AluY elements that were annotated to occur in the hg19 RepeatMasker track near exons from the reference genome. Sequence reads from several exome samples were then aligned to this modified reference. Possible insertions of repetitive sequence were called using a version of RIDE that only took into account soft-clipping and discordant pairs without allowing for an insert-mediated overlap. Sensitivity and the false discovery rate were calculated considering a prediction within 50bp of a simulated insertion to be a true positive, and an unpredicted Alu with coverage in the unaltered genome of 20x a false negative. RIDE achieved a sensitivity of 89.1% with a false discovery rate of 16.1%.

5.6. Discussion and Conclusion

False negative results in exome sequencing experiments are extremely costly, not only are the efforts for the experiment itself wasted, but also any subsequent validation experiments are also futile. False negatives in exome sequencing experiments occur when bioinformatic tools designed to detect common types of variants fail to detect a rare type of variant that causes a patient's disease.

Insertions of repetitive elements are one such class of variants that current tools fail to detect. RIDE is a novel bioinformatic tool designed to detect this rare variant type using the characteristic anomalies present in genomic alignments surrounding the site of the insertion. RIDE will close one analytical hole in current exome sequencing pipelines, and thus reduce the overall false negative rate of the experiment.

CHAPTER 6

CONCLUSION

Exome sequencing provides the ability to simultaneously sequence every known exon in the human genome. This technology has enabled the rapid discovery of many Mendelian disease genes that would not have been otherwise possible. Additionally, genetic testing in genetically heterogeneous diseases that used to require hundreds of individual tests to discover a patient's disease-causing mutations can now be performed in a single and standardized experiment.

Several bioinformatic tools have been presented here to aid in the accurate identification, interpretation, and validation of disease causing variants in exome sequencing experiments. First the Automated Sequence Analysis Pipeline (ASAP) provides automated workflows, variant calling, and variant annotation for the Sanger sequencing validation required when performing genetic testing. Second, the ASAP NGS annotation system aids in the interpretation of variants that affect amino acid sequence or splice sites. Third, tools and strategies to appropriately filter variants to reduce false positives due to systematic sequencing artifacts, variants inconsistent with disease segregation within a family, and variants that fall outside regions of autozygosity in consanguineous lineages were presented. Finally, a Retrotransposon Insertion Detector for Exomes (RIDE) was developed that allows the detection of a class of variants that were previously undetectable using existing tools.

Until the error rates of next-generation sequencing-based experiments decreases, there will still be a need to perform confirmatory Sanger sequencing of identified variants. This confirmatory sequencing is still necessary because genetic testing results can only be used to determine prognosis, treatment options, and family planning decisions if they are based on accurate data. ASAP provides the tools necessary to efficiently perform this confirmatory sequencing. This includes automated base calling,

assembly, alignment and variation calling and variant annotation in addition to tools to manage files and results.

Interpreting the many variants discovered by next-generation sequencing experiments can be one of the most difficult challenges of this work. This interpretation requires both the correct annotation of variants and appropriate filters to reduce the false positive rates of these experiments. The ASAP NGS annotation system is a tool that can help interpret variants in coding exons by predicting the effect a genomic change would have on the amino acid sequence. To interpret coding and non-coding variants also requires separating plausible disease-causing variants from the many false positives produced by these experiments. The systematic error catalog and the filtering based on consistency with disease segregation can help provide this separation.

Ultimately, all of the best variant annotation, filtering and validation efforts are wasted if the variant detection algorithms do not identify the disease-causing mutations. RIDE, a Retrotransposon Detector for Exomes is a novel bioinformatic tool to close one such analytical shortfall. By using the characteristic anomalies present in the genomic sequence alignments of exomes, RIDE can detect the rare retrotransposon insertion events that can be major causes of Mendelian disease.

While these tools enable efficient genetic testing using exome sequencing, much work remains. As genetic testing moves from Sanger sequencing to exome sequencing to whole genome sequencing, an increasing number of variants are discovered, and the interpretation of these variants becomes more challenging. Moving forward, the interpretation of variants in introns, UTR, and intergenic space will become increasingly important. While efforts like ENCODE are providing invaluable knowledge of the normal function of these regions, the affects of variations on the function of these regulatory regions is still poorly understood. Genome-scale studies of gene expression, alternative splicing, and transcription factor binding in affected patients will be

instrumental in building the tools necessary to predict the impact of these non-coding changes.

REFERENCES

1. DeLuca, A. P. ASAP-an Automated Sequence Analysis Pipeline for Clinical Genetic Testing. (2008).

2. DeLuca, A. P. *et al.* Sequencing and Disease Variation Detection Tools and Techniques. *Computer Systems and Applications (AICCSA), 2011 9th IEEE/ACS International Conference on* 80–83 (2011). doi:10.1109/AICCSA.2011.6126607

3. Tucker, B. A. *et al.* Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa. *Proc Natl Acad Sci USA* **108,** E569–76 (2011).

