BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

SUMIT MIDDHA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DR CLAUDIA NEUHAUSER, ADVISOR

SEPTEMBER, 2014

Acknowledgements

I would like to acknowledge my thesis committee for the support and supervision throughout the process. This includes Dr Vipin Kumar and Dr Chad Myers along with Dr Steve Thibodeau who were instrumental in the success of my work. I would also take this opportunity to thank my colleagues and collaborators, many of whom are co-author on the papers that were published from this work.

Dedication

This thesis is dedicated to my wife Mridu whose unwavering love and support made this accomplishment possible.

Abstract

DNA sequencing as an application of Next Generation Sequencing (NGS) is beginning to reshape how physicians diagnose and make treatment decisions for their patients. These NGS technologies provide a great depth of information by bringing along unprecedented throughput of data, huge scalability and speed. The terabytes of data generated has precipitated a need for efficient bioinformatics analysis and interpretation processes. My dissertation provides an end-to-end solution to analyze DNA sequencing data, interpret and deliver results efficiently and effectively. I developed a modular, robust workflow Targeted RE-sequencing Annotation Tool (TREAT) to provide a backbone for NGS DNA analysis, in collaboration with Mayo Clinic's bioinformatics core [1]. TREAT is one of the first bioinformatics solutions to incorporate alignment, variant calling, annotation and visualization of DNA sequencing data. To better evaluate the increasing foray of NGS into the clinical domain, I designed a module for comprehensive depth of coverage evaluation for genes and variants of interest. This module extending upon the TREAT pipeline helps quantify the applicability of NGS for clinical gene panels [2]. With dwindling costs and increasing availability of whole genome sequencing, turnaround time remains a major factor for clinical adaptation of NGS. I developed a novel iterative bioinformatics approach to expedite whole genome analysis by focusing on clinically relevant genomic regions, reporting results in less than 10% of the original processing time [3]. Further research employing additional clinical annotation has given us insight into a comprehensive genotype phenotype correlation

evaluation of clinically reportable variants. Here I report on the characteristics of clinically relevant variants typically expected per individual from whole exome DNA sequencing data. These data highlight challenges that need to be addressed including both phenotype issues of disease penetrance and uncertainty about what is clinically reportable, and sequencing issues like incomplete sequencing coverage, thresholds for data filtering and lack of high quality databases to determine functional annotation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Next Generation DNA Sequencing

Know thyself has a whole new meaning from the standpoint of knowing one's entire DNA sequence for preventative and personalized medicine. DNA quantifies an individual's uniqueness and is useful for disease risk prediction and drug efficacy. Simply put, DNA sequencing is the determination of the string of nucleotides in a DNA molecule. Its knowledge is imperative for understanding a broad range of biological phenomena [5]. Sequence of DNA is useful in studying the genome itself, identifying genes, phenotypes and potential drug targets along with evolutionary history and metagenomic relationships. Identification of sequence differences, i.e. variants, is the starting point to elucidate the mechanism of disease and bring research into clinical practice.

In 2001, the Human Genome Project gave researchers a near complete map of a reference human genome revealing vast expanses of evolutionarily conserved and therefore functional yet novel DNA sequences [6-8]. After decades of Sanger sequencing based techniques for determining the DNA sequence [9], massively parallel sequencing

technologies have democratized the entire field of sequencing [10]. There are some basic differences in these two technologies to generate DNA sequencing data (**Figure 1.1**). The traditional methods of Sanger sequencing, though of better quality, remain time consuming and expensive. Next Generation Sequencing (NGS) on the other hand simultaneously sequences millions of DNA segments [11], each segment referred to as a 'read'. Each nucleotide in the sequenced region may be contained in multiple reads to repeatedly call the nucleotide providing confidence or higher quality in the read out. The sum of sequenced reads mapping at a nucleotide position is referred to as the 'coverage' of that nucleotide.

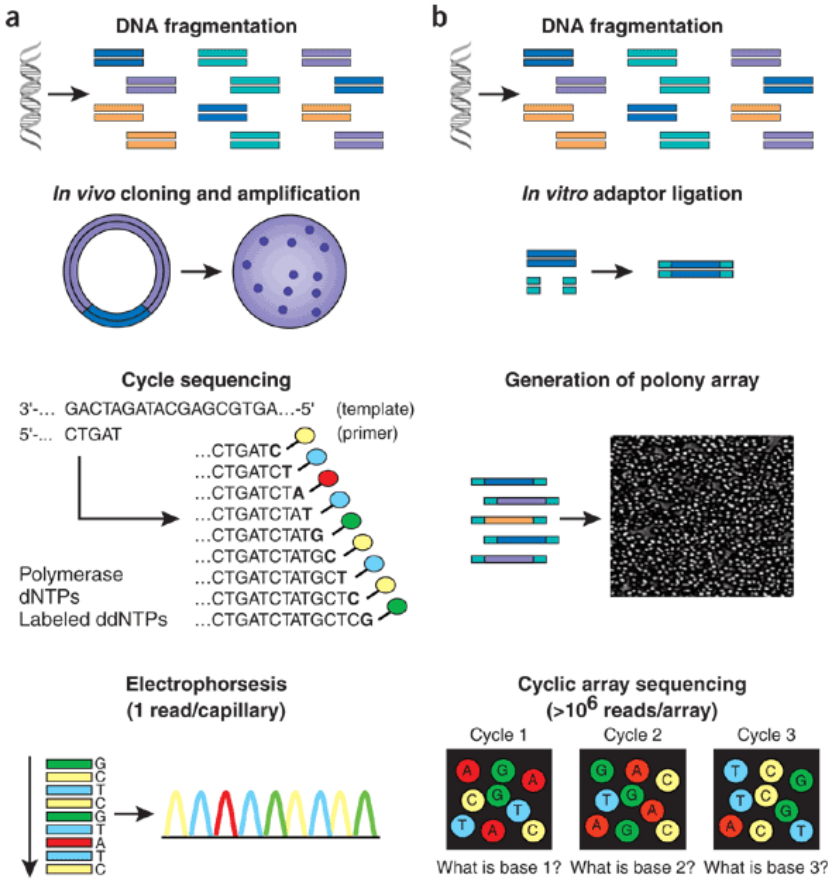

Figure 1.1: Conventional versus second-generation sequencing [12]. (a) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform E. coli. For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated. Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace. (b) In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or 'polonies' [13]. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature. (Permissions to reproduce figure and label obtained from Nature Publishing Group – Appendix A.4)

The major sequencing technologies to gain prominence in the last decade are usually lumped into the category of NGS, sometimes referred as second-generation sequencing. These include sequencing by hybridization [14], single molecule sequencing [15, 16] and cyclic array sequencing [17]. The latter encompasses the popular commercial platforms of Roche 454, Illumina Solexa and Life Technologies SOLiD sequencing platforms. After the first human genome sequenced using traditional Sanger sequencing methods, a plethora of whole genomes were sequenced using these disruptive high throughput technologies [18-22].

DNA sequencing as an application of NGS is in the phase of being adapted towards realizing the goal of personalized medicine. It is gradually replacing the

traditional Sanger sequencing and becoming a robust technology of choice to find novel variants for human disorders [23]. An increasing number of papers have been published on the importance and wide applicability of DNA sequencing analysis in recent years. This includes large international projects like 1000 genomes [24], The Cancer Genome Atlas (TCGA) [25, 26], International Haplotype Mapping (HapMap) project [24], Beijing Genomics Institute (BGI) sequenced 200 Danish exomes [27] and NHLBI exomes [28] setup to study large groups of individuals. Large projects like Encyclopedia of DNA Elements (ENCODE) are leveraging the NGS technology to generate comprehensive maps of genomic and epigenomic elements in normal human and cancer cell lines [29]. The idea of a $1000 genome and ubiquitous access to whole genomes is a looming reality bringing with it many ethical and legal issues that need to be resolved [30]. Despite that, NGS technologies are bound to have a tremendous and far reaching impact on genomic research in the years to come [31].

Figure 1.2 gives a brief overview of the process of NGS experiments [32]. The double stranded DNA template is fragmented and sent to the sequencing reaction after a size selection step. The sequenced reads are then aligned or assembled based on the employed analytic technique.

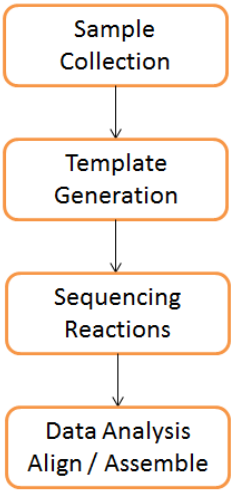

Figure 1.2: NGS workflow depiction of the four major phases carried out in order to generate the data

1.2 Whole Genome and Whole Exome Sequencing

Whole Genome Sequencing (WGS) is the read-out of the entire 3 billion nucleotides of an individual's DNA sequence to report potential variants, but is not always cost effective. Whole Exome Sequencing (WES) on the other hand involves enriching for and sequencing the entire coding DNA, approximately 1-2% of the whole genome and thus less expensive, has spurred a rush for clinical adaptation and commercialized services. Human exome has been shown to harbor more than 85% of the known disease causing mutations [23]. Many recent projects involving exome sequencing were a success because of WES being able to provide higher coverage of multiple samples at a more affordable cost to do mutation discovery ([33], [34], [35], [36], [37],

[38]). WES has also gained widespread application for diagnostic and clinical genetics along with characterizing various Mendelian disorders [39, 40].

With the plummeting costs, WGS is primed to gain wider usage. The biases inherent to the capture and library preparation process of WES may result in poorly covered regions leading to inadequate evaluation. It is also likely that many yet undiscovered disease-causing mutations are located not within the coding exon regions, but within intronic and intergenic regions. Many structural changes such as deletions, inversions and duplications, especially more than a few hundred nucleotides long, are more efficiently identified by WGS.

Essentially the choice of WGS or WES is a subjective matter as both NGS applications have context specific benefits depending on the genomic mechanism being studied. There have been high-profile reports of WES resolving incidental findings and revealing complexity missed by WGS evaluation [41], [42]. At the same time research has highlighted discovery of novel disease causing mutations from WGS that would have likely escaped detection from WES [43]. This dissertation and the identified solutions are applicable to both WGS and WES, providing timely and prioritized high quality results for clinical applicability.

From an analytic standpoint, the processing of both the datasets is similar as it begins with alignment of the short sequenced reads to the entire genome. The quality checks and technical challenges dealing with high error rate of sequencer, multiply mapped reads due to sequence homology or reference sequence bias in variant calling are much the same for WES and WGS [44]. WES does present additional challenges of uneven coverage (due to only the coding regions being targeted and sequenced) making identification of copy number and structural variation challenging. Majority of the bioinformatics workflows are cognizant of these similarities and differences between WES and WGS and are therefore designed to implement the common processes of alignment, re-alignment and variant calling before diversifying into specific modules for Copy Number Variants (CNV) and Structural Variants (SV) identification.

1.3 Workflow infrastructure for NGS analysis

The terabytes of data generated by NGS applications have precipitated the need for efficient bioinformatics processes to conduct high throughput genomic analysis and interpretation. A typical run of the current Illumina HiSeq instrument can generate more than 600 billion bases (Gb) of DNA throughput (Table 1.1). The voluminous and complex nature of NGS data has brought about an explosion of computational and statistical tools to analyze and interpret the sequencing data. These tools are designed for one of the 5 major steps of analyzing DNA sequencing data: 1) quality checks, filtering and basic throughput evaluation of the sequenced reads, 2) alignment of the raw reads to

the reference sequence, 3) identification of differences between the sample being analyzed and the standard reference to call variants, 4) annotation of those variants with known genomic feature information like intronic, stop-gain or missense and 5) visualization of the aligned reads and variants on the genomic scale.

List of NGS sequencing platforms and their expected throughputs, error types and error rates. Each platform has distinct advantages owing to cost, error rate, read length, and so on

Platform	Run time (h)	Read length (bp)	Throughput per run (Mb)	Error type	Error rate (%)
Roche					
454 FLX+	18–20	700	900	Indel	1
454 FLX Titanium	10	400	500	Indel	1
454 GS	10	400	50	Indel	1
Illumina					
GAIIx	14	2 × 150	96,000	Substitution	>0.1
HiSeq 2000	8	2 × 100	400,000	Substitution	>0.1
HiSeq 2000 V3	10	2 × 150	<600,000	Substitution	>0.1
MiSeq	1	2 × 150	1000	Substitution	>0.1
Life technologies					
SOLiD 4	12	50 × 35	71,000	A-T Bias	>0.06
SOLiD 5500xl	8	75 × 35 PE 60 × 60 MP	155,000	A-T Bias	>0.01
Ion torrent					
PGM 314 Chip	3	100	10	Indel	1
PGM 316 Chip	3	100+	100	Indel	1
PGM 318 Chip	3	200	1000	Indel	1
Pacific biosciences					
RS	14/8 Smart Cells	1500	45/SC	Insertions	15

Table 1.1: List of NGS sequencing platforms and their expected throughputs, error types and error rates [45]. (Permissions to reproduce obtained from Elsevier – Appendix A.5)

Presently, there are more than 60 aligners available for mapping the high-throughput of short reads to a reference sequence [46]. This has led to the arduous task of selecting the most appropriate aligner for a particular application [47]. There is an even greater number of genotype calling tools that use base quality, mapping quality, coverage depth, model of known variants and Linkage Disequilibrium among other features [48].

The burgeoning of these tools has been accompanied with a lack of good benchmarking and a lack of automated workflows utilizing these tools for routine analysis. The very popular view of a $1000 genome, but $10000 analysis [49] has been largely true due to the huge informatics challenges faced by research labs. The need for closing the gap between high throughput data generation and the ability to process, analyze and navigate through the resulting data has been highlighted by multiple groups [50].

For a typical NGS analysis, 5-8 or even more separate bioinformatics tools might be required to QC the data, map reads, identify sequence variants, annotate using local or public databases and visualize the data. Usually these tools need to be installed and configured in a way such that NGS data can be sequentially analyzed through the process in an automated fashion (bioinformatics workflow). However this does require end users to download the individual tools, any associated dependencies, files and databases. The dependencies and associated files need to be formatted appropriately to ensure compatibility with the mosaic set of tools. Another challenge with using a set of tools is input output format inconsistencies and the requirement of data parsing to ensure proper functioning of the automated workflow as a whole. The tools have a variety of parameters and features for flexibility of usage in different scenarios but the end users need to understand the impact and tweak the settings for their custom setup. The tools are being actively developed leading to new releases and versions that require a constant commitment to keep the workflow up-to-date and efficiently functional.

As detailed in the later chapters, by overcoming a lot of these hurdles our Targeted RE-sequencing Annotation Tool (TREAT) workflow [1] fills the need for a robust integrated bioinformatics framework. As TREAT was being used other groups have also developed similar workflows, with many focusing on the downstream filtering and interpretation of human sequence variation data ([51], [52], [4], [53], [54], [55], [56], [57], [58], [59]). However, TREAT remains a well-suited application for researchers, labs and programmers and has been adapted by organizations like University of Illinois Urbana Champaign and Appistry©. It has been adopted at Mayo Clinic for tens of thousands of DNA samples sequenced by various investigators and clinicians.

1.4 Depth of Coverage from NGS data

One of the most vital metrics in DNA sequencing is depth of coverage. It is defined as the number of times a particular nucleotide is evaluated, that is, the number of reads mapping to that nucleotide location. Coverage is reported as an average coverage of a region based on the total number of nucleotides sequenced and region length, or as a percentage of the region covered by at-least a threshold number of reads. Higher coverage ensures greater accuracy but also entails higher cost of data generation implying a trade-off decision that is made during the study design phase of the project. Much higher average sequencing depth is necessary to achieve accurate variant calling over the entire genome/exome [60].

As NGS is rapidly moving towards clinical application, its usage to replace an existing Sanger sequencing based gene-panel test requires a comprehensive read coverage analysis. It has been reported that some regulatory regions consistently have lower coverage from NGS data leading to poorer variant detection ([61], [62]). Sequencing costs often limit the amount of sequence reads generated, limiting the coverage and quality of the data. There is a justifiable need to report coverage before claiming genome-wide or exome-wide survey for sequencing projects. This is especially necessary for studies that demonstrate applicability of NGS to replace gene-panel testing. For instance, the Dewey et al paper [63] highlights the inability of WGS to detect variants in vital inherited disease genes due to sub-optimal sequencing coverage. The obvious use-case is to evaluate the depth of coverage observed in regions or genes of interest from a pilot sequencing experiment. Such comprehensive analysis can potentially identify systematic lack of coverage in regions critical for the research or disease being studied. As an extension of the TREAT workflow, I developed an approach to perform an exhaustive evaluation of genes and known disease mutations of interest and report that for a sequencing experiment. This approach was developed in collaboration with Mayo Clinic's Neuropathy workgroup [2] and then also adopted by Colon cancer, Dilated Cardio-myopathy and Sudden Death research groups to report an extensive understanding of screening using NGS data and the potential for diagnosis.

1.5 Speed of WGS bioinformatics analysis for clinical application

A typical WGS bioinformatics analysis workflow takes approximately 3-4 days per sample [64]. The rate at which newer datasets can be generated, analyzed and interpreted is limited less by the instrument time but more by the computational analysis [65]. It has been reported that the central processing unit (CPU) time required on a single 2.1 GHz processor for a single WGS analysis would be about 1701 hours or 0.20 years [66]. Newer methods that rely on specialized and expensive hardware to provide faster solutions are also limited to particular sequencing platform ([66, 67]). Meanwhile, as the cost of WGS is decreasing, clinical laboratories are looking at broadly adopting this technology to screen for variants of clinical significance. To completely leverage this technology in a clinical setting, results need to be reported rapidly, as the turnaround rate could potentially impact patient care.

From a diagnostic and personalized medicine perspective, a very limited number of genomic regions are clinically reportable and actionable. Depending on the computing infrastructure available, the processing of WGS data can take several days, with the majority of computing time devoted to aligning reads to genomics regions that are to date not clinically interpretable [3]. Currently, most of the clinically relevant genomics information is related to protein-coding exon regions where the impact of coding variants can be interpreted in the context of proteins and their function [3]. This focus provides an

opportunity to develop new bioinformatics algorithms that prioritize and swiftly report clinically relevant findings [3]. To expedite the delivery of most clinically relevant results from WGS data, we developed a platform that is sequencing technology and hardware independent. This recent work reports all the clinically actionable variants from WGS data in a matter of 5 hours, only $1/15^{th}$ of the original analytic time [3].

1.6 Clinical filtering and interpretation of NGS data

Even with the high accuracy of NGS data, the enormous throughput implies that a low error rate of 0.1% would still introduce 3 million erroneous base calls within the 3 billion long human genome sequence. Identification and removal of systematic errors of a particular technology can help reduce such false positives. The bioinformatics workflows introduced in Section 1.3 have their own inherent biases due to the tools and parameters or thresholds used for the automated analysis. Despite the adoption of Illumina© as a common sequencing platform, there is wide diversity to the annotation and reports provided as non-standardized passive decision support in electronic health records [68]. From a clinical reporting perspective, a vast majority of the variants called from NGS data are VUS (Variants of Unknown Significance) [69]. VUS could be due to incomplete or missing information on the gene function or lack of knowledge about the effect of the variant on the protein translated from the gene. The VUS are generally confusing for the individuals and their families and challenging for the clinicians.

The abovementioned reasons imply a need for careful and systematic filtering of variants generated from NGS data. There is ongoing debate on the subjective nature of such filtering and the scientific community has not yet settled on an accepted solution. Typically variants are filtered based on the mapping quality of reads at that position, the total number of independent reads showing the variant and frequency of the variant in known repositories like HapMap [70], 1000 Genomes [24] and in-house databases.

A large variety of fragmented annotation sources and databases need to be setup and queried in order to extract out all known information about a variant or the function of a gene. The huge amount of data generated by NGS, although of great value, threatens to be a major obstacle with no readily available methods to accurately interpret all the results [71]. NGS has led to the uncovering of incidental findings that need to be carefully evaluated and in some cases reported back to the individuals. However, it requires multiple hours of manual curation by genetic counselors to refer relevant literature, access multiple fragmented databases and discuss with physicians to reach a decision to report or not report each individual variant [63]. It remains challenging to effectively interpret NGS data and the content of results reported to an individual.

1.7 Challenges and Motivation

Despite almost a decade of advances and successes from NGS DNA sequencing applications, there remain numerous challenges and deficiencies in the current understanding and interpretation of results. There is incomplete coverage of known disease genes, low reproducibility of genetic variation findings and uncertainty about the clinically reportable of findings ([63], [72],[73]) limiting clinical applicability. There are broadly four key challenges for clinical applicability of NGS sequencing that call for more effective and novel computational approaches:

1. **Workflow infrastructure for analysis**: NGS analysis, applications and results are still a moving target with ever-changing needs. It is almost always necessary to bring together a conglomeration of 5-8 or even more computational tools in order to begin to use the sequencing data. This entails installation and availability of IT resources along with knowledge of file formats and scripting in order to make the mosaic set of tools work together. The entire workflow requires sufficient flexibility to allow replacement of tools with newer better alternatives.

2. **Depth of coverage evaluation and reporting:** It is critical to evaluate the depth of coverage of genes, regions and nucleotide positions to ensure sensitivity from NGS DNA sequencing results. Lack of variant call for a particular sample could be due to the genetic variant not being present, or due to the lack of supporting sequence reads. Reporting the coverage information on a pilot set of data for select genes or regions of interest helps identify the

expected yield from NGS DNA sequencing before performing the actual experiment.

3. **Expedite analysis of clinical relevance:** The high throughput of data requires large amounts of computational time for efficient analysis. However, not the entire set of results is immediately pertinent for clinical evaluation, as the phenotypic implications of a large majority of genetic variation remain unknown. Prioritizing return of the most relevant results as the first iteration would be tremendously beneficial for bringing NGS applications into a clinic. This process needs to be implemented with no loss in quality and accuracy of the results.

4. **Clinical filtering and data interpretation**: Although a large percentage of genetic variants identified through NGS are accurate and have a high validation rate, there remains an uncertainty to the assignment of clinical significance to variants. Even before that, due to limited understanding of the deleterious effect of the variants, multiple iterations of variant filtering is commonly done to prioritize the variants important for clinical evaluation. There has been no concerted benchmarking of the yield from WES or WGS for the number and type of interpretable variants identified per individual. It remains unclear how many of the identified variants annotated by genetic counselors or physicians potentially cause a phenotype in the individual.

1.8 Outline of the chapters

To address these challenges, my dissertation entails the development of an end-to-end solution to analyze, interpret and deliver DNA sequencing data, efficiently and swiftly. In precisely the order of these challenges, the next four chapters elaborate on the four major contributions of this dissertation.

Chapter 2 introduces the Targeted RE-sequencing Annotation Tool (TREAT) workflow [1] and how it provides the backbone for DNA sequencing bioinformatics. Specifically, I provide details on the various modules of the TREAT workflow, selection of appropriate set of tools, flexibility of doing analysis from various starting points, efficient reporting and visualization of data and the Amazon EC2 cloud image for utilization by end-users who are not experienced in programming and software installation.

Chapter 3 addresses the critical aspect of sequencing coverage for NGS DNA sequencing studies. The module developed builds upon TREAT to report coverage depth information for clinically relevant genes. This involves a comprehensive sequencing depth coverage analysis of regions-of-interest for prompt data interpretation. I collaborated with Dr Christopher Klein (Associate Professor of Neurology) at Mayo Clinic and evaluated coverage for a set of known neuropathy genes and disease causing variants using exome-sequencing data from 24 samples with Inherited Peripheral

Neuropathy (IPN) [2]. This approach can be generalized to any set of genes or regions of interest and has been adopted by Cardio-myopathy and Sudden Death groups at Mayo Clinic for coverage analysis.

Chapter 4 pursues the challenge of computational time needed for bioinformatics analysis of Whole Genome Sequencing data and its impact on clinical applicability. To be an option of routine clinical diagnostic and screening usage, WGS requires faster turn-around of results. At the same time, only a very limited number of genomic regions are clinically reportable and actionable. From a diagnostic or screening standpoint, the analysis and results of only those relevant regions or genes would be of interest. I present an iterative approach for WGS bioinformatics focuses on such clinically relevant genomic regions and reports results in less than 5 hours [3]. It provides a huge improvement on a generic WGS bioinformatics analysis that takes about 75 hours to complete alignment and variant calling on WGS data. This speed-up is a critical step towards delivering results to the patient and potentially impacting course of treatment.

Chapter 5 explores the complexities in making the field of personalized genomics riding the wave of NGS a reality. The impact and significance of a large majority of genetic variants identified by NGS is currently unknown. Recent literature [63] highlights the large amount of time necessary for genetic counselors and physicians to research a variant before concluding on its appropriate pathogenecity. The scientific community is lacking an understanding of the number of variants to expect per individual

that warrant such rigorous interpretation by geneticists. Moreover, there has been no comprehensive evaluation of how the genetic results correlate with the medical phenotype of an individual. In order to develop capability of establishing NGS as a routine prognostic test for healthy individuals, it would be important to overcome these challenges.

Chapter 6 discusses the major implications of my dissertation including a summary of contributions, limitations and insights into possible future research investigations.

Chapter 2: Targeted RE-sequencing Annotation Tool (TREAT) [1]

2.1 Bioinformatics Workflow

Computational approaches for biological research need to link together meta-data information and the actual analysis in order to setup systematic workflows [74]. Such workflows involve analytic methods that use input files and tool definitions to analyze the data and compute the results to be presented in some output files. Such workflows have become a commonplace necessity for even small bioinformatics groups as they ensure standardization, automation, scalability and efficient personnel utilization. A person or team is usually involved in the nine major stages of workflow development and deployment:

1. Identifying the set of components or tasks needed for the entire analytic process
2. Researching the available tools for each of those tasks and evaluating and comparing the performance, accuracy, usability and other appropriate metrics

[1] Text and figures reproduced with permissions from Oxford University Press (Appendix Figure A.1)

3. Installing the set of tools identified in the previous step and their associated dependencies and ensuring optimum performance in the local computational environment
4. Analyzing sets of test or pilot datasets through the iterative process of the analyses identified in the first step and evaluating output results at various end points
5. Ensuring appropriate results from each of the steps and evaluating the overall performance
6. Programming custom software and parsing scripts to ensure compatibility of all the tools in order to be able to execute all the steps without interruption
7. Assembling the execution of all the tools into a single process, a conveyor belt of bioinformatics tools, in order to have the complete workflow and generate the end results
8. Analyzing real sized and complexity datasets using the workflow to evaluate performance, accuracy and efficient functioning
9. Further optimizing each of the tasks and the process as a whole using parallelization, latest version of the tools, efficient data formatting and lookup and other techniques

As is clear, this is a very tedious and complicated process, made even more difficult by the lack of documentation, benchmarking and robust test datasets. It also

requires extensive maintenance to keep up with the latest versions of the tools, format changes, bug fixes or parameter optimizations and updates to best practices in the general community based on newer evaluations.

2.3 Introduction to the TREAT Workflow

Biomedical research has been revolutionized by the emergence of NGS technology [75], which enables massive parallel sequencing of biological samples at highly informative depths [76, 77]. Technological innovations continue at an unprecedented rate, providing better chemistry for more accurate sequencing with ever-increasing throughput and reducing cost. Sequencing platforms offered by Roche, Illumina, Life Technologies, and others have expanded the utility of NGS beyond limited number of large genome centers. While WGS is still cost prohibitive for most investigators, targeted re-sequencing has gained vast popularity by focusing on genomic regions of interest. For instance, exome capture and sequencing or WES, which targets the coding regions of the genome, has facilitated the identification of both causal mutations in rare genetic diseases [78-81], and disease associated variants in complex diseases [82, 83].

Evolution of sequencing instrumentation is very dynamic leading to critical needs of standardized and centralized analytic expertise for medium to large bioinformatics groups [84]. Research groups have had short reaction time in which to develop processes,

best practices, rigorous QC/QA and automated management system environments to cope with the increasing demand due to NGS. There is growing demand to have the eventual data and results delivered from NGS analysis in more useful, interpretable formats rather than simply the alignments or lists of variants [84].

Data management and analysis also remain challenging for NGS projects due to the sheer volume of the data. Filtering the tens of thousands of identified single nucleotide variants (SNV) and small insertions/deletions (INDEL) down to a small number of disease relevant variants can be laborious and time consuming. To help mitigate the human effort, several annotation tools have been offered for the variants identified from the NGS data. ANNOVAR [85] was the first such tool published to functionally annotate genetic variants, which selects non-synonymous SNV or frame-shifting INDEL in conserved regions that are not reported in dbSNP or 1000 genome projects, and focuses on genes with multiple variants. Another tool, GAMES [86], annotates variants using UCSC annotations tables and KEGG pathways, and visualizes the variants using UCSC tracks. Other tools emphasize on the integration of a data management system into the annotation workflow, or the implementation of a web-interface, while providing similar annotations to the variant list [87, 88]. These open source tools were the first attempts to automate the variant filtering process for the targeted re-sequencing data.

Another stumbling block of NGS projects is the requirement of powerful computational infrastructure. For example, a single run from an Illumina HiSeq 2000 sequencer produces up to one billion 100 to 200 nucleotide long reads and the alignment of these reads alone takes days on a single Linux node using the fast alignment tool BWA [89]. Without access to advanced IT infrastructure, investigators face fiscal challenges and frustration with managing and processing of the profusion of data in order to extract biologically relevant information [90]. One solution to alleviate the infrastructure challenges is the recent development of Cloud Computing technology that can host analytic workflows for a nominal fee, eliminating the need to locally manage the complex infrastructure required to efficiently analyze NGS data.

As part of this dissertation, I collaborated with Mayo Clinic's bioinformatics core to develop an open source comprehensive workflow, Targeted RE-sequencing Annotation Tool (TREAT), for read alignment, genotype calling, annotation, and visualization of the targeted re-sequencing data. An Amazon Cloud Image of TREAT is also available for investigators without sufficient local IT and or bioinformatics infrastructure.

2.4 Methods

Website: The source code and executables (for both the parallelized and non-parallelized versions of the tool), a detailed user manual and other supplementary files

can be downloaded from http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/bioinformatics-software-packages. An instruction for installing the tool locally or launching the Amazon Cloud image is also available at the web site.

Gene Definition: The TREAT workflow uses human reference Genome Build 37 (hg19) and gene definitions from UCSC Genome Browser (*refFlat* file) by default. The gene definition from Build 36 (hg18) is also available.

Sequence Alignment and Variant Calling: The read qualities are examined by FastQC [91] which generates QC matrix from the FASTQ files including per-base sequence qualities, per-sequence quality scores, per-base nucleotide content, and sequence duplication levels. The FastQC tool also provides warnings for parameters failing to pass QC thresholds. The reads are aligned to the human reference genome and the mitochondrial genome using Burrows-Wheeler Aligner (BWA) [89]. This requires indexing the reference genome. If the sequence duplication levels failed to pass the FastQC threshold, the duplicated reads are removed using the SAMtools's *rmdup* method [92]. The BWA alignment is then locally re-aligned using the Genome Analysis ToolKit (GATK) [93, 94] in order to correct mis-alignments due to the presence of INDEL. SNV are called using SNVMix [95] with a cut-off probability score of 0.8 based on our preliminary testing using a HapMap CEPH subject sequenced by the 1000 genome project, and INDEL are called by GATK with default parameters setting.

Functional Prediction Tools: Two open-source annotation tools are included in TREAT. SIFT [96] is a sequence homology based tool that predicts whether amino acid substitutions in protein have phenotypic effects. It uses multiple alignment information to predict tolerated and deleterious (affecting protein function) substitution at a given variant position. SeattleSeq [97] is a variant annotation tool based on GVS (Genome Variation Server) database, which contains 4.5 million variants with the corresponding genotype data. Its annotations include gene/transcript information along with predicted functional consequence of the input variant.

Variant Annotation: The read depths of each of the A, C, G, T bases at each variant position, as well as the average mapping quality score are provided by curating the BAM pile-up files using SAMtools [92]. If an identified SNV is a known variant from dbSNP [98] or 1000 Genome Project [24], the allele frequencies of Caucasian (CEU), Yoruban (YRI), and East Asian (CHB/JPT) populations from HapMap [70] and 1000 Genome Project are provided. Both SNV and INDEL are annotated by batch submission to the SeattleSeq server [97], and for SNV additional annotations are acquired using a locally or cloud installed SIFT [99]. The SNV or INDEL within a user defined distance (default: 5 base-pairs) to exon-intron splice boundaries are flagged as potential splice variants and the corresponding transcript IDs are provided.

Gene Annotation: Information is reported on genes hosting SNV and INDEL, including (i) the KEGG pathway(s) [100],[101] to which the gene belongs; and (ii) tissue expression specificity of the gene. For the latter, three Affymetrix© microarray data sets were downloaded from Gene Expression Omnibus (GEO) [102],[103]: (i) GSE7307 consisting of 677 samples from 130 different normal and diseased tissue types; (ii) GSE2109 from the Expression Project for Oncology (expO) [104] which has 1426 cancer or tumor samples of 130 tissue origins; and (iii) GSE3526 consisting of 353 normal human tissue samples from ten post-mortem donors which totals 65 different tissue types. These data sets were normalized individually using RMA [105] and the outlier samples were removed. The distribution of the log2 expression values of all probe sets from all samples within the study was used to define the noise level. Specifically, we observed a bi-modal distribution from each study and defined the noise level as the mode of the first peak in the histogram. The average expression level (log2 scale) and standard deviation of a gene across multiple samples of the same tissue type within the study were calculated. If the average expression level of a gene was below the defined noise level, the gene was defined as "not expressed".

Visualization:

Visualization of the variant positions: The variant report provides a hyper-link to a sample-specific Integrated Genome Viewer (IGV) [106, 107] view of aligned reads at the variant position for each SNV and INDEL. The IGV sessions also include track views of the RefSeq genes, the targeted genomic regions, and pre-compiled UCSC Genome

Browser tables [108]. The UCSC track includes: (i) regulatory regions from the tables of firstEF, vista Enhancers, regPotentail7X, tfbsConsSites, and switchDBTss; (ii) evolutionarily conserved domains.

Visualization of the tissue expression specificity: for each of the microarray studies, the average log2 expression level and standard deviation of a gene across multiple samples of the same tissue type are plotted using R [109]. Each gene from the variant report is hyper-linked to a HTML page that contains three plots for each of the three microarray data sets.

Visualization of the gene-pathway association: for each of the variant-hosting genes, a hyper-link is provided to a Microsoft Excel© file that contains all pathways the host gene belongs to. In addition, each pathway in the Excel file is hyper-linked directly to the specific pathway at the KEGG [100],[101] web site.

Default variant filtering: A list of potential disease associated SNV or INDEL is provided consisting of variants with at least one of the following characteristics: non-synonymous or non-sense variants, "novel" variants not observed in dbSNP and 1000 Genome Project, potential splice-variants near the exon-intron boundaries, variants predicted by SIFT [99]/PolyPhen [110] as damaging, or INDEL causing frame-shifting.

Workflow Integration: The different components of the workflow are integrated and the data files are parsed using custom Perl, Java, Groovy and shell scripts. The sequence alignment and variant calling modules were parallelized using the Sun Grid Engine. Parallelization focused on distributing sample and chromosome levels processing on individual CPUs. Note that a non-parallelized version is also available for end users with no access to a Linux cluster.

2.5 Results

TREAT components/workflow: TREAT consists of three key analytic components (**Figure 2.1**). These three modules, in order of execution, are: (A) the read QC and alignment module, (B) the variant calling module, and (C) the variant and gene annotation module. There are three separate entry points to the workflow. A user can either: (I) upload the FASTQ files to run all three modules, (II) supply the aligned reads (BAM files) to take advantage of the variant calling and the annotation modules, or (III) provide the variant lists (INDEL and SNV variants in separate files) in the defined format and run the annotation module only. The selection of BWA for alignment of short-read data was due to its fast and superior performance on paired-end sequenced reads that we routinely encountered [89]. Assessment was done using both real and simulated data utilizing a variety of read-length to determine the best tool as default in the TREAT workflow (**Table 2.1** and **Table 2.2**). The evaluation of variant callers demonstrated SNVMix [95] as the most efficient (**Figure 2.2**) and was thus selected for the workflow.

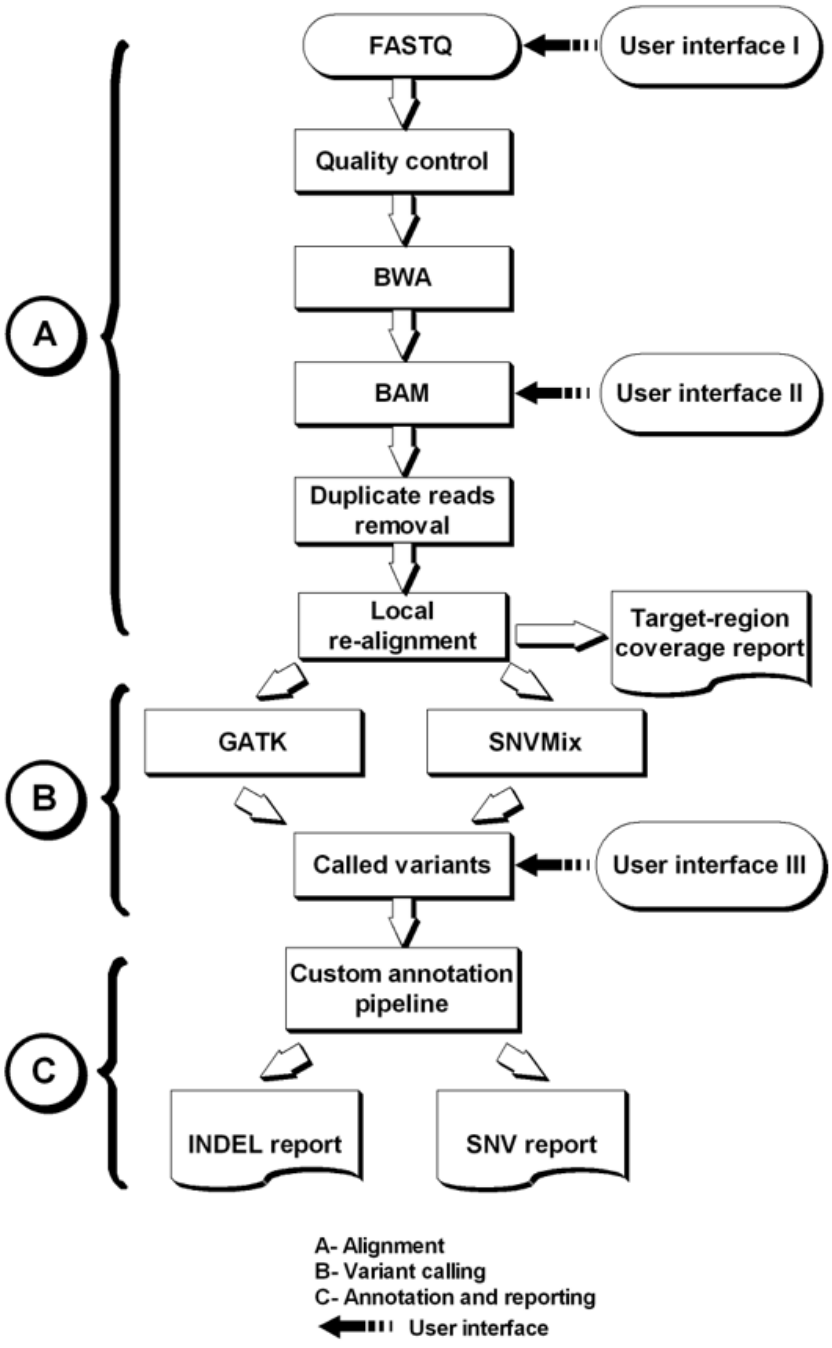

Figure 2.1: Basic flowchart of the TREAT workflow. The three major sections of alignment, variant calling and annotation are highlighted as A, B and C. The three entry points for utilizing the TREAT workflow are also highlighted as user interface options.