4. Rothberg, J. M. & Leamon, J. H. The development and impact of 454 sequencing. *Nat Biotechnol* **26,** 1117–1124 (2008).

5. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26,** 1135–1145 (2008).

6. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39,** 1522–1527 (2007).

7. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27,** 182–189 (2009).

8. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

9. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).

10. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

11. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

12. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

13. dbSNP: the NCBI database of genetic variation. 1–4 (2000).

14. Stone, E. Leber Congenital Amaurosis–A Model for Efficient Genetic Testing of Heterogeneous Disorders: LXIV Edward Jackson Memorial Lecture. *Am J Ophthalmol* (2007).

15. Antonarakis, S. E. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum Mutat* **11,** 1–3 (1998).

16. Kent, W. BLAT-the BLAST-like alignment tool. *Genome Res* **12,** 656–664 (2002).

17. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8,** 175–185 (1998).

18. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8,** 186–194 (1998).

19. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res* **8,** 195–202 (1998).

20. Bhangale, T. R., Stephens, M. & Nickerson, D. A. Automating resequencing-based detection of insertion-deletion polymorphisms. *Nat Genet* **38,** 1457–1462 (2006).

21. Nickerson, D. A., Tobe, V. O. & Taylor, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **25,** 2745–2751 (1997).

22. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41,** D64–D69 (2012).

23. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37,** D32–D36 (2009).

24. Shearer, A. E. *et al.* Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. *Proceedings of the National Academy of Sciences* **107,** 21104–21109 (2010).

25. Hildebrand, M. S. *et al.* Audioprofile-directed screening identifies novel mutations in KCNQ4 causing hearing loss at the DFNA2 locus. *Genet. Med.* **10,** 797–804 (2008).

26. Hildebrand, M., DeLuca, A., Taylor, K. & et al. AudioGene Audioprofiling: A Machine-based Candidate Gene Prediction …. *The* ... (2009).

27. Taylor, K. R. *et al.* AudioGene: Predicting Hearing Loss Genotypes from Phenotypes to Guide Genetic Screening. *Hum Mutat* (2012). doi:10.1002/humu.22268

28. Smith, K. R. *et al.* Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biology* **12,** R85 (2011).

29. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164–e164 (2010).

30. Ng, P. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31,** 3812 (2003).

31. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genom. Human Genet.* **7,** 61–80 (2006).

32. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1081 (2009).

33. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12,** 436–446 (2002).

34. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Meth* **7,** 248–249 (2010).

35. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19,** 1553–1561 (2009).

36. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth* **7,** 575–576 (2010).

37. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology* **8,** R232 (2007).

38. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39,** e118–e118 (2011).

39. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32,** 894–899 (2011).

40. Makarov, V. *et al.* AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics* **28,** 724–725 (2012).

41. Garla, V., Kong, Y., Szpakowski, S. & Krauthammer, M. MU2A--reconciling the genome and transcriptome to determine the effects of base substitutions. *Bioinformatics* **27,** 416–418 (2011).

42. Shetty, A. *et al.* SeqAnt: A web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics* **11,** 471 (2010).

43. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *fly* **6,** 80–92 (2012).

44. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19,** Unit 19.10.1–21 (2010).

45. Giardine, B. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res* **15,** 1451–1455 (2005).

46. Ge, D. *et al.* SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics* **27,** 1998–2000 (2011).

47. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* **21,** 1529–1542 (2011).

48. Rope, A. F. *et al.* Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am. J. Hum. Genet.* **89,** 28–43 (2011).

49. Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28,** 2267–2269 (2012).

50. Medina, I. *et al.* VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Res* **40,** W54–W58 (2012).

51. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

52. Durbin, R. & Haussler, D. GFF (General Feature Format) specifications document. *http://www. sanger. ac. uk/Software/formats/GFF/GFF_Spec. shtml.]* (2000).

53. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).

54. Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4,** e7767 (2009).

55. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32,** 493D–496 (2004).

56. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89,** 10915–10919 (1992).

57. Zheng, J. *et al.* Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with alpha-tectorin and is mutated in autosomal dominant hearing loss (DFNA4). *Proceedings of the National Academy of Sciences* **108,** 4218–4223 (2011).

58. Mahajan, V. B. *et al.* Calpain-5 Mutations Cause Autoimmune Uveitis, Retinal Neovascularization, and Photoreceptor Degeneration. *PLoS Genet* **8,** e1003001 (2012).