Program	Time (h)	Conf (%)	Paired (%)
Bowtie	5.2	84.4	96.3
BWA	4.0	88.9	98.8
MAQ	94.9	86.1	98.7
SOAP2	3.4	88.3	97.5

Table 2.1: Evaluation of short-read aligners on a real dataset using about 12 million sequencing reads mapped to the human genome (Reproduced with permissions from Oxford University Press – Appendix A.6)

	Single-end			Paired-end		
Program	Time (s)	Conf (%)	Err (%)	Time (s)	Conf (%)	Err (%)
Bowtie-32	1271	79.0	0.76	1391	85.7	0.57
BWA-32	823	80.6	0.30	1224	89.6	0.32
MAQ-32	19797	81.0	0.14	21589	87.2	0.07
SOAP2-32	256	78.6	1.16	1909	86.8	0.78
Bowtie-70	1726	86.3	0.20	1580	90.7	0.43
BWA-70	1599	90.7	0.12	1619	96.2	0.11
MAQ-70	17928	91.0	0.13	19046	94.6	0.05
SOAP2-70	317	90.3	0.39	708	94.5	0.34
bowtie-125	1966	88.0	0.07	1701	91.0	0.37
BWA-125	3021	93.0	0.05	3059	97.6	0.04
MAQ-125	17506	92.7	0.08	19388	96.3	0.02
SOAP2-125	555	91.5	0.17	1187	90.8	0.14

Table 2.2: Evaluation of short-read aligners on a simulated datasets using 1 million sequencing reads of lengths 32bp, 70bp and 125bp mapped to the human genome (Reproduced with permissions from Oxford University Press – Appendix A.6)

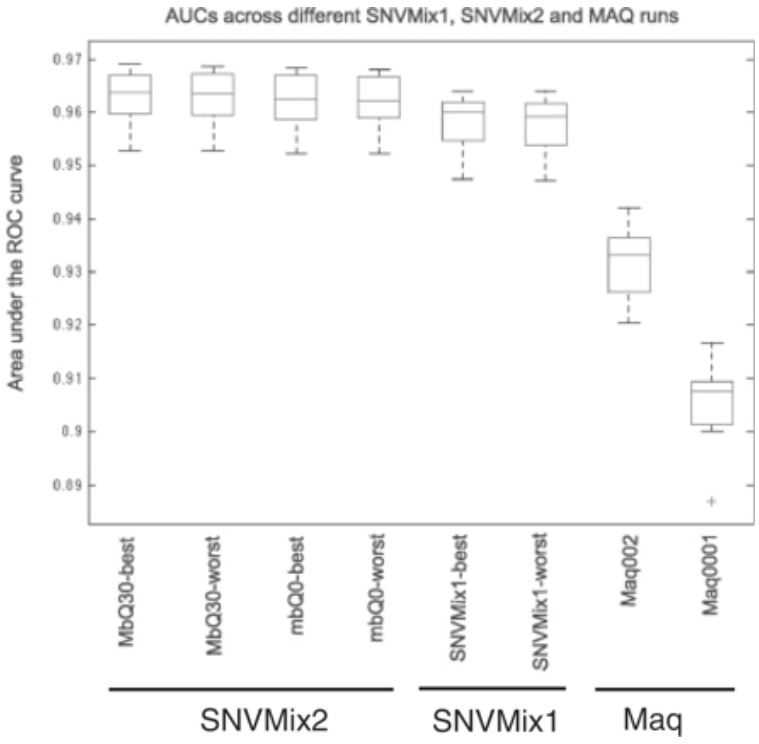

Figure 2.2: Evaluation of SNVMix using cross-validation runs highlights the better performance even with worst-case runs [95] (Reproduced with permissions from Oxford University Press – Appendix A.7)

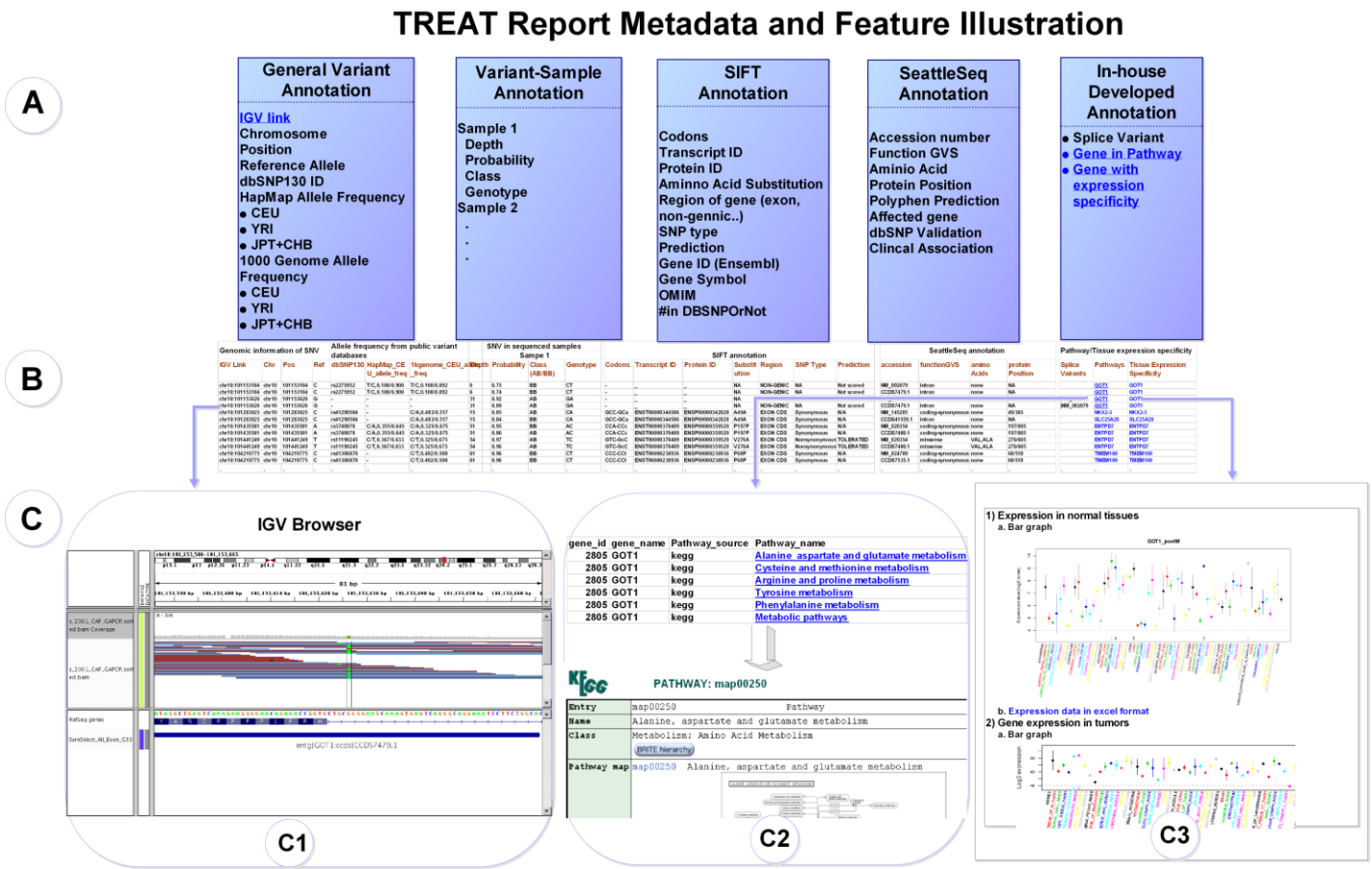

Figure 2.3: Annotation and Visualization Features from the TREAT Variant Report**.** (A). Four annotation categories provided by TREAT: the general variant annotations, the sample-related variant annotations, SIFT and SeattleSeq annotations, and in-house developed annotations. The specific annotation items are listed under each category. (B). A snapshot view of the Excel variant report. The variants are in rows, and annotations are in columns. (C). Three visualization options provided by TREAT: the IGV browser view of the variant positions, the hyper-linked KEGG pathway for each variant-hosting genes, and the tissue expression specificity figures based on three microarray studies.

HTML-based results report: the results from TREAT, including the project description, QC reports, coverage and sequencing depth information, summary of the

SIFT [99] and SeattleSeq [97] annotations, as well as links to the SNV and INDEL reports, are presented in one easy-to-navigate HTML page. Different sections of the HTML page are explained in detail in the TREAT user manual.

Variant Reports: TREAT outputs four variant report files, including the complete INDEL and SNV reports as well as the filtered INDEL and SNV reports using the default filtering criteria described above. These variant reports are provided in Microsoft Excel© format (note that for bigger report files, the users will need Excel 2007). As shown in Figure 2.2.A, there are 5 categories of annotations for each reported variant. The general annotations include the variant physical positions on the chromosome, the reference base at the position, the dbSNP ID when available, and the allele frequencies in CEU, YRI, and JPT/CHB of the SNV reported in both HapMap and the 1000 Genome Project. The sample-related annotations give the read depths at the variant positions, the SNVMix [111] probability score for SNV, and the called genotype of the sample. The SIFT and SeattleSeq annotations including transcript and gene IDs, variant type (synonymous or non-synonymous), and the SIFT and PolyPhen predictions. Our in-house developed annotations provide the distances of each variant to the closest intron-exon boundaries, the tissue expression specificities of the host genes based on multiple microarray experiments, and KEGG pathways each host gene belongs to.

The TREAT workflow also implemented flexible and easy-to-use visualization options (Figure 2.2.C). It automatically outputs a pre-formatted IGV session file for each

sample, and each variant position is visualized with a click-able hyper-link from the Excel file. This enables close examination of the multiple sequence alignment of short reads and the read/base qualities at each reported variant positions (Figure 2.2.C1). Hyper-links are also used to display additional annotations on each variant hosting gene, including the list of KEGG pathways a host gene belongs to and the tissue expression specificity of the gene. The list of pathways is itself hyper-linked to the KEGG website (Figure 2.2.C2) for detailed visualization. Tissue expression specificity is displayed in a separate HTML page which provides additional links to the detailed descriptions of the 3 microarray experiments (GEO link), links to the Excel tables of the actual expression values, and links to the visualization of the expression specificities of each gene across different tissue types (Figure 2.2.C3). For users who choose to locally install TREAT, the ~21,000 Excel files (one for each gene) for the KEGG pathway associations, and ~63,000 PDF files (three microarray studies per gene) for the tissue specificities can be downloaded from the TREAT website. For those who use TREAT on Amazon Cloud, these annotation files are pre-loaded.

Redundant Sources for Variant Annotation: When evaluating SeattleSeq and SIFT, we observed that for regions with two overlapping genes, one of the tools only reports the annotations on one randomly selected gene. Although rarely, we also observed that one of the tools changes the input alleles of some variants for no obvious reason. Other differences between SIFT and SeattleSeq where observed when we submitted 22,135 SNV from an Exome-Seq sample, together with two known causal mutations

from Freeman-Sheldon Syndrome (FSS) [112] and Van Den Ende-Gupta Syndrome (VDEGS) [113], to both SeattleSeq and SIFT. As shown in Figure 2.3, SIFT predicted the FSS mutation as "damaging" and the VDEGS mutations as "damaging with low confidence", while SeattleSeq which uses PolyPhen as the prediction method categorizes the FSS mutation as the "probably damaging" but didn't give a prediction for the VDEGS mutation. Both mutations are correctly classified as "novel" non-synonymous variants by SIFT and SeattleSeq, although the number of coding and non-coding variants defined by the two tools are slightly different. However, since none of these two tools could be fairly assessed as being more accurate than the other, we decided to provide the annotations from both.

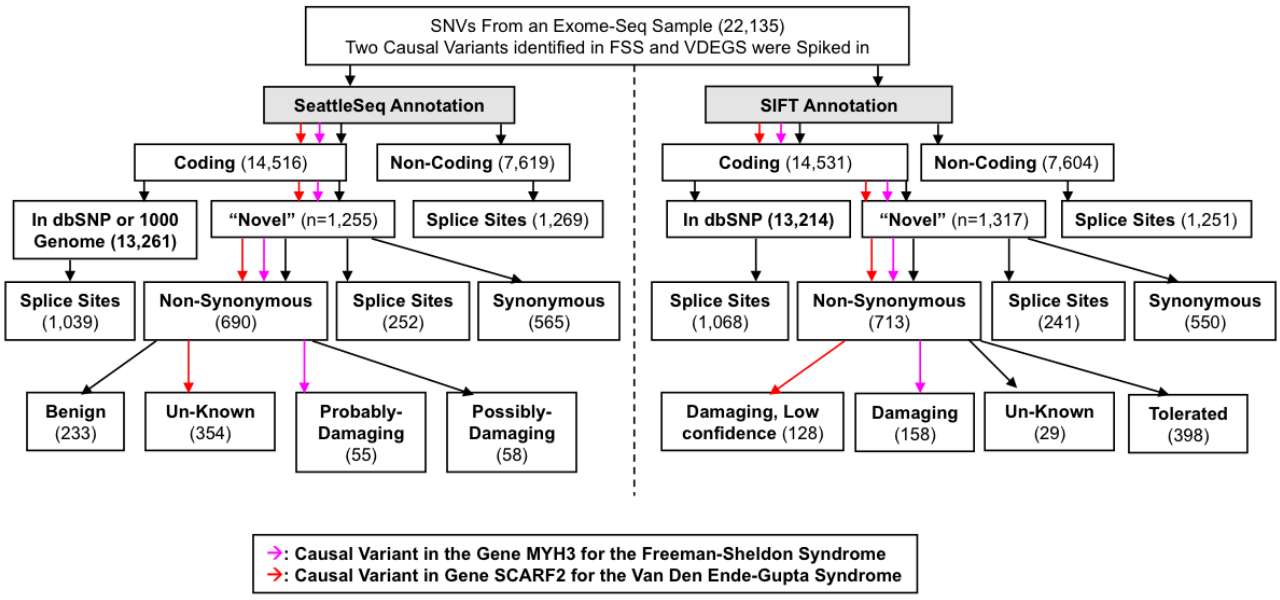

Figure 2.4: Evaluation of the TREAT workflow using spiked-in known variants. The variants for Freeman-Sheldon Syndrome (FSS) and Van Den Ende-Gupta Syndrome (VDEGS) were spiked into an exome dataset and run through the TREAT annotation to evaluate how the diverse annotation tools recover the variants.

Amazon Machine Image: We have built a non-parallelized Machine Image that can be launched on the Amazon Elastic Compute Cloud (EC2). The Machine Image is loaded with the all the open-source tools and necessary annotation files for the direct execution of TREAT. **Table 2.1** summarizes a run performed on 75 million, 100bp single-end (SE) WES reads from an Illumina Sequencer. It took 106 hours for TREAT to upload the FASTQ files, align the reads, call the genotypes, and annotate the called variants. The process cost ~$75. For comparison, a run performed on 500 million 100-base SE reads using the locally installed TREAT that have been parallelized on SGE took 10 hours. On the Amazon Cloud, the most time-consuming step was the local re-alignment of the BAM files before the genotype calling (100 hours). It should be noted that the annotation module was executed in less than one hour with a cost of less than a dollar. Since most genome centers provide called variants as part of the output, the Amazon Cloud implementation of TREAT provides a very cost effective means to annotate these called variants.

	Run Time (hrs)	Cost ($)
FASTQ Upload	3.00	$ 1.90
BWA Alignment	4.00	$ 2.72
Local Re-alignment	100.00	$ 68.00
INDEL Calling	1.50	$ 1.02
SNV Calling	0.33	$ 0.68
Annotation	0.67	$ 0.68
Total	**106.50**	**$ 75.00**

Table 2.3: Cost Estimate of Running TREAT on Amazon Cloud. The FASTQ files of an Exome sequencing run of 75 million 100-base pair-end reads were submitted (single node), actual run times and the associated costs were recorded.

Availability: The source code and executables (for both the parallelized and non-parallelized versions of the tool), a detailed user manual, and other supplementary files can be found at: http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/bioinformatics-software-packages. In addition, an Amazon Machine Image is also available at the website for users who don't have access to a LINUX cluster.

2.6 Discussions

We have developed an analytical workflow TREAT [1], which addresses the current challenges in analyzing and interpreting targeted re-sequencing data. The modular design of the workflow allows the users to interface with TREAT and initiate the analyses at different levels: (i) to start with FASTQ files and utilizes all three modules; (ii) to start using the aligned reads in BAM format and use TREAT for variant calling and annotation; (iii) to use the variant annotation only. Compared to other variant annotating tools such as ANNOVAR [85], GAMES [86], and SeqAnt [87] that tend to focus on certain aspects of sequence analysis, TREAT provides the largest set of annotations, including the tissue specificity and KEGG [100],[101] pathway associations to each variant-hosting genes, both SIFT [99] and SeattleSeq [97] variant annotations, and allele frequencies of known variants in multiple populations using both HapMap [70] and 1000 genome project [24] data. At the time of publication, TREAT was the only workflow with modular flexibility, excel formatted end-reports that investigators can directly work

with, integrated visualization option along with variant and annotation information and an amazon cloud instance to ensure capability to kick-start the analysis without comprehensive local bioinformatics or IT infrastructure.

The superior accuracy and efficiency of TREAT workflow is due to the component applications utilized in the entire process. These applications included in TREAT have been carefully evaluated and selected from a pool of available open source applications. For read alignment, Bowtie [114], BWA, MAQ and SOAP [115] were the top aligners we identified from a usability and support stand-point. BWA was more accurate in both the fraction of confidently mapped reads and the error rate of confident mappings [89]. So we decided to use BWA for our workflow. For variant calling, SNVMix, MAQ and GATK were selected. We found SNVMix to be the most accurate for SNV calling and GATK for INDEL and thus selected those for our default workflow. In some circumstances the output of more than one application providing similar annotations has been included in TREAT since no fair criteria could be used to rank the quality/accuracy of one application above the other. This is the case of SeattleSeq and SIFT for variant annotations with observed discrepancies. Both annotations are includes in the final report to let the user choose the more appropriate one or perform consensus selection.

The various reports and graphic interfaces provided by TREAT make it a flexible and easy to use application. The reporting of results in Excel format integrates the

visualizations of the sequence alignment at variant positions, pathways and expression specificity of the variant hosting genes via click-able hyper-links for each reported INDEL and SNV. In addition, the summary of the targeted re-sequencing results is stored in a centralized HTML report with links to the TREAT website, the targeted region coverage report, the read QC report, the description of the TREAT workflow, and links to the website of the annotation tools and databases.

For maximum flexibility, three versions of TREAT were implemented: (i) a parallelized version for users with access to Sun Grid Engine LINUX clusters; (ii) a non-parallelized version for users with a single LINUX node; (ii) and an Amazon Cloud version for users with no access to local bioinformatics infrastructures. By targeting all user groups and enabling rapid integration of emerging analytic methods, we believe that TREAT provides a sustainable NGS analytic workflow with wide applicability to the research community.

We continue to add new functionality and features to TREAT to make it a comprehensive tool for targeted re-sequencing analysis, annotation and interpretation. These include the accommodation of allele frequencies from the large list of populations from dbSNP database, and adding more splice-site prediction data from various tools for variants near exon-intron boundaries. In addition, we are constantly comparing the new alignment and variant-calling tools with the ones used in TREAT and will switch to better tools in our workflow when validated. Most importantly, we are in the process of

developing an in-house variant database that collects all variants detected from hundreds of individuals with various types of diseases using exome and whole-genome sequencing. This database will provide critical annotations whether the observed variants are truly "novel" or disease-specific.

In summary, TREAT integrates sequence alignment, variant calling, variant annotation, variant filtering, and visualization in one comprehensive analytic workflow. The rich set of annotations provided by TREAT, the easy to use, centralized HTML summary report, and the Excel-formatted variant reports with hyper-linked visualization utilities enable the filtering of detected variants based on their functional characteristics, and allow the researchers to navigate, filter, and elucidate tens of thousands of variants to focus on potential disease-associated variant(s). It is a modular, efficient and streamlined workflow that scales well, is used by multiple institutes including Mayo Clinic, UIUC, Appistry© and is capable of evolving with the changing needs of NGS Bioinformatics.

Chapter 3: Depth of Coverage Evaluation - A Case Study of Inherited Neuropathy[2]

3.1 Coverage from Sequencing Data

Depth of coverage, or simply coverage, is one of the most vital metrics in DNA sequencing data evaluation. It is defined as the number of times a particular nucleotide is sequenced, that is, the number of reads mapping to that nucleotide location. Coverage is generally reported as an average coverage of a region based on the total number of reads and region length, or as a percentage of the region covered by at-least a threshold number of reads. Much higher average sequencing depth is necessary to achieve accurate variant calling over the entire genome/exome [60]. It has been reported that some regulatory regions consistently have lower coverage from NGS data leading to poorer SNP detection [62]. However, higher coverage also entails greater cost of data generation implying making a trade-off decision.

To illustrate this, Figure 3.1 demonstrates a set of sequencing data with varying coverage. In most cases, coverage profiles are reproducible among samples sequenced using comparable technology and throughput, with the coverage variability being

[2] Text and figures reproduced with permission from BMJ Publishing Group Ltd (Appendix Figure A.2)

sequence (homology to genomic regions) and GC (percentage of Guanine and Cytosine nucleotides) content dependent.

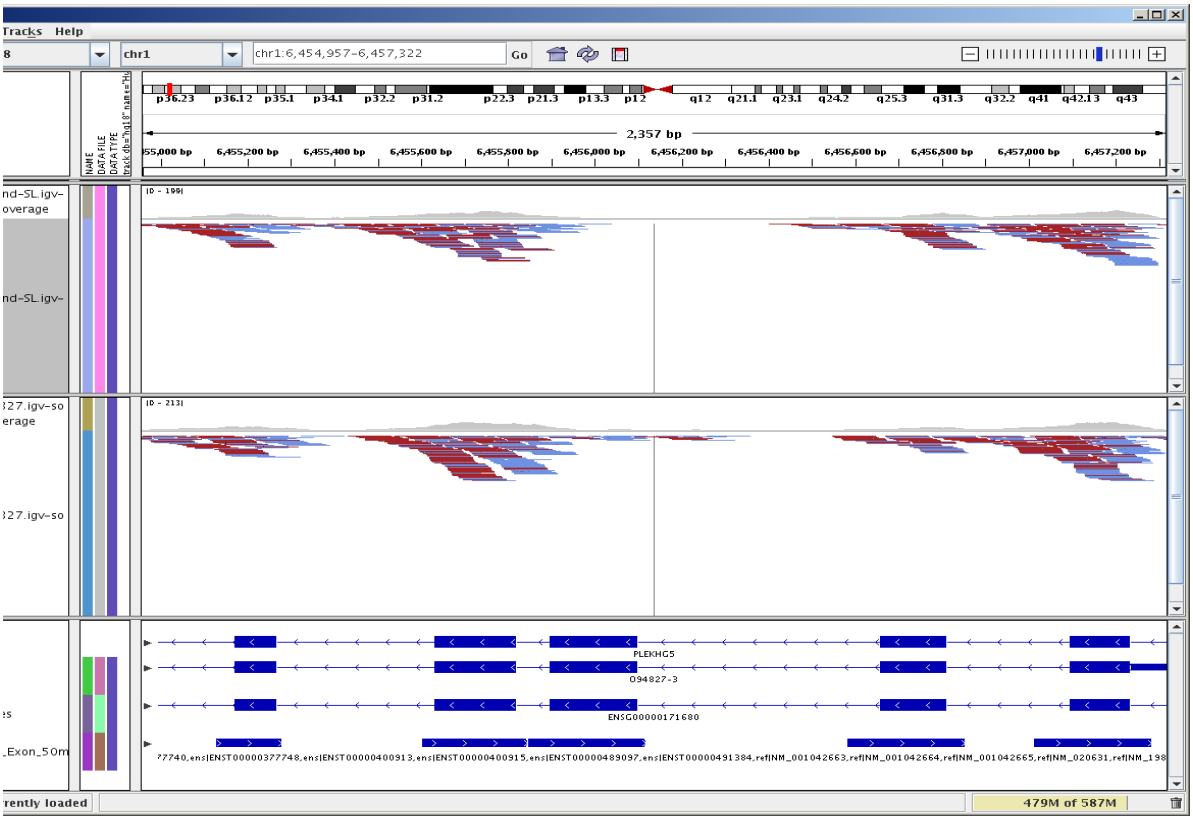

Figure 3.1: Variability in coverage of exons from WES data visualized using IGV. The red and blue segments are short Illumina reads, while the solid blue bars at the bottom are coding exons of the gene. Two samples of the same batch with similar coverage profiles are depicted.

3.2 Inherited Peripheral Neuropathy

Peripheral neuropathy is a common medical problem suffered by adults with a prevalence reported to be 15% in persons older than 40 years [116]. A wide range of

neuropathy screening tests such as electrophysiological examinations (nerve conductions and EMG), blood tests, and MRI imaging studies are performed in routine clinical evaluation [117]. Nerve and skin biopsies may also be performed in select patients [118, 119]. Most recently reviewed mean Medicare expenditures per neuropathy patient in the year of diagnosis were $14,362, and despite the expense, many patients (48% of 12, 673) went without specific diagnosis [117]. The ordered tests objectify the neuropathy and focus on determining inflammatory, autoimmune; metabolic, toxic causes and exclusion of alternative symptomatic causes. Highlighted to explain the high rate of undiagnosed patients is the lack of standard evaluations including for early hyperglycemia [120]. However, also contributing in the low rate of diagnosis is the lack of any standard test to simultaneously address a wide array of inherited neuropathy causes.

Inherited neuropathies are common and their genetic and clinical heterogeneity make diagnosis difficult. The phenotypic similarity to acquired neuropathies further adds to the complexity in diagnosis. For example, a large prospective clinical trial of 205 undiagnosed neuropathy patients found that only by intensive clinical evaluation including by kindred studies were diagnosis possible and inherited neuropathies accounted for 42% of these patients [121]. Inherited causes of neuropathy are diverse and currently Online Mendelian Inheritance in Man (OMIM) [122] lists 593 entries linking chromosomal locus or mutated genes with neuropathy. If identified as inherited comprehensive genetic testing can lead to specific genetic diagnosis. In one common form of inherited neuropathy, Charcot-Marie-Tooth (CMT), also known as hereditary

motor and sensory neuropathy (HMSN), a recent large study demonstrated that extensive genetic testing was able to find genetic causes for 67% of HMSN patients [123].

In an attempt to reduce cost in an ever-growing list of causative neuropathy genes helpful algorithmic candidate gene selection approaches have evolved [123-126]. Inherent in the effectiveness of candidate gene approaches, however, are: 1) expertise in bedside neuropathy clinical examinations, 2) medical genetics understanding for patterns of inheritance, and knowledge of family history; 3) accurate nerve conductions to segregate axonal from demyelinating forms; and 4) the assumption there is a limited phenotypic range with the ordered candidate genes [127]. Unfortunately current review of standard U.S. neuropathy evaluations would suggest these candidate gene approaches to be problematic for most patients as only 19.8% have been evaluated by nerve conductions, [117] and the ability of physicians to properly classify neuropathy types has been seriously questioned [120]. Additional complication is that gene mutations may have a wider phenotypic range than initially appreciated thereby preventing proper candidate gene selection. Compounding the problem is that even if a genetic cause is suspected the cost of even a limited number of candidate genes is typically prohibitive, with an average cost per gene ranging from $500-1000.

3.3 NGS in Inherited Neuropathy

With recent advent of NGS technology, WES has emerged as an exciting new tool with potential to diagnose heterogeneous genetic neurological disorders [128]. Single-family examples of next generation application in neuropathy gene identification are now reported including by targeted candidate exome regional analysis [129-132]. WES screens nearly all exons in the genome with unprecedented speed, and the cost is continuously decreasing. Despite the excitement of applying WES in routine neurological practice, a comprehensive study to establish its effectiveness in the many categories of inherited neuropathy has not been reported. WES has capability of screening a large number of genetic heterogeneities simultaneously when clinical similarities exist among different disorders. In this report, we investigate the utility of WES as a screening test in genetic diagnosis of five common inherited neuropathy categories: 1) hereditary motor and sensory neuropathy (HMSN); 2) distal hereditary motor neuropathy (dHMN); 3) hereditary sensory and autonomic neuropathy (HSAN); 3) complicated-hereditary spastic paraplegia (c-HSP); 4) various metabolic neuropathies.

The number of causal genes for inherited neuropathies is predicted to accelerate, largely due to the application of WES [133]. WES can readily cover any newly discovered causal genes and expand our understanding of phenotype–genotype associations. For example, mutations in ATL1 have been largely linked with pure HSP and only rarely with HMSN or HSAN neuropathies [134, 135].

3.4 Our WES Data to Study Inherited Neuropathy

Twenty-four indexed patient samples from 15 kindred underwent WES and bioinformatics analysis for coverage and sequencing depth of genes responsible for inherited neuropathies. These kindred had earlier unsuccessful candidate gene testing and the five probands were initially thought to have acquired neuropathy. Five major classifications of inherited neuropathies (HMSN, dHMN, HSAN, complicated-HSP, and genetic metabolic neuropathies) were chosen for evaluation. A total of 74 known genes were reviewed including specific interrogation of their 5195 previously reported pathogenic mutations. The most commonly reported mutated genes were also specifically reviewed. These 74 neuropathy genes (Table 3.1) were generated based on a recent review [136], GeneReviews [137], Online Mendelian Inheritance in Man (OMIM) [122] and the Human Genome Mutation Database (HGMD) [138]. Simply extracting the average coverage sequencing depth may miss nucleotide-level details, so we further evaluated each nucleotide in the coding exons of these genes. Nucleotide bases with at least 1× (one-fold) coverage were considered sequenced, and those with coverage of ≥ 10× (≥ 10-fold) were considered genotyped [139].

Neuropathy	Gene*
Hereditary motor and sensory neuropathy type 1 (AD)	*PMP22, MPZ, LITAF, EGR2, NEFL, GDAP1, INF2*
Hereditary motor and sensory neuropathy intermediate (AD)	*MPZ, DNM2, YARS, NEFL, GJB1*
Hereditary motor and sensory neuropathy type 2 (AD)	*MFN2, MPZ, GARS, BSCL2, GDAP1, HSPB1, HSPB8, TRPV4, RAB7 KIF1B, AARS, DYNC1H, LRSAM1, TFG*
Hereditary motor and sensory neuropathy type 2 (AR)	*LMNA, MED25, GDAP1, PRPS1, HINT2*
Hereditary motor and sensory neuropathy type 4 (AR)	*GDAP1, MTMR2, SBF2, FGD4, SH3TC2, NDRG1, EGR2, CTDP1, PRX*
Distal hereditary motor neuropathy	*HSPB8, HSPB1, GARS, DCTN1, BSCL2, TRPV4, DYNC1H1, SETX, IGHMBP2, ATP7A*
Complicated hereditary spastic paraplegia (AD)	*ATL1, SPAST, NIPA1, KIA0196, BSCL2, REEP1, ZFYVE27*
Complicated hereditary spastic paraplegia (X linked, AR)	*PLP1, CYP7B1, SPG7, SPG11, ZFYVE26, SPG20, SPG21, L1CAM, SLC16A2*
Hereditary sensory and autonomic neuropathy	*SPTLC1, RAB7, SPTLC2, ATL1, DNMT1, WNK1, FAM134B, KIF1A, IKBKAP, NTRK1, NGF, CCT5, DST*
Metabolic neuropathy	*TTR, ARSA, GALC, ABCD1, PHYH, GLA, CYP27A1, PBGD (HMBS), XPC, ATM, TTMP*

*Genes studied and their associated mutations were derived from a recent review,[8] Gene Reviews, Online Mendelian Inheritance in Man and the Human Genome Mutation Database.
AD, autosomal dominant; AR, autosomal recessive.

Table 3.1: Genes studied by Neuropathy type. These neuropathy genes were generated based on a recent review [136], GeneReviews [137], OMIM and HGMD. (Reproduced with permission from BMJ Publishing Group Ltd Appendix Figure A.2)

For the selected 24 WES cases, we generated 12–14 billion nucleotide (Gb) of sequencing data per sample to achieve an average sequencing depth of 140× with high quality genotype calls for targeted region.

3.5 WES Data Analysis

In-solution exome capture was performed using SureSelect Human All Exon Kits V4 (51Mb) according to the manufacturer's protocol (Agilent Technologies, Santa Clara, CA). Illumina Hiseq2000 (Illumina, San Diego, CA) 101-bp paired-end read sequencing was performed and data was analyzed with the in-house workflow. Briefly, the sequence reads were aligned to the human genome build 37.1 using Novoalign (v2.07.13) [140] followed by re-alignment, re-calibration and variant calling using GATK (v1.2-26) [93, 94]. The called variants were filtered using variant quality score recalibration (VQSR) and annotated using the latest TREAT workflow [1]. The dbSNP135 [98], 1KGenome [24] and EPS6400 Exome database [28] annotation were used as additional filtering tool on the lists of variants.

3.6 Coverage Report from TREAT

An output from the TREAT workflow for WES is a coverage plot of the targeted region, reporting the percentage of region covered at a particular depth of coverage. WES

like all other NGS applications has non-uniform sequence depth, requiring larger amounts of sequencing to better cover majority of the targeted region. I developed a comprehensive read coverage analysis package that uses the alignment data to evaluate a list of prioritized genes from DNA sequencing data. The analysis reports gene, exon and nucleotide level coverage of the genomic regions along with assessment of the known mutations for the disease being studied. Moreover, the coverage reports are provided as the alignment completes, before the steps of variant calling and annotation. It is useful to get that early access information in order to gauge the sequencing performance. This work was done by way of focused bioinformatics analysis to examine the efficacy of WES as a screening tool for the five major types of Inherited Peripheral Neuropathy.

3.7 Coverage among Inherited Neuropathy Genes

To further investigate the utility of exome sequencing as a screening tool for diverse neuropathies, we comprehensively analyzed the WES on 24 unrelated Caucasians for the percent and depth of coverage of causal genes and known mutations in 5 common classifications of inherited neuropathies (Table 3.1, Figure 3.2). Our analyses include (1) HMSN: 33 genes, 1342 mutations; (2) dHMN: 10 genes, 443 mutations; (3) c-HSP: 17 genes, 1037 mutations; (4) HSAN: 13 genes, 142 mutations; and (5) select genetic metabolic neuropathies: 11 genes, 2331 mutations (Table 3.2). Among the metabolic neuropathies, we focused on those with treatment prospects [136].

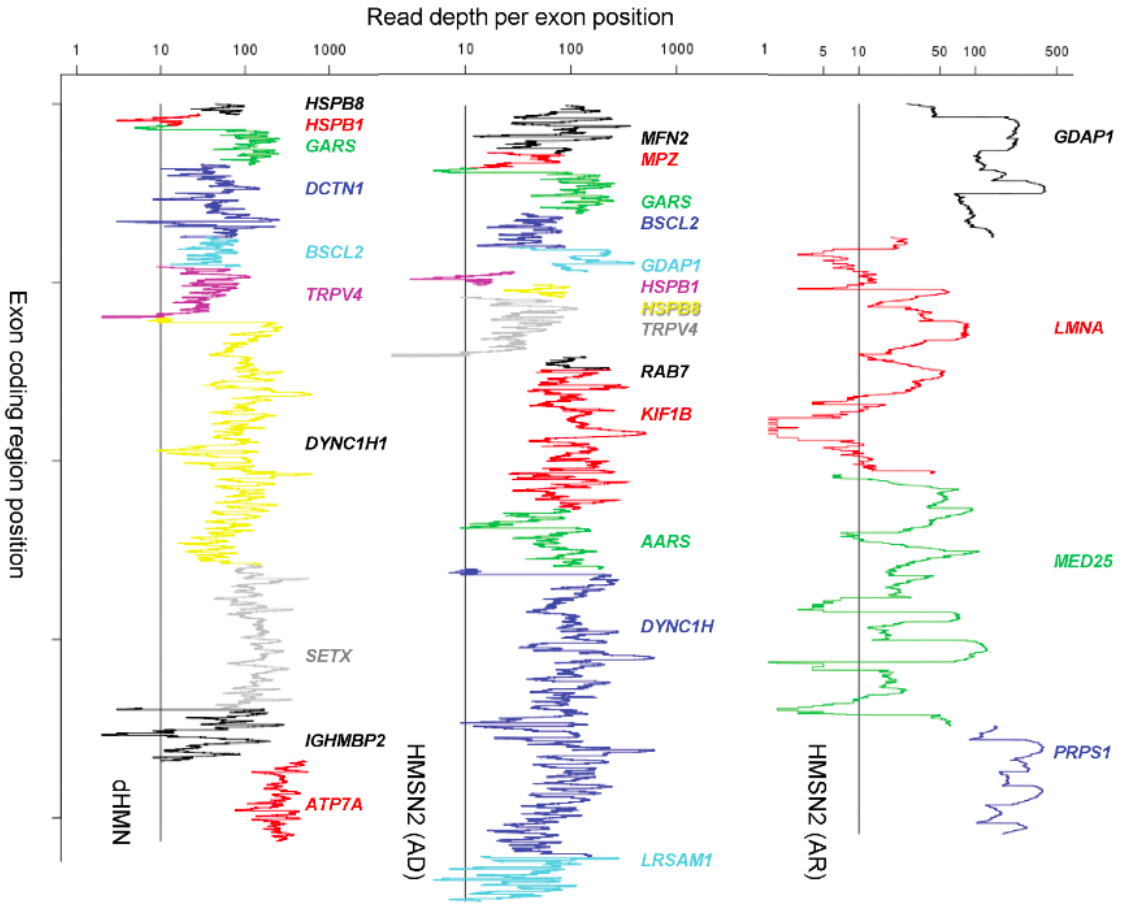

Figure 3.2: Coverage of coding regions for genes implicated in axonal motor neuropathies. It displays the sequencing depth of coverage across coding region of axonal neuropathy genes for a WES sample. The solid black line demarcates 10x (ten-fold) coverage that is sufficient for efficient genotyping, and the read-depth is plotted on a log-scale. The results shown are from Agilent Sure Select All Exon Kit on HiSeq 2000 with 100 base paired end sequencing. AD denotes autosomal dominant; AR, autosomal recessive; dHMN distal hereditary motor neuropathy; HMSN2, hereditary motor and sensory neuropathy type 2; WES, whole exome sequencing. (Reproduced with permission from BMJ Publishing Group Ltd Appendix Figure A.2)

We conducted three steps of coverage analysis. First, we checked the coverage of Agilent All-Exon capture kit for the Consensus Coding Sequence (CCDS) [141] exons of these 74 unique neuropathy genes. The Agilent All-Exon kit is designed to capture almost all exon regions of entire genome, but not all the bases in the exons are captured

due to the difficulties in designing probes for certain regions such as repetitive elements and GC rich regions. We found that >98% of CCDS exons of these 74 neuropathy genes are targeted by Agilent All-Exon capture kit. This means that theoretically WES should generate sequencing coverage data for more than 98% of exons in these 74 genes.

Clinical neuropathy	No. of genes	No. of mutations evaluated	No. of CCDS exons covered*	Per cent of CCDS exons with >10× coverage, mean (range)	Per cent of mutations with 10× coverage, mean (range)	Per cent of mutations with 5× coverage, mean (range)	Per cent of mutations with 1× coverage, mean (range)
HMSN	33	1342	497/502 (99%)	93 (89–95)	89 (80–96)	96 (89–98)	99 (98–99.8)
dHMN	10	443	224/226 (99%)	95 (95–96)	94 (93–95)	96 (95–96)	98 (96–99.7)
c-HSP	16	1037	268/271 (99%)	93 (90–95)	97 (94–99)	99 (98–99.6)	99.7 (99.6–99.8)
HSAN	13	142	338/344 (98%)	96 (94–97)	90 (84–94)	95 (91–98)	99 (98–100)
Metabolic neuropathies, limited set	11	2231	169/173 (98%)	89 (85–90)	91 (85–95)	96 (93–97)	99 (99–99.8)

*Covered by the SureSelect V3 (50 Mb) capture kit.
CCDS, consensus coding sequence; c-HSP, complicated hereditary spastic paraplegia; dHMN, distal hereditary motor neuropathy; HMSN, hereditary motor and sensory neuropathy; HSAN, hereditary sensory and autonomic neuropathy.

Table 3.2: Coverage of neuropathy genes by clinical subtype using WES. The genes and mutations implicated in various clinical neuropathies were gleaned from literature reviews. The 5th column for percent of CCDS exons with more than 10x coverage is assuming at-least 90% of the exon coding region at 10x or more coverage. Reproduced with permission from BMJ Publishing Group Ltd Appendix Figure A.2)

Second, we performed analysis on actual coverage of exons in 74 neuropathy genes based on the sequencing data of 24 exomes. We found that 89-97% of nucleotide bases of the exons are sequenced at >10X depth (Table 3.2). Based on our WES analysis experience, we consider 10X depth an adequate threshold of confident variant calling for

germ line mutation, especially when we plan to use Sanger sequencing to confirm the identified mutations tracking with the affected status in the kindred.

Third, we investigated whether all known 5195 unique mutations in the non-redundant list of 74 neuropathy genes were covered. The results showed that on average, 89% of 1342 mutations for HMSN, 94% of 443 mutations for dHMN, 97% of 1037 mutations for complicates HSP, 90% of 142 mutations for HSAN, 91% of 2231 mutations for metabolic neuropathy were sequenced at >10X depth.

3.8 Discussions

The comprehensive bioinformatics analysis revealed the diverse forms of inherited neuropathy: HMSN, dHMN, HSAN, c-HSP and select metabolic neuropathies with >98% coverage and >10× sequencing depth for 93% (range 89%–96%) of all exons and 5195 known mutations in 74 neuropathy genes.

This analysis is a perfect use-case for demonstrating the importance of comprehensive coverage analysis. It highlights the importance of evaluating regions of low coverage from WES data that may result in suboptimal interrogation of the genes of interest. WES combined with Sanger sequencing fill-in of inadequately covered exons

may be necessary to improve a genetic test's sensitivity to detect presence of disease associated mutations.

Another observation is the considerable variability in coverage of genes among different samples, even though some regions may have consistently low coverage. This makes it important to run the coverage evaluation on each sample especially for the genes or regions of interest. The TREAT workflow [1] generates a generic coverage plot of targeted region, reporting percentage of region covered at a particular depth of coverage. The incorporation of focused coverage analysis reports on gene, exon and nucleotide level coverage of genomic regions along with an assessment of known mutations for the disease being studied.

Chapter 4: Fast Reporting of Clinically Relevant WGS Variants[3]

4.1 Introduction

Whole Genome Sequencing is beginning to reshape the understanding of how DNA variants work and influence disease risk from the perspective of personalized treatment. WGS has the potential to transform diagnostic testing in the very near future. As the cost of sequencing continues to decrease, the broader adoption of this protocol by clinical laboratories is expected. Sequencing platforms are being redesigned to accelerate the sequencing of whole genomes. For instance, the Illumina Hi-Seq 2500 platform can perform this task in almost a day, shifting the rate-limiting step to data processing. The computationally expensive step of aligning millions of short reads to the whole genome could be prohibitive for routine use of WGS in a clinical setting where the speed of analysis can impact patient outcome. Clinical applicability can be improved by expediting WGS variant reporting based on relevance for clinical decision-making. Currently, most of the clinically relevant genomics information is related to protein-coding exome regions [38] where the impact of coding variants can be interpreted in the context of proteins and their function [71, 142]. This current focus opens opportunities to

[3] Text and figures reproduced under the terms of Creative Commons Attribution License (Appendix Figure A.3)

develop new bioinformatics algorithms that prioritize and swiftly report clinically relevant findings.

Recently, an ultra-fast preprocessing workflow was published: ISAAC [67]. This workflow completes the whole genome alignment and variant calling in 7–8 hours. Although, ISAAC is the fastest solution currently to our knowledge, its deployment requires specific hardware and is, at least for now, limited to Illumina © sequencing data.

We explored another approach that does not require specific hardware or software solution and is independent of the next generation sequencing platform used. Instead of expediting the whole alignment and calling process, our proposed approach prioritizes read alignment and variant calling in genomic regions of clinical relevance (referred to as the Target Reference Genome) before reporting variants in genomic regions of lower clinical significance. The proposed workflow operates in three steps. First, clinically relevant reads are selected by aligning all the sequencing data to the Target Reference Genome. Then, this reduced set of aligned reads is aligned to the whole reference genome to correct for alignment artifacts. These artifacts arise from reads forcibly aligned to the Target Reference Genome that align more accurately to non-targeted regions. After the second alignment step, reads that remain aligned on the Target Reference Genome are re-aligned and recalibrated followed by variant calling. Variants are immediately reported to clinical experts for interpretation and decision support. The final step, which can be

deferred or executed at a slower pace, handles the remaining reads that are aligned on the whole reference genome.

The gain of reporting speed obtained with this iterative workflow is due to the significantly smaller size of the Target Reference Genome compared to the whole reference genome. If the targeted region corresponds to the whole exome, read alignment in the first step would be limited to less than 2% of the reference genome. Similarly, assuming even coverage, only 2% of the reads will be aligned on the whole reference genome in the second step.

Although conceptually very simple and straightforward to implement, the question of results accuracy remains to be addressed. In this manuscript, we compare results obtained by the target workflow with a generic whole genome sequencing workflow. In the process, we have also compared the impact on our iterative workflow of two aligners, BWA [89, 143] and Novoalign [140], on results accuracy.

4.2 Methods

4.2.1 Datasets

To test out approach, we selected a CEPH family trio from the 1000 Genomes project [24] consisting of NA12878 (Child), NA12891 (Father), and NA12892 (Mother). Each sample was sequenced using the Illumina Next Generation Sequencing Platform (HiSeq 2000) with the pair-end protocol that produced on average 397 bp long sequence fragments from which 100 bp were sequenced at both ends. Sequencing of these samples resulted in more than 2.4 billion 100 bp long reads with an average coverage of 80x across the entire genome. The Binary Alignment Map (BAM) files obtained for these samples were converted to FASTQ reads format for further analysis. The same individuals have been genotyped with a combination of Illumina and Affymetrix SNP chips for HapMap Phase III [70]. This genotype data was used to validate variants calls from sequencing data.