59. Stone, E. M. *et al.* Autosomal Recessive Retinitis Pigmentosa Caused by Mutations in the MAK Gene. *Invest Ophthalmol Vis Sci* **52,** 9665–9673 (2011).

60. Züchner, S. *et al.* Whole-Exome Sequencing Links a Variant in DHDDS to Retinitis Pigmentosa. *The American Journal of Human Genetics* **88,** 201–206 (2011).

61. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236,** 1567–1570 (1987).

62. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–1303 (2010).

63. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* **21,** 952–960 (2011).

64. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).

65. Johnson, J. O. *et al.* Exome sequencing reveals riboflavin transporter mutations as a cause of motor neuron disease. *Brain* **135,** 2875–2882 (2012).

66. Schossig, A. *et al.* Mutations in ROGDI Cause KohlschUtter-TOnz Syndrome. *The American Journal of Human Genetics* **90,** 701–707 (2012).

67. Shamseldin, H. E., Elfaki, M. & Alkuraya, F. S. Exome sequencing reveals a novel Fanconi group defined by XRCC2 mutation. *Journal of medical genetics* **49,** 184–186 (2012).

68. Sirmaci, A., Edwards, Y. J. K., Akay, H. & Tekin, M. Challenges in Whole Exome Sequencing: An Example from Hereditary Deafness. *PLoS ONE* **7,** e32000 (2012).

69. Puffenberger, E. G. *et al.* Genetic Mapping and Exome Sequencing Identify Variants Associated with Five Novel Diseases. *PLoS ONE* **7,** e28936 (2012).

70. Topaloglu, A. K. *et al.* Inactivating KISS1 mutation and hypogonadotropic hypogonadism. *New England Journal of Medicine* **366,** 629–635 (2012).

71. Aldahmesh, M. A. *et al.* Identification of ADAMTS18 as a gene mutated in Knobloch syndrome. *Journal of medical genetics* **48,** 597–601 (2011).

72. Hanson, D. *et al.* Exome Sequencing Identifies CCDC8 Mutations in 3-M Syndrome, Suggesting that CCDC8 Contributes in a Pathway with CUL7 and OBSL1 to Control Human Growth. *The American Journal of Human Genetics* **89,** 148–153 (2011).

73. O'Sullivan, J. *et al.* Whole-Exome Sequencing Identifies FAM20A Mutations as a Cause of Amelogenesis Imperfecta and Gingival Hyperplasia Syndrome. *The American Journal of Human Genetics* **88,** 616–620 (2011).

74. Walsh, T. *et al.* Whole Exome Sequencing and Homozygosity Mapping Identify Mutation in the Cell Polarity Protein GPSM2 as the Cause of Nonsyndromic Hearing Loss DFNB82. *The American Journal of Human Genetics* **87,** 90–94 (2010).

75. Bischof, J. M. *et al.* Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum Mutat* **27,** 545–552 (2006).

76. Rödelsperger, C. *et al.* Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics* **27,** 829–836 (2011).

77. Krawitz, P. M. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* **42,** 827–829 (2010).

78. Carr, I. M. *et al.* Autozygosity Mapping with Exome Sequence Data. *Hum Mutat* **34,** 50–56 (2012).

79. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

80. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38,** 1767–1771 (2010).

81. Purcell, S., Neale, B., Todd-Brown, K. & Thomas, L. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of ...* (2007).

82. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27,** 718–719 (2011).

83. Vaché, C. *et al.* Usher syndrome type 2 caused by activation of an USH2A pseudoexon: Implications for diagnosis and therapy. *Hum Mutat* **33,** 104–108 (2011).

84. Deininger, P. L. & Batzer, M. A. Alu repeats and human disease. *Mol Genet Metab* **67,** 183–193 (1999).

85. Hancks, D. C. & Kazazian, H. H., Jr. Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development* 1–13 (2012). doi:10.1016/j.gde.2012.02.006

86. Hajirasouliha, I. *et al.* Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26,** 1277–1283 (2010).

87. Stewart, C. *et al.* A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genet* **7,** e1002236 (2011).

88. Sveinbjörnsson, J. I. & Halldórsson, B. V. PAIR: polymorphic Alu insertion recognition. *BMC Bioinformatics* **13 Suppl 6,** S7 (2012).

89. Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S. & Nagasaki, M. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* **12 Suppl 14,** S7 (2011).

90. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21,** 974–984 (2011).

91. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res* **19,** 1117–1123 (2009).

92. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).

93. Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26,** i350–i357 (2010).

94. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Meth* **7,** 576–577 (2010).

95. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* (2012). doi:10.1093/bib/bbs017

96. Zhang, J. & Wu, Y. SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data. *Bioinformatics* **27,** 3228–3234 (2011).