Data Availability

1. Sequencing data: ftp://ftp-trace.ncbi.nih.gov/1000genomes /ftp/technical/working/20120117_ceu_trio_b37_decoy/
2. Genotyping data: ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phaseIIIII/forward/

4.2.2 Target Reference Genome

We selected the set of 2638 clinically relevant genes from the Clinical Genomic Database [144]. It should be noted a smaller set similar to clinical gene panels could have

been selected as well. The 2638 genes include 42048 unique exons in the UCSC RefFlat annotation. The average length of these exons is 280 bp with a standard deviation of 635 bp. The boundary of each exon was extended by 550 bp to account for the sequencing protocol that produced 100bp long paired-end read from about 400 bp long sequence fragments. The extended sequence of each exon was extracted from the Human Reference Genome (Build 37) and concatenated into a single Target Reference Genome FASTA file.

4.2.3 Standard sequence alignment and variant calling workflow

As the standard whole genome alignment workflow, we used Novoalign [140] for initial alignment of sequence reads followed by GATK [93, 94] for re-alignment, re-calibration and variant calling (**Figure 4.1**). This has been the de-facto analysis standard for high quality NGS data alignment and variant calling on DNA sequenced data.

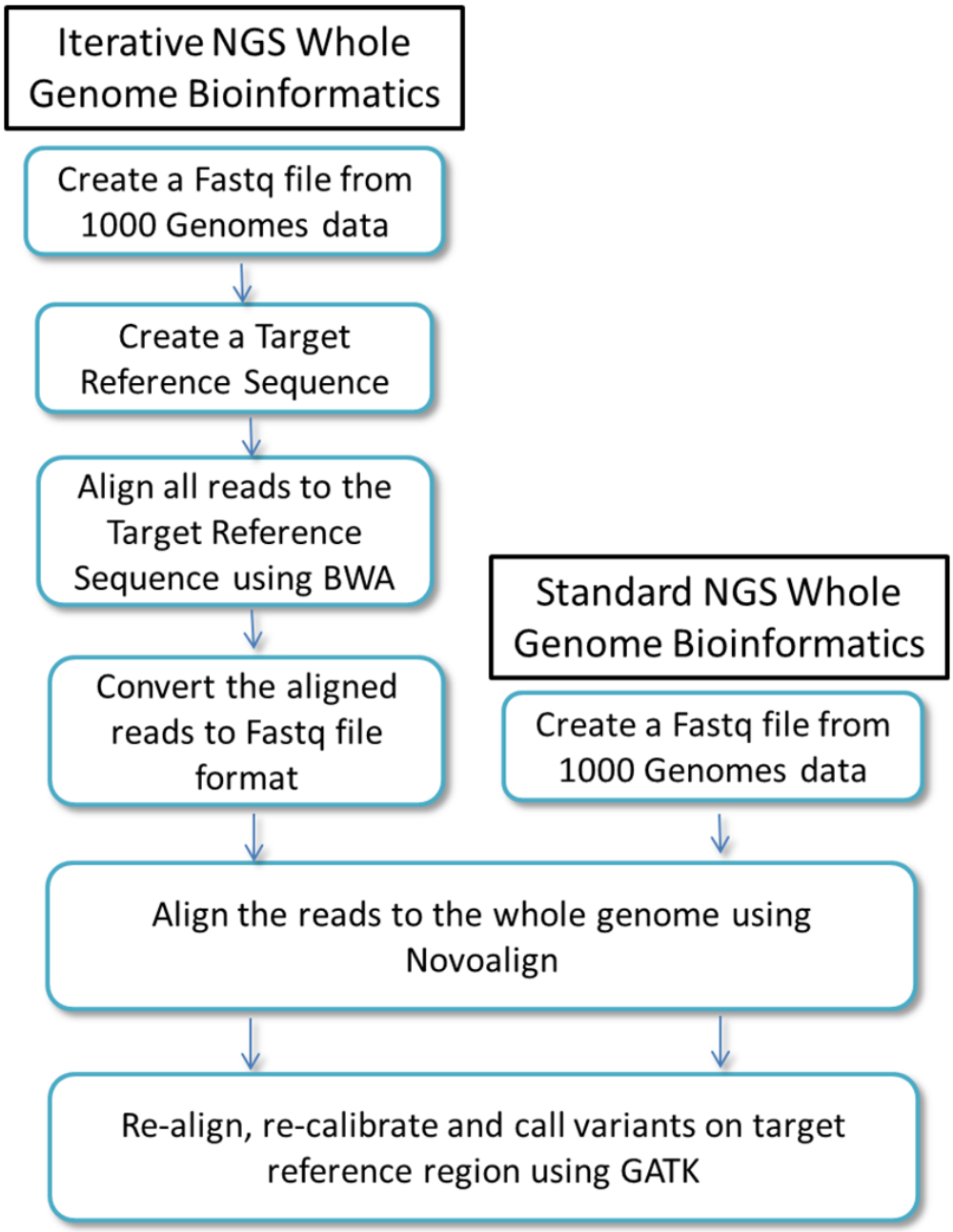

Figure 4.1: Basic components of the iterative workflow as compared to a standard NGS whole genome analysis. Using a smaller target reference sequence as the first step, the iterative workflow reduces the time to report variants on WGS data by 14-fold, taking a mere 5 CPU hours compared to the original time of 75 CPU hours. (Reproduced under the terms of Creative Commons Attribution License Appendix Figure A.3)

4.2.4 Iterative workflow

The different steps of the iterative workflow are displayed in **Figure 4.1**. The first step filters out the reads that do not map on the Target Reference Genome while the second step refines the alignment of the mapped reads by aligning them on to the Human Reference Genome. As previously explained, this step eliminates reads that have been forcibly mapped on the Target Reference Genome but would have aligned more accurately to another location of the Human Reference Genome. Since the first alignment step produced a BAM file with mapped reads information, the BAM file was converted in FASTQ format to perform the second alignment step.

We tested the iterative workflow with two aligners BWA and Novoalign. BWA is known to be faster than Novoalign, however, from our internal benchmark Novoalign produces slightly better read alignments. Since the workflow includes two alignment steps, we ran the workflow with different combinations of the two aligners. For any investigated combination of aligners, GATK was used to call variants.

4.3 Results

The CEPH family FASTQ files were processed with both the standard and iterative workflows. The two sets of results were compared with the genotypes obtained from the OMNI SNP platform reported by 1000 Genomes project. No additional processing was done on these reported data that were used as the gold standard in this study. 18634 OMNI genotypes included in the Target Reference Genome were used for accuracy estimates.

4.3.1 Results accuracy estimated from genotype calls

Using the 18634 genotypes of the OMNI SNP platform as 'truth', we assessed the accuracy of genotypes called by the standard workflow and the iterative workflow (**Table 4.1**). The iterative workflow results were produced with different combinations of aligners. Apart from one SNP on chromosome Y, all genotypes had adequate coverage. Results in **Table 4.1** highlight that the BWA-Novoalign workflow has slightly higher performance accuracy than the Novoalign-Novoalign workflow. Although not necessarily significant, this result suggests that the accuracy difference that we have observed between BWA and Novoalign in the first alignment steep has little impact on the quality of the final results. However, since BWA is significantly faster than Novoalign, the BWA-Novoalign workflow completes the task more than 5 times faster than the Novoalign-Novoalign workflow. Based on these findings, our remaining analysis is limited to the results obtained with the BWA-Novoalign workflow.

Workflow	Aligner used in step 1	Aligner used in step 2	Number of Concordant SNV	Number of Discordant SNV	% Concordance	Execution time (hrs)
Standard	Novoalign	-	18344	130	99.29	73.86
Iterative	BWA	BWA	17459	947	94.88	3.09
Iterative	**BWA**	**Novoalign**	**18435**	**129**	**99.30**	**4.98**
Iterative	Novoalign	Novoalign	18435	129	99.30	14.09
Iterative	Novoalign	BWA	18324	172	99.07	10.03

Table 4.1: Concordance of SNP data with variants from standard and iterative workflow. The numbers were calculated for sample the HapMap NA12878, but we got similar results for the other two sample evaluated. (Reproduced under the terms of Creative Commons Attribution License Appendix Figure A.3)

4.3.2 Genotyping calls missed by the standard and iterative workflows

About 99.3% of the genotypes called accurately by both workflows. When comparing the overlap between the 0.7% miscalled genotypes (i.e. 130 with the standard workflow and 129 with the iterative workflow), all but one of the genotypes were identical. This result reinforces the very similar performance of the two workflows and suggests that no significant bias was introduced by the iterative approach.

4.3.3 Performance accuracy of iterative and standard workflows on SNV and INDEL

We demonstrated that both the standard and iterative workflows had similar accuracy when compared to the OMNI SNP genotype calls. We then investigated the overlap between all the variants reported by the standard and the iterative workflow. These variants include single nucleotide variants (SNV) and insertions/deletions (INDEL). The large majority of the variants were called by both workflows (Table 4.2). We further analyzed discordant variants not called by the two workflows. Using the basic quality metrics of quality-by-depth (QD), strand bias and low read-depth coverage of less than 10 we observed that the majority of discordant variants had poor quality. For reference, less than 3% (236 out of 8754) of the concordant SNV had QD<5 or strand bias. To highlight the segregation that a basic quality metric like QD can achieve on low quality variant calls, **Figure 4.2** highlights the poor QD for discordant variant calls. We reviewed the 35 (out of 118) exclusive SNV/INDEL variant calls with QD>5. Out of these 35 variants, 19 have a clear strand bias. Of the remaining 16, 9 have a low coverage depth of less than 10 reads and 6 fall in a region with multiple (> = 5) homologous regions in the whole genome. This leaves just one exclusive variant of good quality that was called by our iterative workflow but not called by the standard workflow. Thus, we concluded that variants exclusively called by only one of the approaches are of low quality.

Workflow	Variant	Type	NA12878	NA12891	NA12892
Iterative	SNV	Shared	8754	8506	8809
Standard	SNV	Shared	8754	8506	8809
Iterative	SNV	Exclusive	38	34	39
Standard	SNV	Exclusive	62	57	70
Iterative	INDEL	Shared	975	902	905
Standard	INDEL	Shared	975	902	905
Iterative	INDEL	Exclusive	5	5	9
Standard	INDEL	Exclusive	13	11	14

Table 4.2: Evaluation of SNV and INDEL called by the iterative and standard workflow. Almost identical numbers were observed for the three samples evaluated with the majority of SNV and INDEL calls shared by both iterative and standard workflows. (Reproduced under the terms of Creative Commons Attribution License Appendix Figure A.3)

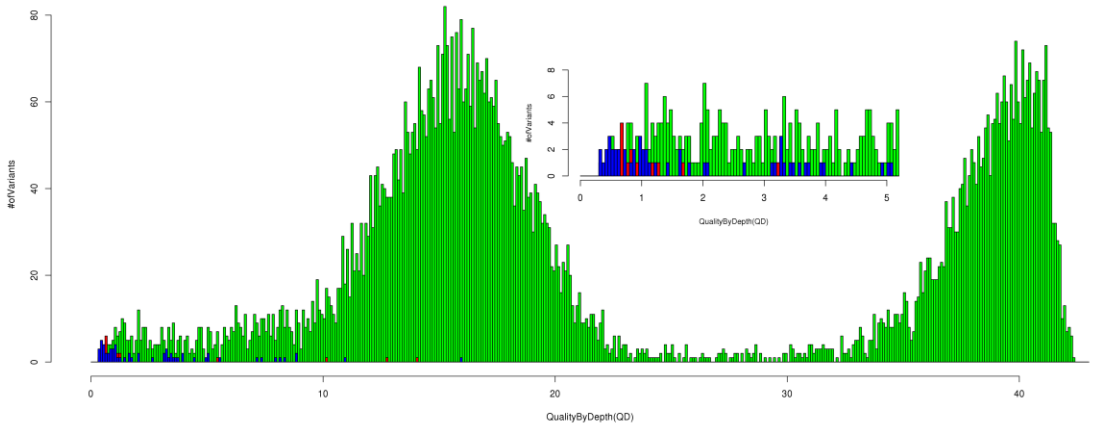

Figure 4.2: Distribution of QD (Quality by Depth) scores of variants, with the variants shared by both workflows in green, those identified by only the standard workflow in

blue and those identified by only the iterative workflow in red. Y-axis is the count of variants and the inset shows a zoomed-in view of QD<5 region.

4.3.4 Importance of the second alignment step

We explore the contribution of the second alignment step to the accuracy of variant calling. When using our iterative workflow, 27.5% of the reads aligned from 1st step to the CGD genes are aligned to a different location in the 2nd step. When calling variants directly after the first alignment step, only 83.25% concordance is obtained with the SNP chip data compared to 99.3% concordance when the reads are processed by the second alignment step. We also observed that more than 15,000 exclusive variants are reported after the 1st alignment step, this number dropping to 100 after the second alignment step. The second alignment step in the iterative workflow is therefore critical for accurate variant calling.

4.3.5 Reporting speed of clinically relevant variants

As shown in **Table 4.1**, the preferred iterative workflow takes less than 5 CPU hours to complete the alignment on the target reference genome and calling of the variants. The alignment of the remaining reads and variant calling took ~71 CPU hours.

A total of ~76 CPU hours was therefore needed to complete the full preprocessing of the whole genome experiment. In comparison, it also took ~76 CPU hours for the standard workflow to complete. We believe that this CPU overhead is acceptable in a clinical setup where the fast reporting of clinical variants could have a critical impact on patient's fate.

As a test, we extended Target Reference Genome to include all gene exons. The variants calls were reported in ~15 CPU hours, still an acceptable time compared to the 76 CPU hours needed for alignment of the whole genome using standard workflow.

4.4 Discussions

We developed and tested an iterative whole genome sequencing workflow designed to rapidly report variants in target genomic locations. The approach first focuses on aligning all the sequence reads on the target genomic locations and then realigning this subset of mapped reads to the reference genome. We benchmarked the accuracy of the iterative workflow against genotype data used a gold standard and also compared reported SNV to those reported by our standard whole genome sequencing workflow. Our results indicate that the standard and iterative workflows performed similarly well, with 99.3% accurate genotypes called The overlap between any variants (SNV and

INDEL) called by the standard and iterative workflow is also very high (98.8%), with most of the non-concordant calls being of low confidence (low QD score).

From this analysis, we can conclude that the iterative approach does not introduce significant noise or bias that would have a negative impact on the downstream calling of variants. With regards to time, using the Target Reference Genome, which included 2638 genes, allowed for the reporting of variants called in these regions in less than 5 hours. When extending the alignment to the whole exome, results were obtained in ~76 hours.

This iterative workflow can be particularly useful clinically when only a limited set of actionable variants need to be rapidly reported to clinicians. As compared to other published approaches, our iterative workflow does not require any additional investment in software or hardware. It is independent of the sequenced organism and the sequencing platform used as long as a reference genome is used to align the reads. Moreover, the iterative workflow can be implemented with any aligner or target reference region to swiftly report variants in those regions from whole genome sequencing data. This is relevant for applications related to regulatory regions, micro-RNA regions and potentially novel coding regions that may need a focused but fast evaluation.

Finally, the third step of the alignment, which consists of aligning the remaining reads, is the most time consuming. Interestingly, in our example, these reads are now naturally organized in independent islands covering the intergenic and intronic regions of the genome, facilitating the parallel processing of read realignment in these regions. Parallelization could be a means to significantly accelerate this final step. This option, however, was not investigated in this study.

Chapter 5: Clinical Filtering and Data Interpretation[4]

5.1 Introduction

There is growing availability of personal genomes and exomes in lieu of the decreasing cost and turn-around time of sequencing from Next Generation Sequencing (NGS) [24, 145]. Although potential for interpreting the functional consequences of results from sequencing data is lagging. There is great optimism and expectation that personal WGS/WES will benefit numerous individuals. This has led multiple institutions on the path to aggressively pursue application of NGS, assuming the evidence basis supporting this approach will evolve with experience.

Exome is ~1% of the whole genome but harbors more than 85% of known disease causing mutations [23]. WES provides good return on investment by being less expensive and faster than WGS. This has led to the widespread popularity and application of WES for diagnostic and clinical genetics along with characterization of various Mendelian disorders [39].

[4] This part of the work has been submitted for publication, the citation will be added as a supplemental file

Currently, few large-scale studies have comprehensively evaluated even the number of clinically interpretable variants from WES of an individual. There has been no in-depth assessment of how much or little the genetic variants from WES results correlate with the medical phenotype of an individual. The ClinSeq project aims to sequence 1000 subjects in order to determine genotype-phenotype association of variants and modes of returning results to individual subjects [146, 147]. To date, the group has identified 12 participants (1.2%) with gene mutation that leads to markedly increased risk of cancer. The next phase of this study would be to return carrier risk results to participants. A normal individual has been estimated to have 50–100 mutations in the heterozygous state that can cause a recessive Mendelian disorder as a homozygous genotype [24]. In a recent study, a team from Harvard Medical School utilized published recommendations of the National Heart, Lung, and Blood Institute (NHLBI) group [148] for the return of results [149]. They evaluated a representative sample of 160 disease-associated variants and extrapolated a conservative genome-wide estimate of 3955-12,579 variants per individual to be reported back. In another study, the group from Stanford [63] analyzed 12 WGS samples highlighting the lack of coverage in some of the 56 ACMG-reportable genes and large discordance of INDEL from two sequencing technologies. They curate 90 – 127 variants per person yielding only 2 – 6 personal disease risk findings per individual.

The goals of our study were to address these major clinical gaps, utilizing the Mayo Clinic Biobank. WES was conducted on 89 individuals who had donated blood to the Biobank, a research resource that has enrolled over 40,000 Mayo Clinic patient

volunteers since 2009 [150]. The 89 individuals selected were deceased at the time they were selected for sequencing, and all had a long history of medical care at Mayo Clinic. There was extensive information about what medical issues they had encountered in their lifetime. Analysis of WES data from this cohort provides baseline information on quality, filtering strategies, expected number of variants and depth of coverage from sequencing data along with correlation of this output with medical diagnoses on an individual basis.

5.2 Methods

5.2.1 Sample selection criteria

The Mayo Clinic Biobank resource is described in detail elsewhere [150]. Briefly, patients at Mayo Clinic who are 18 years or older, English speaking, have mental capacity to consent, and are residents of the United States are eligible for the Mayo Clinic Biobank. Recruitment was conducted via a mailed invitation to people scheduled for an appointment in general internal medicine, primary care internal medicine, family medicine, and preventive medicine and the specialty areas of obstetrics/gynecology and executive health. No threshold for health or disease was required to enroll in the Biobank.

The first group (group-1) of 39 Biobank participants (**Table 5.1**), including 24 males and 15 females, was selected for WES based on three major criteria: a) being deceased; b) long period of electronic medical record (EMR) information (median 15 and

mean 13 years); and c) later age of death. Preference was given to those with a death certificate available at the time of selection to confirm cause of death. Fifty-three deceased subjects were available at the time of the group-1 selection. Of note, nearly all of the confirmed causes of death were due to diseases common in the USA (cancer, heart/lung disease or trauma) which is consistent with causes of death in the general population of this age group. Average age of this group was 77.0 years at death with a range from 53-93 years.

As further funding became available for the project, the second group (group-2) of 50 Biobank participants (**Table 5.1**), including 27 males and 23 females, was selected from a total of 146 using the same criteria. In addition, we attempted to diversify the medical diagnoses among this group. To do this, we gave preference to individuals without a history of cancer, non-smokers and those with a younger age of death. This group had an average age of 72 years at death ranging from 28 to 93.

For the entire set of 89 subjects, there were 51 males and 38 females with an average age at death of 74.5 years (range 28 to 93, median 78 years).

88 Biobank WES samples	
74 years at death (range 28-93)	
51 Males	**38 Females**
75 years at death	74 years at death
32-91 years at death	28-93 years at death
Group 1	**Group 2**

39 samples		50 samples	
~77 years at death		~72 years at death	
53-93 years at death		28-93 years at death	
24 Males	15 Females	27 Males	23 Females

Table 5.1: Age and gender information of the 89 WES Biobank samples. The age information is shown by gender and also by group, as the 89 samples were sequenced in two groups or batches based on availability of resources and funding.

5.2.2 Patient Phenotype

We were interested in taking a high level view of how genotype might correlate with phenotype. To obtain the phenotype, I collaborated with a medical geneticist who abstracted all significant medical diagnoses from the electronic medical record (EMR) at Mayo Clinic for each study participant. 61% (n=55) of participants have more than 15 years of EMR while the remaining had median EMR of 12 years (inter-quantile range of 8 to 14 years). An average of 12 diagnoses per participant (range 2-20) were entered into a free-text field. Diagnoses made only as part of the terminal event (low blood pressure, low cardiac output, low renal output, respiratory failure, etc) were not included when they reflected end-of-life situation that could describe most terminally ill individuals. Many, but not all participants had seen multiple specialists. Undoubtedly this sort of chart audit misses some diagnoses and clinical findings depending on the reasons for each medical visit, but given the routine use of the self-reported past medical illnesses and review of systems forms, the records were fairly comprehensive in scope.

5.2.3 Sample preparation and DNA exome capture

The two groups of Mayo Clinic Biobank DNA samples were sequenced a year apart, based on resources becoming available for WES and analysis. The first group (n=39) was captured using Agilent's 50 Mb SureSelect Human All Exon chip while the second group (n=50) was captured using Agilent's SureSelect V4 + UTR kit. These enriched DNA samples were sequenced as one sample per lane on Illumina Genome Analyzer IIx flow cell and as three samples per lane on the Illumina HiSeq 2000, respectively. The sequencing was performed as 101 bp × 2 paired-end reads using TruSeq SBS sequencing kit version 1 and data collection version 1.1.37.0 followed by base-calling using Illumina's RTA version 1.7.45.0.

5.2.4 Bioinformatics Analysis and Annotation

The data was analyzed using our in-house workflow and updated TREAT annotation package [1]. Briefly, the sequencing reads were quality-checked using the FASTQC [91] and custom tools, aligned using Novoalign [140], re-aligned and re-calibrated using GATK [93, 94], followed by base-quality and variant-quality score recalibration and Single Nucleotide Variant (SNV), Insertion/Deletion (INDEL) calling using GATK **(Figure 5.1)**. The variants were then annotated using SeattleSeq [97, 112], SIFT [99], PolyPhen [110], Variant Effect Predictor and internal annotation databases and reported in VCF and Excel formats. Custom parsing scripts were used to include

Human Gene Mutation Database (HGMD) [138] and Online Mendelian Inheritance in Man (OMIM) [151] annotation. The list of data sources used for variant annotation is provided in **Table 5.2**.

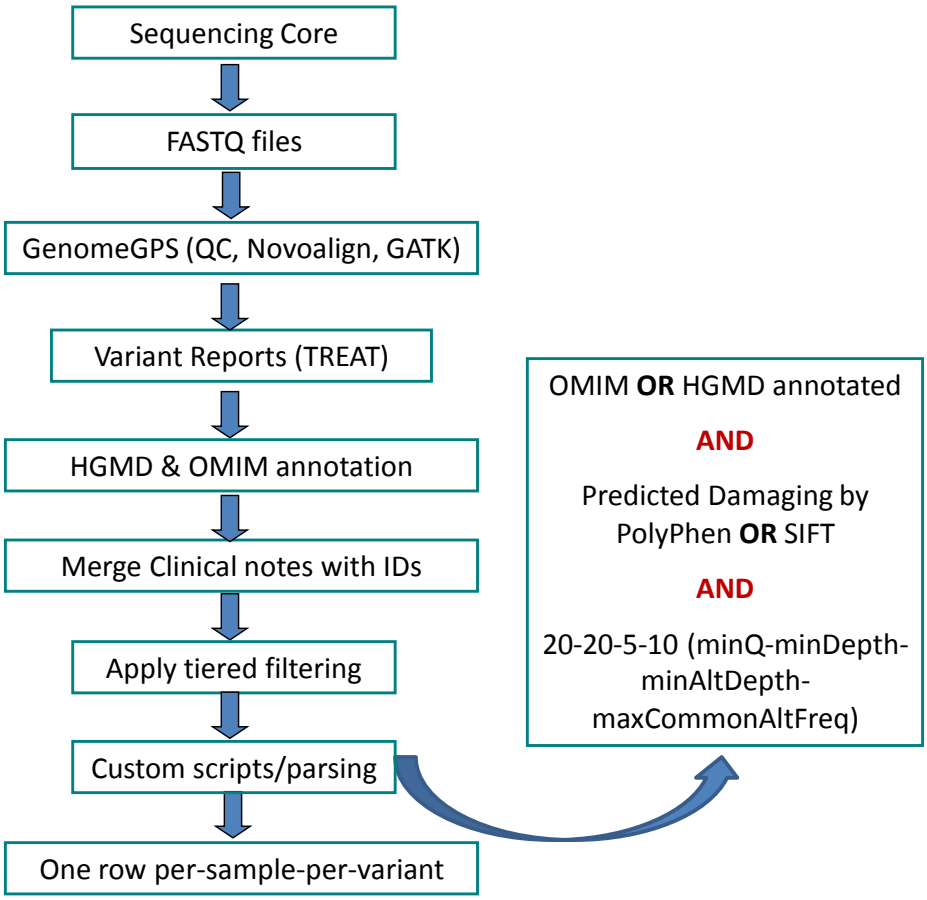

Figure 5.1: Flowchart for data analysis stages starting with sequencing to variant calling, filtering and annotation. The one row per-sample-per-variant was then used for medical phenotype evaluation

Data Sources used for variant annotation

Allele Frequency estimation

NCBI database of Single Nucleotide Polymorphisms database, version 135
Hapmap 2 and 3 Utah residents with Northern & Western European ancestry from CEPH collection

Hapmap 2 and 3 Youruba in Ibadan, Nigeria
Hapmap 3 and 3 Han Chinese in Beijing, China
1000 genomes project phase 1, European ancestry
1000 genomes project phase 1, West African ancestry
1001 genomes project phase 1, East Asian ancestry
National Heart Lung Blood Institute Exome Sequencing Project 5400 European Americans
National Heart Lung Blood Institute Exome Sequencing Project 5400 African Americans
200 Danish Exomes sequenced at BGI

Gene Annotation

ENSEMBL Gene ID
NCBI RefSeq
ENTREZ Gene ID
UCSC known gene

Functional Effect Prediction

Sorting Intolerant From Tolerant (SIFT)
PolyPhen2
SNP Effect Predictor

Phenotype Association

Human Gene Mutation Database (HGMD)
Online Mendelian Inheritance in Man (OMIM)
NHGRI Genome Wide Association Study catalog
Catalog Of Somatic Mutations In Cancer (COSMIC)
CLINVAR
Leiden Open Variation Database (LOVD)
Exome Variant Server
ALAMUT
Breast Cancer Information Core

Table 5.2: Sources used for Variant Annotation. The collection of various resources used and split by the type of information queried from the annotation sources

5.2.5 Concordance with Array Genotypes

Samples in group-1 were genotyped using either the Illumina Infinium HumanOmni 2.5-8 plus HumanOmni 2.5-S-8 arrays (N=4) or the Illumina Infinium Human Omni5-Quad array (N=35), samples in group-2 were genotyped using the Illumina Infinium HumanOmni 2.5v1.1 array (N=50). Comprehensive quality control (QC) was performed in order to verify sample quality. Relationship checking analysis of

the SNP array data identified 4 samples from group-1 that were incorrectly pipetted twice on the SNP array resulting in 4 of the WES samples having no array SNPs available for concordance analysis. Concordance rates comparing WES variant calls to array genotypes were calculated for each subject including all array genotypes in the WES capture region. All WES variant calls with read-depth > 10 were included in the concordance analysis.

5.2.6 Custom Variant Filtering

Because clinical correlation was an eventual goal, a customized filtering strategy was devised for SNV that included variants only in genes that had a listed HGMD or OMIM phenotype. They were required to have a minimum (PHRED-scale) mapping quality score of 20 (probability of being accurate > 99%), a minimum depth of 20 mapped reads and a minimum alternate (non-reference allele) read depth of 5 **(Figure 5.1)**. The variants with minor allele frequency of less than 10% in the 1000 genome [152], HapMap [153], NHLBI ESP exomes [28, 154] and 200 BGI Danish exomes [27] were included. In case of missense variants a deleterious *in-silico* prediction was required from either PolyPhen or SIFT. Variants common (>10%) to this dataset of 89 individuals were also filtered out. In HGMD, there are a variety of variants and genes that have had some functional work conducted but have not been associated with any disease state and these were removed as un-interpretable. In addition, reported non-disease associations such as improved memory performance were also removed. We defined the SNV variants as follows: 1) Tier I SNV – nonsense, loss of stop or splice site variants; and 2) Tier II SNV – missense variants.

With respect to INDEL, only those identified in genes with HGMD or OMIM phenotype and a minimum alternate (INDEL-supporting) read depth of 5 reads were included for interpretation. INDEL were also split by potential impact as: 1) Tier I INDEL – frame-shift or splice site; and 2) Tier II INDEL – codon change or codon deletion/insertion.

5.2.7 Gene Inheritance mode

Prior to evaluating genotype-phenotype correlation, it was necessary to assign each genetic entry to the inheritance pattern generally associated with disorders due to alterations in that gene. For each gene containing a variant included in the Tier 1 or Tier 2 files a medical geneticist assigned that gene to one of 7 groups: 1) autosomal dominant (AD); 2) autosomal dominant or autosomal recessive (AD/AR); 3) autosomal recessive only (AR): 4) X-linked recessive (XLR); 5) X-linked dominant (XLD); 6) Y-linked (YL); and 7) GWAS association only (SNP). For example, if the disorder was AD one might see a phenotype. For AR disorder with only one allele altered, no phenotype would be expected clinically. If the disorder were XLR, we would not generally expect females to manifest a phenotype. If the only association reported with a gene was a coding region GWAS hit, then we might see some tendency, but most likely not. Assigning these genes to one of these 7 groups was a very inexact science: some classical autosomal dominant disorders have also been associated with GWAS hits for entirely different phenotypes. In

general, a good faith effort was made to assign each gene into the most established category for that gene.

5.2.8 Genotype-phenotype correlation scoring

Once the inheritance pattern and the clinical phenotypes were added to the Tier 1 and Tier 2 variants, a medical geneticist manually scored each genetic variant by comparing the participants' disease phenotypes with all of the phenotypes that had been reported in that gene (not restricted to a specific variant). The phenotype listings used were obtained from both HGMD and OMIM and were compared side by side with the patient disease diagnosis list. A "Yes" score meant the participant phenotype overlapped in some way with one of the phenotypes reported in that gene. A "No" meant there was no overlap seen. An "X" indicated inability to assess for genotype-phenotype correlations. For example a gene variant associated with prostate cancer found in a woman; or a gene resulting in abnormal sperm shape, which would not have been identified on typical medical visits; or variants that reduced risk for various conditions—realistically there was no way to assess any effect.

5.2.9 Coverage analysis of 56 ACMG-reportable genes

Individual gene level coverage analysis was performed to evaluate efficient reporting of variants from the 56 genes for which clinical reporting has been

recommended by the ACMG (ACMG-reportable genes) [155]. This was done using BEDTools [156] and in-house developed scripts on the aligned BAM files from WES data. The per-nucleotide coverage was used to identify coding regions in ACMG-reportable genes with less than 10x read-depth coverage.

5.2.10 Variants in the 56 ACMG-reportable genes

The resultant variants from our custom filtering were subset to only those found in any of the 56 ACMG reportable genes [155] using custom perl scripts.

5.2.11 Cancer specific genes and cancer phenotypes

The subset of genes known to be linked to cancer in the ACMG-reportable list of 56 genes [155] and those now included on some of the cancer-specific NGS panels offered clinically (**Table 5.3**) were collected. The resulting 58 genes were used to select all Tier 1 or Tier 2 SNV and INDEL potentially disrupting the function of these genes. Manual curation was conducted for each variant in this list and variants were scored using the scale: 1=neutral/non-pathogenic, 2=likely neutral/non-pathogenic, 3=variant of uncertain significance; 4=likely pathogenic and 5=pathogenic. The 89 Biobank participants were separated into those with cancer (excluding non-melanoma skin cancers) and those without cancers to compare the genetic results.

Gene ID	Cancer NGS Panels	ACMG reportable genes	Other genes		Gene ID	Cancer NGS Panels	ACMG reportable genes	Other genes
AKT1	1	0	0		*PMS2*	1	1	0
APC	1	1	0		*POLD1*	1	0	0
ATM	1	0	0		*POLE*	1	0	0
ATR	1	0	0		*PRSS1*	1	0	0
BAP1	1	0	0		*PTEN*	1	1	0
BARD1	1	0	0		*RAD50*	1	0	0
BMPR1A	1	0	0		*RAD51*	1	0	0
BRCA1	1	1	0		*RAD51C*	1	0	0
BRCA2	1	1	0		*RAD51D*	1	0	0
BRIP1	1	0	0		*RET*	1	1	0
CDH1	1	0	0		*SDHB*	1	1	0
CDK4	1	0	0		*SDHC*	1	1	0
CDKN2A	1	0	0		*SDHD*	1	1	0
CHEK1	1	0	0		*SMAD4*	1	0	0
CHEK2	1	0	0		*STK11*	1	1	0
CTNNA1	1	0	0		*TP53*	1	1	0
FAM175A	1	0	0		*TP53BP1*	1	0	0
GALNT12	1	0	0		*VHL*	1	1	0
GEN1	1	0	0		*XRCC2*	1	0	0
GREM1	1	0	0		*MEN1*	0	1	0
HOXB13	1	0	0		*RB1*	0	1	0
MLH1	1	1	0		*SDHARF2*	0	1	0
MRE11A	1	0	0		*TSC1*	0	1	0
MSH2	1	1	0		*TSC2*	0	1	0
MSH6	1	1	0		*WT1*	0	1	0
MUTYH	1	1	0		*NF2*	0	1	0
NBN	1	0	0		*FLCN*	0	0	1
PALB2	1	0	0		*BAP1*	0	0	1
PIK3CA	1	0	0		*FH*	0	0	1

Table 5.3: List of 58 cancer related genes evaluated for the 89 WES samples. The three columns denote binary presence or absence of these cancer pre-disposition genes in the various clinical NGS gene panels, the list of 56 ACMG-reportable genes and other genes selected based on our experience.

5.3 Results

5.3.1 Data Metrics

An average of 270 million reads (140 – 421 million) and 116 million reads (69 – 147 million) of sequence data were obtained for the 39-sample (group 1) and 50-sample (group 2), respectively **(Table 5.4)**. The difference in throughput is due primarily to differences in the number of samples sequenced per lane; one sample per lane sequenced for group 1 compared to three samples per lane sequenced for group 2. More than 96% of the targeted region was covered with at least 10 reads (10x coverage) for group 2 samples compared to 88% for group 1 samples. Overall, the Agilent SureSelect V4+UTR capture kit used for the group 2 had much better capture efficiency with a greater number of all sequenced reads mapping to the intended capture region. This kit also provided improved performance by way of greater balance and uniformity in the overall coverage of the capture region. Due to the larger capture region in Agilent SureSelect V4+UTR, there were a greater number of variants (SNV and INDEL) reported for group 2 WES samples. However, when evaluating the final filtered lists of Tier1 and Tier2 variants, there were minimal differences in the number of variants from the two groups of WES samples **(Table 5.4).**

We found an average of 149 (range 134-172) and 191 (range 169-215) Tier 1 SNVs in group 1 and group 2 samples, respectively. Following our custom variant

filtering strategy described in the methods section, we found an average of 4 (range 0-7) and 4 (range 1-9) Tier 1 SNV per sample for group 1 and group 2, respectively. For Tier 2 SNVs, we observed an average of 8751 (range 8164-9504) and 11039 (range 10650-11387) variants per sample in group 1 and group 2, respectively that reduced to an average of 64 (range 44-88) and 63 (range 47-95) per sample, respectively after filtering.

	Group 1 (39 samples)			**Group 2 (50 samples)**		
	Mean	Max	Min	Mean	Max	Min
Total Reads (in millions)	270	421	141	116	147	69
Mapped Reads (in millions)	259	405	134	115	146	68
Mapped Reads within Targeted Region (in millions)	134	212	58	92	117	38
% Coverage of Targeted Region at 5x	94.34	97.3	91.98	98.9	98.93	98.76
% Coverage of Targeted Region at 10x	91.13	94.7	88.16	98.1	98.78	96.05
% Coverage of Targeted Region at 20x	87.52	91.8	82.98	94.4	96.52	87.62
Total SNV in the Coding Regions	42,661	49,065	38,444	64,696	68,875	61,996
Tier1 = Stop gained/lost / Start lost / splice acceptor/donor	149	172	134	191	215	169
Tier1 after filtering	4	7	0	4	9	1
Tier2 = missense	8751	9504	8164	11039	11387	10650
Tier2 after filtering	65	88	44	63	95	47
Total INDELs in the Coding Regions	3,087	3,860	2,517	7,332	7,862	6,694
Tier 1 after filtering	3	7	0	3	10	0
Tier 2 after filtering	6	15	0	9	15	3
Total Variants after filtering (Tier 1 & 2, SNV & INDEL)	78	107	56	79	119	62

Table 5.4: Number of reads and variants per sample from the 89 WES individuals. The per-sample data is separated into the two groups in which the actual sequencing was performed. Mean, maximum and minimum metrics are shown for each field. The Tier 1 and Tier 2 fields are defined in the text and were selected based on the tool SNP Effect Predictor's [4] effect severity as being high and medium respectively. The filtering refers to the custom thresholds defined in Figure 5.1 as the gene having OMIM or HGMD annotation, and minimum variant phred-scale quality 20 (>99% probability of being accurate), minimum read-depth 20, minimum non-reference allele depth 5 and maximum alternate frequency 0.1 (and predicted damaging by SIFT or PolyPhen in case of missense SNV)

5.3.2 Concordance with array genotype calls

Comparison of genotype calls from array chips provides a good assessment of sample quality. After restricting to variants in the target capture region we calculated the fraction of variant genotypes in the array that are also NGS called variants as well as the genotype concordance rate. All samples had high quality data having sample call rates > 95%. The concordance between NGS called variants and array genotypes was indicative of high quality sequencing (concordance > 99% for all samples).

5.3.3 Genotype-Phenotype correlation

The clinical notes of the sequenced individuals and the phenotype annotations of genetic variants from WES data were manually evaluated. **Table 5.5** shows a high level summary of the proportion of variants that correlated with any known phenotypic finding. The majority of medical diagnoses observed for these Mayo Clinic Biobank

individuals were common complex genetic disorders, similar to that seen in the general population (atherosclerotic cardiovascular disease, Type 2 Diabetes, obesity, degenerative joint disease, cataracts, osteoporosis, etc.), for which there is little useful genotypic information. Overall, 3% (N=202) of the total 7046 Tier-1 and Tier-2 SNV/INDEL variants had a matching phenotype from clinical chart review while 53% (N=3710) variants did not exhibit a correlating phenotype. The remaining 44% (N=3134) variants were unable to be assessed for genotype-phenotype correlations.

Focusing on the variants identified in genes known to have autosomal dominant expression (AD or AD/AR), there were 129 Tier-1 variants (73 SNVs and 56 INDELs) identified. Of these, 4 Tier-1 SNVs and 5 Tier-1 INDELs were in genes for which there was a phenotypic match (**Table 5.6**) On the other hand, 66 Tier-1 SNVs and 50 Tier-1 INDELs in AD or AD/AR genes did not have an apparent phenotypic match to the individual's medical record (**Table 5.7**). Among the 1091 Tier-2 SNVs in AD or AD/AR genes, we observed 42 with phenotypic matches (**Table 5.8**) compared with 1006 Tier-2 SNVs with no apparent phenotype match to the individual's medical record.

Tier-1 SNV	# of variants	Match	No Match	Cannot assess	**Tier-1 INDEL**	# of variants	Match	No Match	Cannot assess
AD	60	3	55	2	AD	39	4	34	1
AD/AR	13	1	11	1	AD/AR	17	1	16	0
Digenic	2	0	2	0	Digenic	7	1	6	0
AR	92	2	0	90	AR	84	0	0	84
XLR	11	1	2	8	XLR	0	0	0	0
XLD	2	0	2	0	XLD	0	0	0	0
SNP	181	11	152	18	SNP	120	4	103	13

Others	13	0	0	13	Others	3	0	0	3
Total	374	18	224	132	**Total**	270	10	159	101
Tier-2 SNV	# of variants	Match	No Match	Cannot assess	**Tier-2 INDEL**	# of variants	Match	No Match	Cannot assess
AD	914	27	861	26	AD	212	2	179	31
AD/AR	177	15	145	17	AD/AR	6	4	2	0
Digenic	58	4	50	4	Digenic	0	0	0	0
AR	2129	2	0	2127	AR	113	2	0	111
XLR	51	0	12	39	XLR	35	0	11	24
XLD	11	0	8	3	XLD	1	0	1	0
SNP	2326	107	1897	322	SNP	198	11	161	26
Others	37	0	0	37	Others	134	0	0	134
Total	5703	155	2973	2575	**Total**	699	19	354	326

Table 5.5: Phenotypic overlap of any type with gene containing the variant of interest. Tier 1 variants are most likely to be significant; Tier 2 variants contain many variants of uncertain clinical significance (See text for definitions). Autosomal Dominant (AD) and Recessive (AR), X-linked Dominant (XLD) and recessive (XLR) and GWAS associated Single Nucleotide Polymorphism (SNP) were compiled as separate lists

Gene	HGMD and OMIM descriptions	Matching finding in Biobank participant
SMAD3	***Aneurysms***-osteoarthritis syndrome|Aortic aneurysms & dissections with early-onset osteoarthritis|**Osteoarthritis**|Thoracic aortic aneurysms & dissections; Loeys-Dietz syndrome, type 3	**degenerative joint disease**, ***aneurysm***
MSR1	***Atherosclerosis***, increased risk, association with|Barrett oesophagus/oesophageal adenocarcinoma|Chronic obstructive pulmonary disease, in smokers, association with|Prostate cancer|Prostate cancer, association with.	***atherosclerosis***
TULP3	***Glaucoma***, primary open angle (due to copy number variant in this gene)	***glaucoma suspect***
FLG	***Eczema*** |Eczema, association with|Eczema, association with and Asthma, association with|Fissured skin on hands of patients without dermatitis|Genetic modifier in pachyonychia congenita|Hand eczema, association|Ichthyosis vulgaris|Peanut allergy, association with|Psoriasis|Psoriasis vulgaris|Psoriasis,	***eczema***

	increased risk, association	

Table 5.6: Four Autosomal Dominant (AD) genes or AD/AR genes with Tier 1 SNV variants for which there was a match (shown in *bold*) with phenotype in a biobank participant

Gene	HGMD and/or OMIM descriptions (some truncated)	Frequency
AADAC	Tourette syndrome |Reduced enzyme activity	1
ALK	Neuroblastoma	1
ANO7	Glaucoma, primary congenital	1
ASXL1	Bohring-Opitz syndrome|Systemic mastocytosis with associated non-mast cell lineage disease	2
BCMO1	Hypercarotenemia and hypovitaminosis A|Altered beta-carotene metabolism, association with	1
CARD14	Psoriasis, association with|Psoriasis|Pityriasis rubra pilaris	1
CATSPER2	Asthenoteratozoospermia & deafness, non-syndromic	1
COL8A2	Glaucoma, primary open angle|Fuchs corneal dystrophy	1
COMP	Pseudoachondroplasia|Multiple epiphyseal dysplasia|Early-onset osteoarthritis	1
CRYBA4	Cataract and microcornea|Cataract, lamellar|Microphthalmia	1
DPP6	Autism spectrum disorder |Ventricular fibrillation, idiopathic	1
EFHC1	Myoclonic epilepsy, juvenile|Intractable epilepsy of infancy|Idiopathic epilepsy, generalised	1
FAM83H	Amelogenesis imperfecta, hypocalcified|Amelogenesis imperfecta, hypoplastic local	1
FREM1	Bifid nose, renal agenesis & anorectal malformations syndrome|Craniosynostosis, isolated metopic|Manitoba-oculo-tricho-anal syndrome	1
GON4L	Intellectual disability	1
HBM	Thalassaemia alpha	1
KRT83	Monilethrix	12
MSR1	Atherosclerosis, increased risk, association with|Barrett oesophagus/oesophageal adenocarcinoma|Chronic obstructive pulmonary disease, in smokers, association with|Prostate cancer|Prostate cancer, association with	4
MYBPC3	Hypertrophic cardiomyopathy with inclusion body myositis|Increased left ventricular wall thickness|Left ventricle dysfunction in CAD, association with|Skeletal myopathy, association with|Sudden infant death syndrome |Dilated cardiomyopathy|Cardiomyopathy, left-ventricular noncompaction|Cardiomyopathy, left ventricular noncompaction|Cardiomyopathy, hypertrophic/dilated|Cardiomyopathy, hypertrophic|Cardiomyopathy, dilated|Cardiomyopathy, association with|Cadiomyopathy, dilated	4
MYO1A	Sensorineural deafness, nonsyndromic	1
NBAS	Short stature, optic atrophy & Pelger-Huet	1

NOL3	Cortical myoclonus	1
OBSCN	Cardiomyopathy, hypertrophic|Glioblastoma|Potential protein deficiency	1
PITPNM 3	Cone dystrophy, autosomal dominant|Cone dystrophy	1
PLCB4	Auriculocondylar syndrome	1
POLR1C	Treacher-Collins syndrome	1
PRPH	Amyotrophic lateral sclerosis|High myopia	1
RAD21	Cornelia de Lange-like syndrome	1
RASA1	5q14.3 neurocutaneous syndrome|Arteriovenous fistula|Arteriovenous malformation |Capillary malformation-arteriovenous malformation|Capillary malformations|Sturge-Weber syndrome	1
RNASEL	Ribonuclease L deficiency, association with|Ribonuclease L deficiency|Prostate, cancer, protection against, association with|Prostate cancer, association with |Prostate cancer	1
RP1L1	Macular dystrophy, occult|Potential protein deficiency	1
SLC6A2	Reduced gene expression|Orthostatic intolerance and tachycardia|Major depression|Decreased transport activity|Attention-deficit hyperactivity disorder, association with	1
TBC1D4	Insulin resistance	1
TRPA1	Episodic pain syndrome|Paradoxical heat sensation, association with	1
TRPM2	Amyotrophic lateral sclerosis and parkinson disease	1
TTF2	Autism	1
TTN	Tibial muscular dystrophy|Potential protein deficiency|Myopathy with early respiratory failure|Myopathy with cellular aggregates|Myopathy|Muscular dystrophy |Cardiomyopathy, hypertrophic|Cardiomyopathy, dilated|Arrhythmogenic right ventricular cardiomyopathy	1
AADAC	Tourette syndrome |Reduced enzyme activity	1
ALK	Neuroblastoma	1
FLG	Eczema |Eczema, association with|Eczema, association with and Asthma, association with|Fissured skin on hands of patients without dermatitis|Genetic modifier in pachyonychia congenita|Hand eczema, association|Ichthyosis vulgaris|Peanut allergy, association with|Psoriasis|Psoriasis vulgaris|Psoriasis, increased risk, association …	4
SH3TC2	Charcot-Marie-Tooth disease 1|Charcot-Marie-Tooth disease 4C|Hereditary motor & sensory neuropathy	1
VWF	Von Willebrand disease 2n/1|Von Willebrand disease 2n|Von Willebrand disease 2m |Von Willebrand disease 2c|Von Willebrand disease 2b-like|Von Willebrand disease 2b|Von Willebrand disease 2u|Von Willebrand disease 3 |Von Willebrand disease, association with|Von Willebrand disease, quantitative type, association with|Von Willebrand, …	1
SEMA3E	CHARGE syndrome	1
APOB	Hypobetalipoproteinaemia|Hypobetalipoproteinemia-induced nonalcoholic steatohepatitis|Hypocholesterolaemia |Hypocholesterolaemia, association with|Increased apoB and cholesterol levels, association with|Increased cholesterol levels|Ischaemic stroke, association with |Oligoasthenoteratozoospermia, association with|Hypertriglyceridaemia |Hypercholesterolaemia |Altered APOB	1

	levels |Aortic stenosis, association with|Apolipoprotein B deficiency|Coronary artery disease, association with|Coronary heart disease|Coronary heart disease, association with|HDL cholesterol, association with |Hepatitis C virus infection, association with	
DOCK8	Mental retardation|Immunodeficiency, combined|Hyper-IgE syndrome, autosomal recessive	1
BRCA2	Ovarian / peritoneal carcinoma|Oesophageal squamous cell carcinoma|Oesophageal carcinoma |Oesophageal cancer, association with|Ocular melanoma|Medulloblastoma |Male BC risk|Lung cancer |Lunc cancer|Liver cancer|Ovarian cancer|Ovarian carcinoma|Ovarian insufficiency, primary |Reactive lymphoid hyperplasia |Prostate cancer, high-grade|Prostate cancer |Promyelocytic leukemia |Potential protein deficiency|Poorer survival in prostate cancer patients|Peritoneal carcinoma|Pancreatic cancer |…	1
LDLR	Stroke, increased risk, association with|Reduced plasma LDL cholesterol, association with|Increased plasma LDL cholesterol|Hypercholesterolaemia|Coronary artery disease, increased risk in low BMI individuals|Coronary artery disease, association with|Altered transcription	1
EDN3	Waardenburg-Hirschsprung disease|Waardenburg syndrome 4B|Waardenburg syndrome 4|Shah-Waardenburg syndrome|Phenotype modification in HSCR|Hirschsprung disease|Central hypoventilation syndrome	1

Table 5.7: AD genes or AD/AR genes that are either dominant or recessive, with Tier 1 SNV variants for which there was a NO match with phenotype (n=55 examples)

Gene	HGMD and OMIM descriptions (some truncated)	Matching phenotype
ACTN4	Glomerulosclerosis, focal and segmental	Maybe: chronic renal failure
ASB10	***Glaucoma,*** primary open angle	glaucoma
BFSP2	***Cataract***, progressive, juvenile onset|Cataract, Y-suture|Congenital cataract|Diffuse cortical cataract with scattered lens opacities	cataracts
BRCA1	***Breast-ovarian cancer***, familial ; Papillary thyroid cancer, reduced risk|Pancreatic cancer |Pancreatic adenocarcinoma|Ovarian carcinoma|Ovarian cancer, association with|Ovarian cancer|Ovarian / peritoneal carcinoma|Neuronal migration defect|Mean number of breaks per cell, association with|Peritoneal carcinoma|…	Breast ca
CORIN	***Hypertension***, association with|Impaired brain natriuretic peptide processing	high BP
CRYBA4	***Cataract*** and microcornea|Cataract, lamellar|Microphthalmia	cataracts
DIRC2	***Renal cancer***	renal ca
ENPP1	Myelopathy (OPLL)|Myelopathy (OPLL), association with|Obesity & type 2 diabetes, association with|***Obesity in metabolic syndrome***, association with|Obesity, association with|Pseudoxanthoma elasticum|Rickets, hypophosphataemic|Rickets, hypophosphataemic & OPLL|Rickets,	metabolic syndrome

	hypophosphataemic, autosomal recessive|Major cardiovascular events in high risk individuals|Liver damage in NAFLD|Decreased kidney function|Diabetes, association with |Diabetic nephropathy, increased risk, association with|Generalized arterial calcification of infancy|Generalized arterial calcification of infancy and pseudoxanthoma elasticum|Hypertriglyceridaemia in males, association with|Hypertriglyceridemia in males, association with|Idiopathic infantile arterial calcification|Insulin resistance, association with	
EPHA2	**Cataracts**; posterior polar, 1, 116600 (3); Cataract, age-related	cataracts
EPHA2	**Cataracts**; posterior polar, 1, 116600 (3); Cataract, age-related	cataracts
EPHA2	**Cataracts**; posterior polar, 1, 116600 (3); Cataract, age-related	cataracts
EPHA2	**Cataracts**; posterior polar, 1, 116600 (3); Cataract, age-related	cataracts
ESPN	***Hearing loss,*** non-syndromic|Hearing loss, autosomal dominant|Deafness and vestibular areflexia	SNHL
FLNC	***Arrhythmia*** & myofibrillar myopathy, late-onset|Distal myopathy|Myopathy, myofibrillar	arrhythmias
FN1	Glomerulopathy with fibronectin deposits|Autism	maybe microhematuria
FN1	Glomerulopathy with fibronectin deposits|Autism	maybe microhematuria
KCNE2	***Cardiac arrhythmia***|Long QT interval, drug induced, association with|Long QT syndrome|QTc interval, association with	maybe LBBB/Vtach
LIPI	Hypertriglyceridaemia|Plasma HDL cholesterol|Plasma HDL cholesterol, association with	high lipids
LIPI	Hypertriglyceridaemia|Plasma HDL cholesterol|Plasma HDL cholesterol, association with	high lipids
LIPI	Hypertriglyceridaemia|Plasma HDL cholesterol|Plasma HDL cholesterol, association with	high lipids
MED13L	Autism|***Colorectal cancer***, increased risk, association with |Congenital heart defect|Intellectual disability, nonsyndromic, no cardiac involvement	rectal ca
MET	***Papillary renal carcinoma***|Lymphoedema |Gastric cancer|Diffuse large B-cell lymphoma |Colorectal cancer |Autism, association with	kidney cancer but not papillary type
MLH3	Oesophageal cancer |Endometrial cancer|***Colorectal cancer***, non-polyposis|Colorectal cancer, increased risk	colon ca
MSH3	Radiosensitivity in breast cancer patients, association with|Proximal colon cancer, increased risk, association with|Colorectal cancer, increased risk, association with|***Colorectal cancer*** |Colon cancer, association with|Colon cancer ; Endometrial carcinoma	colon ca
MYH6	***Sick sinus syndrome***, increased risk, association with|Congenital heart defects|Cardiomyopathy, hypertrophic|Cardiomyopathy, dilated |Atrial septal defect	maybe LBBB/Vtach
MYH6	***Sick sinus syndrome***, increased risk, association with|Congenital heart	sick sinus

	defects|Cardiomyopathy, hypertrophic|Cardiomyopathy, dilated |Atrial septal defect	syndrome
MYPN	Cardiomyopathy, dilated|Cardiomyopathy, dilated / ***hypertrophic|Cardiomyopathy***, hypertrophic|Cardiomyopathy, restrictive	possible hypertrophic cardiomyopathy
NEUROD1	***Diabetes, type 2,*** early-onset |Diabetes, permanent neonatal|Diabetes, MODY|Diabetes mellitus, type 2, association with|Diabetes mellitus, type 2|Diabetes mellitus, type 1, association with	DM2
PTCH1	Multiple ***basal cell carcinoma*** |Nevoid basal cell carcinoma syndrome|Odontogenic keratocysts|Short stature, intellectual disability & facial dysmorphism|Skin cancer, association with|Microcephaly and developmental delay |Keratocystic odontogenic tumours, non-syndromic|Basal cell carcinoma|Gorlin syndrome |Gorlin syndrome and autism|Gorlin-syndrome-related odontogenic keratocysts|Holoprosencephaly	nonmelanoma skin ca
RUNX1	Thrombocytopaenia |Thrombocytopaenia and acute myeloid leukaemia|Thrombocytopaenia, association with|Thrombocytopaenia, non-syndromic with ***myelodysplasia***|Rheumatoid arthritis, susceptibility, association|Platelet disorder, familial & myeloid leukaemia|Platelet disorder, familial|Mental retardation, short stature & thrombocytopaenia|Leukaemia, chronic myelomonocytic|Developmental delay, congenital anomalies & thrombocytopenia|Acute myeloid leukaemia, myelodysplastic syndrome-related	yes myelodysplasia
SERPIND1	Heparin cofactor 2 deficiency	maybe-thombophilia but in context of pancreatic cancer
TTN	Tibial muscular dystrophy|Potential protein deficiency|Myopathy with early respiratory failure|Myopathy with cellular aggregates|Myopathy|Muscular dystrophy |Cardiomyopathy, hypertrophic|Cardiomyopathy, dilated|Arrhythmogenic right ventricular cardiomyopathy	possible hypertrophic cardiomyopathy
TTN	Tibial muscular dystrophy|Potential protein deficiency|Myopathy with early respiratory failure|Myopathy with cellular aggregates|Myopathy|Muscular dystrophy |***Cardiomyopathy, hypertrophic***|Cardiomyopathy, dilated|Arrhythmogenic right ventricular cardiomyopathy	possible hypertrophic cardiomyopathy
AMPD1	Adenosine monophosphate deaminase deficiency|Features of ***metabolic syndrome*** in coronary artery disease, association with	metabolic syndrome
BRCA2	Breast-ovarian cancer, familial; Fanconi anemia; **prostate** and pancreatic cancer…	prostate ca
BRCA2	Breast-ovarian cancer, familial; Fanconi anemia; prostate and ***pancreatic cancer***…	pancreatic ca

BRCA2	***Breast***-ovarian cancer, familial; Fanconi anemia; prostate and pancreatic cancer…	breast ca
BRCA2	***Breast***-ovarian cancer, familial; Fanconi anemia; prostate and pancreatic cancer…	breast ca
CFH	Membranoproliferative glomerulonephritis|Macular degeneration, exudative age-related, association with|***Macular degeneration, age-related, association*** with|Lung cancer, increased risk|Kidney function, association with|Inflammation, visual impairment, and cardiovascular mortality, association with|IgA nephropathy |Hemolysis, elevated liver enzymes & low platelet count|Haemolytic uraemic syndrome, atypical|Membranoproliferative glomerulonephritis, association|Membranoproliferative glomerulonephritis, association with|Meningococcal disease, lower risk, association with|Thrombotic microangiopathy following transplantation|Thrombotic microangiopathy following kidney transplantation|Stargardt disease …	macular degeneratio n
FLG	***Eczema*** |Eczema, association with|Eczema, association with and Asthma, association with|Fissured skin on hands of patients without dermatitis|Genetic modifier in pachyonychia congenita|Hand eczema, association|Ichthyosis vulgaris|Peanut allergy, association with|Psoriasis|Psoriasis vulgaris|Psoriasis, increased risk, association with|Autism|Atopic eczema |Allergen sensitization, association with|Atopic asthma |Atopic asthma and dermatitis, asssociation with|Atopic asthma, association with|Atopic dermatitis |Atopic dermatitis / eczema herpeticum|Atopic dermatitis & asthma, increased risk, association with|Atopic dermatitis and asthma|Atopic dermatitis, increased risk, association with|Atopic dermatitis, reduced risk, association with|Atopic disease, association with	eczema
LRP6	Fragility fractures, increased risk, association with|Crohn's disease, early-onset ileal, association with|Coronary artery disease, early|Carotid artery ***atherosclerosis in hypertension, incr risk***, association|Alzheimer disease, late onset, association with	extensive peripheral vascular disease
MC1R	Vitiligo protection|UV-induced skin damage, vulnerability to|Red hair, increased risk|Photoaging, association with|Melanoma, in CDKN2A mutation carriers, association with|Melanoma, association with|***Melanoma*** |Increasing size of congenital melanocytic nevi, association with|Impaired activity|Basal cell carcinoma|Depression, association with|Ephelides, increased risk, association with|Fair hair, association with|Functional melanin, lower levels, association with|Glucocorticoid deficiency without pigmentation	Melanoma x2
NBN	Hepatic cancer, association with.|Lung cancer, association with |Lung cancer, increased risk, association with|Medulloblastoma|Melanoma|Nasopharyngeal carcinoma, increased risk, association with|NBS severity|Nijmegen breakage syndrome|Nijmegen breakage syndrome with macrocephaly, schizencephaly and large CSF spaces.|Ovarian carcinoma|Solid tumours|Gastrointestinal cancer, association with|Ganglioglioma|Acute lymphoblastic leukaemia|Acute lymphoblastic leukaemia, association with|Acute lymphoblastic leukaemia, increased risk|Aplastic anaemia|Bladder cancer in smokers / meta-analysis, association with|Breast cancer|Breast cancer, increased risk, association with |Breast cancer, reduced risk|Cancer, increased risk, association with |Colorectal cancer|Fertility defects[**prostate** not listed but is in literature]	prostate ca
PPARG	***Insulin resistance, diabetes*** and hypertension|Insulin sensitivity in	DM2

	normoglycaemia and type 2 diabetes, association|Maternal obesity, association with|Obesity|Obesity, association with|Obstructive sleep apnea, association with|Partial lipodystrophy|Periodontitis, association with|Plasma resistin levels, association with|Polycystic ovary syndrome modifier|Reduced serum HDL-C levels, association with|Ulcerative colitis, association with |Insulin resistance in type 2 diabetes, association with|Insulin resistance|Increased plasma leptin levels in obesity, association…	
RP1	***Hypertriglyceridaemia***, association with|Potential protein deficiency|Retinitis pigmentosa|Retinitis pigmentosa, autosomal recessive	high lipids
RP1	***Hypertriglyceridaemia***, association with|Potential protein deficiency|Retinitis pigmentosa|Retinitis pigmentosa, autosomal recessive	high lipids
RP1	***Hypertriglyceridaemia***, association with|Potential protein deficiency|Retinitis pigmentosa|Retinitis pigmentosa, autosomal recessive	hyperlipidemia
SERPINC1	Myocardial infarction, inceased risk, association with|***Deep vein thrombosis*** |Antithrombin deficiency, type 1|Antithrombin deficiency & brachydactyly|Antithrombin deficiency|Altered promoter activity	Maybe: retinal vein occlusion
SERPINC1	Myocardial infarction, inceased risk, association with|Deep vein thrombosis |Antithrombin deficiency, type 1|Antithrombin deficiency & brachydactyly|Antithrombin deficiency|Altered promoter activity	Maybe: retinal vein occlusion

Table 5.8: Number of autosomal dominant genes or dominant/recessive genes with Tier 2 SNV for which there was a match (shown in *bold*) with phenotype in a biobank participant. A total of 1091 variants of this type were noted and this was the subset with any phenotypic overlap or match. The other 834 are not shown

Viewing the genotype-phenotype correlation from the other perspective, we assessed the phenotypes available from chart review that match the genotype from WES analysis. 16 of the 89 participants had none of their phenotypes potentially explained by the filtered WES genotypes. For the remaining 73 participants, 146 (23%) phenotype matches (average 2 per person, range 1-7 matches) were observed from a total of 636 phenotypes. A maximum of 7 phenotypic matches were observed in an individual with 8 phenotypes obtained from chart review. The genotypes contributing to phenotype matches were all Tier-2 SNVs in genes with AD or AD/AR inheritance or GWAS SNP candidates.

5.3.4 ACMG-reportable Gene Coverage Analysis

For the purposes of this study, coverage of 10 reads mapped at a nucleotide position was considered sufficient for high quality variant calling. Utilizing these minimal criteria, we wanted to examine the extent of gene coverage for a typical set of clinically actionable genes. Using the 56 ACMG-reportable [157] gene list, we found on average 9 (range 4-17) of the 56 ACMG genes with poor coverage, having less than 90% of the coding region covered with more than 10 reads in a WES sample (**Figure 5.2**). Coverage of three genes (*PMS2*, *SDHC* and *SDHD*) deteriorated in the WES samples from the second group that utilized the newer Agilent SureSelect V4+UTR capture kit compared to the first group that used Agilent SureSelect 50Mb capture kit (**Figure 5.3**).

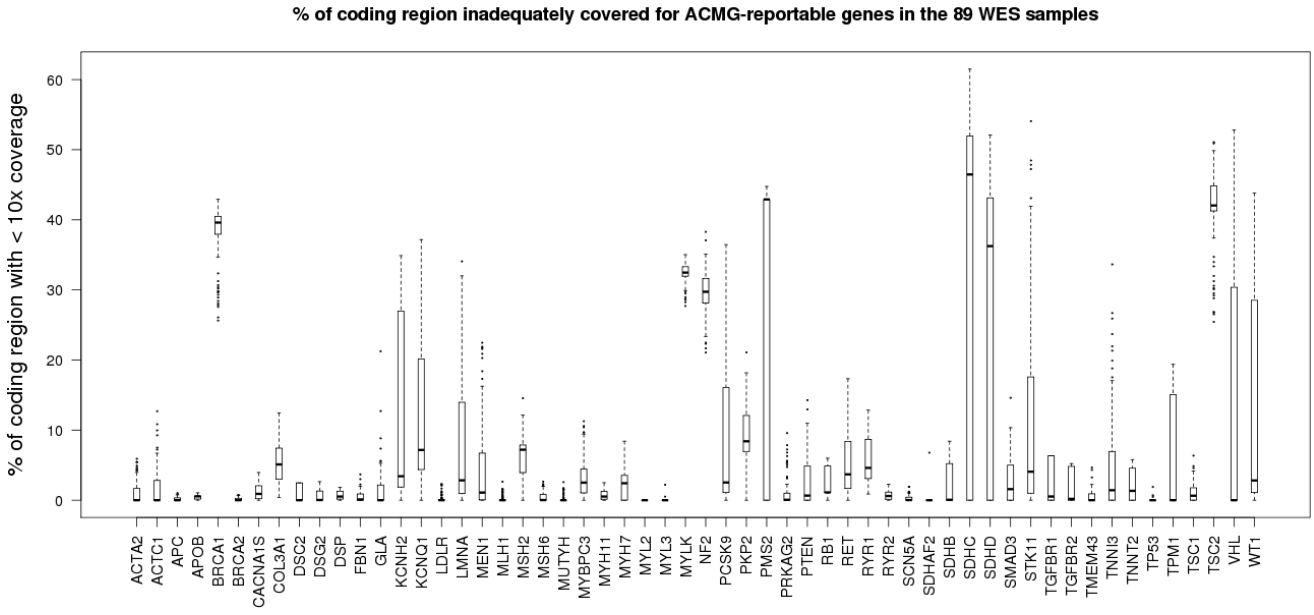

Figure 5.2: Lack of sequencing coverage of the coding region for 56 ACMG-reportable genes in the 89 WES samples. Y-axis represents percentage of coding region with less than 10x (ten-fold) coverage.

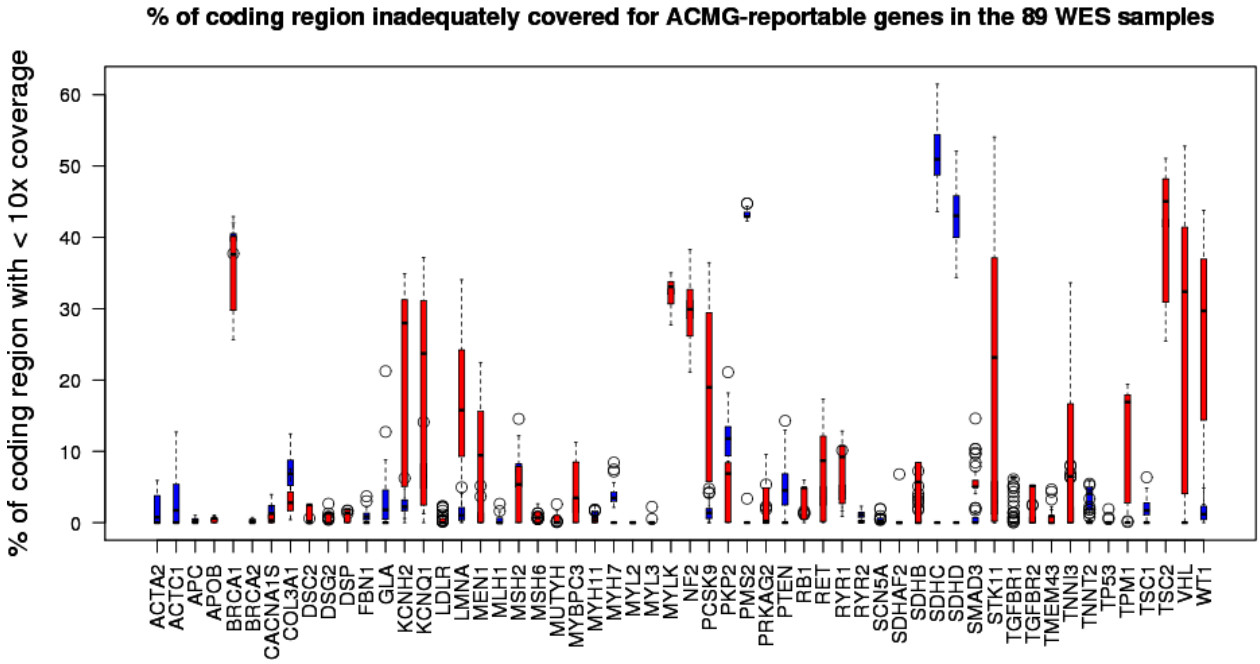

Figure 5.3: Average Gene coding-region coverage by capture-kit (Red is group 1 and blue is group 2) in the 56 ACMG-reportable genes. Y-axis represents percentage of coding region with less than 10x (ten-fold) coverage.

5.3.5 Clinically Significant Variants in ACMG-reportable Gene

We found an average of 1.8 OMIM/HGMD annotated Tier1 or Tier2 filtered variants (range 0-6, median 2) per sample among the list of 56 ACMG-reportable genes. 15 samples had no variants. The 161 variants (13 Tier 1 and 148 Tier 2) found in 74 of biobank samples affect 27 of the 56 ACMG reportable genes. The 8 Tier 1 SNV are from 5 genes (*APOB*, *BRCA2*, *LDLR*, *MYBPC3*, *SMAD3*) and are comprised of 7 stop-gain and 1 splice variant from 7 samples. The 5 Tier 1 INDEL (in 5 samples) are all frame-shift from 4 genes (*BRCA2*, *DSC2*, *PCSK9*, *DSP*). The BRCA2 variants were both found in males, and one was a stop-gain mutation one a frame-shift INDEL. All these variants

were reported as heterozygous and no confirmatory testing was conducted on any of these variants.

5.3.6 Evaluation of variants in Cancer Predisposition Genes

We also examined the frequency of variants among 58 cancer genes, as defined in the methods section. For this group, a total of 115 genetic Tier 1 or Tier 2 variants were found. The distribution of these variants by cancer history and gender is shown in **Figure 5.4** and **Table 5.9**. There were approximately 1.3 (range 0-3) variants per subject with cancer and about 1.1 (range 0-4) variants in subjects without cancer. Even separating by gender, there is not a significant difference in the average number of variants found per person. However, after manual curation of pathogenicity scores using International Agency for Research in Cancer (IARC) guidelines [158, 159], there are 3 variants in the subjects with cancer with a score of 4 (likely pathogenic) compared to none in the subjects without cancer. These 3 variants, all reported as heterozygous, are a missense SNV in *BRIP1* gene and frame-shift mutations in *ATR* and *BRCA2* genes.

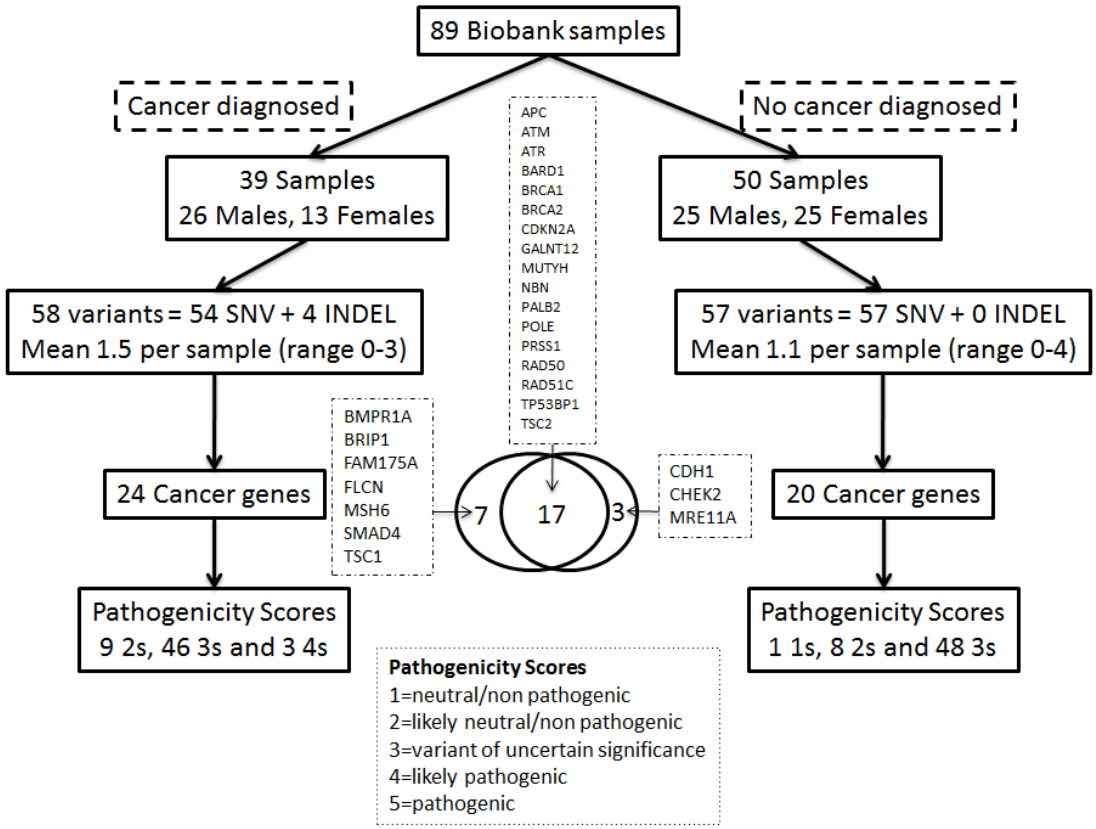

Figure 5.4: Distribution of Tier 1 / Tier 2 variants by cancer history and gender for the cancer predisposition genes is illustrated. The lists of cancer genes exclusively affected in individuals who had cancer in their lifetime, those who did not and the common ones are depicted by the Venn diagram accompanied with actual gene ID. The pathogenicity scores were assigned using International Agency for Research in Cancer (IARC) guidelines which assigns 1 for benign and 5 for pathogenic variants

	89 Biobank Samples			
	Cancer diagnosed		**No cancer diagnosed**	
# of samples	39		50	
Total Variants	58		57	
Mean, Range	1.3, 0-3		1.1, 0-4	
SNV	54		57	
INDEL	4		0	
Cancer Genes	24		20	
	Male	**Female**	**Male**	**Female**

# of samples	26	13	25	25
Total Variants	41	17	31	26
Mean, Range	1.6, 0-3	1.3, 0-3	1.2, 0-4	1, 0-3
SNV	39	15	31	26
INDEL	2	2	0	0
Cancer Genes	20	10	19	13
Lists of Cancer Genes	APC	ATM	APC	APC
	ATM	ATR	ATM	ATM
	BRCA1	BARD1	ATR	ATR
	BRCA2	BMPR1A	BRCA1	BARD1
	BRIP1	BRCA1	BRCA2	BRCA1
	CDKN2A	BRCA2	CDH1	BRCA2
	FAM175A	MUTYH	CDKN2A	GALNT12
	FLCN	PALB2	CHEK2	MRE11A
	GALNT12	TP53BP1	GALNT12	NBN
	MSH6	TSC1	MRE11A	PALB2
	MUTYH		MUTYH	PRSS1
	NBN		NBN	RAD50
	PALB2		PALB2	TSC2
	POLE		POLE	
	PRSS1		PRSS1	
	RAD50		RAD50	
	RAD51C		RAD51C	
	SMAD4		TP53BP1	
	TP53BP1		TSC2	
	TSC2			

Table 5.9: Distribution of 89 WES samples by cancer diagnosis and gender. Also included are metrics on Tier 1 / Tier 2 SNV & INDEL along with the list of cancer predisposition genes found in the groups

5.4 Discussions

This first of its kind study evaluated the WES findings of potentially significant coding region DNA variants in genes of reported significance in 89 individuals who had lived out their entire lifespan and whose medical records were available for correlation. The majority of their medical diagnoses, which were those of the general population, (atherosclerotic cardiovascular disease, Type 2 Diabetes, obesity, degenerative joint disease, cataracts, osteoporosis, etc) were not accounted for by highly penetrant Mendelian gene variants, which is not unexpected. The contribution of WES in providing information that allows people to be proactive for these multi-factorial disorders is likely to be minimal. Of more interest to this study was the degree to which mutations in genes of more Mendelian/single gene disorders did or did not correlate with medical events in these individuals' lifetimes.

We tried to select a set of samples from the Mayo Clinic Biobank to moderately represent the general population. Overall, for the entire set of 89 subjects, there were 51 males and 38 females with an average age at death of 74.47 years (range 28 to 93, median 78 years). In comparison, the average age of death in US is 79 years, while it is 80.9 years in the state of Minnesota and 82.4 years in the Olmsted County [160] where Rochester, MN is located. The lower average age of death in our 89 subjects is primarily due to the selection of cases with younger age at death.

Although different exome enrichment capture kits and different sequencing throughput per sample were used for the two groups of WES samples, the final numbers

of filtered variants were comparable. This is due to the much better capture efficiency of the newer capture kit used for group 2 generating enough sequencing coverage using one-third of a sequencing lane per sample. Overall, our WES dataset of 89 samples generated about 79 (range 56-119) filtered variants per individual (**Table 5.4**). This number is lower than the results from a recent work on clinical interpretation and implication of WGS data that estimates a median of 108 (range 90-127) variants identified for curation per person [63]. These were extracted from the close to 3 million variants identified for each of the 12 WGS samples using the Sequence to Medical Phenotypes identification software developed by their group. Upon manual curation, median of 5 (range 2-6) reportable variants with personal inherited disease risk were identified.

The slightly lower number of variant from WES data compared to the WGS study could be due to differences in sequencing coverage, along with stringency of filtering methods used. The 12 sample WGS dataset [63] was too small to use an internal common variant filtering and remove bioinformatics analyses biases, which was performed for our WES data. This step of filtering variants that were seen in 10% or more of the 89 WES samples removed more than 30% of the called variants in all categories. This highlights the need for a dynamic local repository of variants identified from sequencing to filter the commonly reported ones as likely artifacts of the bioinformatics processing.

In the Tier 1 SNV (stop codon-gain/loss, start codon-loss or splice altering), 4 genes were found to contain variants for which there was a phenotypic match (**Table**

5.6). The *SMAD3* variant is intriguing: it is novel, and while the affected individual did not have a Loeys-Dietz phenotype, he did have a small abdominal aortic aneurysm, degenerative joint disease in his 60s. In addition he had idiopathic pulmonary fibrosis, which has not been associated with *SMAD3* in humans. However, animal studies have suggested a role for fibrosing disorders in mice [161, 162]. Had this variant been discovered during life, it would have been concerning but defining optimal clinical management would have been challenging.

Of the other three variants with overlap between gene and patient phenotype, the *FLG1* mutation (**Table 5.6**) likely contributes significantly to this person's eczema, but this knowledge is not particularly beneficial. Regarding the genes associated with atherosclerosis and glaucoma it is unlikely that these gene variants were the major contributors to these common and complex phenotypes and prior knowledge of these variants would not likely have led to medical interventions. **Table 5.8** looks at the Tier 2 SNV phenotypic (missense predicted damaging) correlations and a longer list of matches shows up. However, the reader will quickly conclude that most of these are multifactorial disorders and the match called out is unlikely to be direct cause and effect. Two exceptions to this may the *RUNX1* mutation in a person with myelodysplastic disorder and the *MC1R* mutation in an individual with 2 melanomas.

Table 5.7 looks at the other side of the coin: the DNA variants with Tier 1 SNV for which no phenotypic match was apparent. Even if some of these genes are not well

established as disease causative or reported to have more associations/low penetrance than typical Mendelian, it is still notable that the list of genes with no evident phenotype is 10 times longer than the list of genes with phenotypic matches. A clinician undertaking testing of an individual would not know which of the Tier 1 SNV might actually be of relevance to this person and which would not.

The majority of these variants would be classified as VUS, not as truly pathogenic using guidelines proposed by the ACMG at the 2013 NSGC meeting, in which in-silico analysis was deemed insufficient alone to distinguish pathogenic from non-pathogenic variants. We observed a stark contrast between the number of variants of different types discovered and the low number of times for which a phenotypic match, even very leniently defined, could be found (**Table 5.5**). In this dataset there were thousands of variants that have been reported before or are novel, in OMIM/HGMD genes, that in-silico analysis defines as likely damaging but the evidence for that effect in the lives of these individuals was absent in the vast majority of instances.

Because the presence or absence of a cancer diagnosis is more straightforward to categorize from a chart review than many sorts of medical disorders (e.g., limited ability to determine if diagnoses like cardiomyopathy or renal failure are primary or secondary on most chart reviews), a deeper genotype-phenotype evaluation of known Mendelian cancer predisposition genes was conducted. There were approximately 1.3 (range 0-3) variants per individual with cancer and about 1.1 (range 0-4) variants in individuals

without cancer. Even separating by gender, there is not a significant difference in the average number of variants found per individual. However, after manual curation of pathogenicity scores, there are 3 variants in the subjects with cancer with a score of 4 (likely pathogenic) compared to none in the individuals without cancer. Presently our ability to determine which DNA variants are pathogenic and which are benign is a major limiting factor in tapping into the clinical utility of WES. This sub-analysis of cancer genes does suggest that a subset of the genetic variants might be contributing to disease, but that most missense variants, which were present in equal numbers in those with and without cancer diagnoses, are not creating apparent risk.

A technical limitation to sensitive genetic variation assessment is the missing coverage of important genes. An average of 9 (17%), with a range 4-17 (7% - 30%) out of the 56 ACMG-reportable genes had sub-optimal coverage per individual for efficient variant calling in our WES data. These numbers are of poorer quality compared to 9-17% ACMG-reportable genes with low coverage identified in the recent WGS study [63]. There are several possibilities for this finding, including variable exome capture efficiency in the library preparation and the usage of multiply-mapped reads during the initial bioinformatics analysis. A sequence read aligning to multiple locations on the genome is commonly mapped randomly to one of those locations, discarded or mapped to a location with a probability based on general read coverage of the region. This is unclear for the WGS study [63] whereas we discarded multiply-mapped reads from our WES data analysis. The large variation in coding region coverage for some genes is due to the

different exome enrichment capture kits used in our two groups. Three genes in particular (PMS2, SDHC and SDHD) had different capture probe design in the Agilent V4+UTR exome kit in order to avoid the homologous pseudo-gene regions, thus yielding much lower overall coverage (**Figure 5.3**).

Notable challenges of this analytic approach include personnel time needed for manual literature review, the subjective nature of bioinformatics filtering thresholds, ambiguity of assigning gene inheritance mode and uncertainty about variant pathogenicity classification. Though not timed, we would agree with recent reports that expert review of each variant to score for pathogenicity could take around an hour per variant [63]. Despite stringent bioinformatics filtering there are a large number of variants, especially missense, requiring classification. Working groups of experts in genomic research, analysis and clinical diagnostic sequencing are collaboratively looking for recommendations and guidelines for investigating genetic variants' causality in human disease [163] and databases of curated variants are needed even more urgently than ever as WES/WGS launches.

Our study is limited to SNV and small INDEL identified from WES data. Compared to WGS, WES is not optimal for detecting Copy Number Variation (CNV) and large structural variants and most of the available tools suffer from limited power to detect CNVs [164]. Our project involved a single medical geneticist expert evaluating the gene inheritance and pathogenicity classification as opposed to a group of experts

engaged in other studies [63, 165]. The EMR at Mayo Clinic may have omitted some important diagnoses as patients may have received care elsewhere and not recorded significant findings on their intake forms. The bioinformatics tools used are not clinically validated and arbitrary quality and read-depth thresholds were used for data filtering. The data we analyzed are from self-reported Caucasian individuals and some of the challenges are further aggravated in studying individuals from other populations. Lastly, our filtering had a heavy reliance on HGMD and OMIM for gene filtering and pathogenic mutation identification.

This study provides new information and begins to quantitate the limited correlation between DNA variants and clinical manifestations on an individual basis, and as such, provides a cautionary note regarding the current predictive value of most DNA variants in the setting of a non-disease selected population. There were many technical challenges that likely affected the results, such as incomplete sequencing coverage, inability to provide clinical interpretation to most DNA variants, diagnoses missing from the EMR, but these alone are unlikely to account for the gap between variants found and absent medical diagnoses. Scores of genes whose functions remain unknown or under-appreciated were excluded from this study. A large number of the variants assigned *in silico* as pathogenic may be neutral. To develop a disorder, multiple genes may need to be involved—single gene disorders may be rather rare in reality. Resolving and understanding these issues will require sustained and large scale collaborative research.

Chapter 6: Conclusions and Discussion

The advent of next generation sequencing has brought about a revolution in biological sciences, stoking novel ideas to pursue research that could potentially transform clinical care. DNA sequencing is by far the most popular of the NGS applications and includes Whole Genome Sequencing, Whole Exome Sequencing and sequencing of custom gene panels or targeted regions. However, the direct output from NGS instruments is uninformative and requires extensive infrastructure, analytics and interpretation. This dissertation highlights the major challenges associated with efficiently utilizing DNA sequencing data and provides precise bioinformatics solutions.

There are four major aspects to this dissertation. We developed TREAT [1], an integrated workflow infrastructure to automate the basic analysis of NGS DNA data (**Chapter 2**). The workflow incorporates basic quality checks, alignment, variant calling, visualization and reporting. TREAT is a modular, efficient and streamlined workflow that scales well, is used by multiple institutes including Mayo Clinic, UIUC, Appistry© and is capable of evolving with the changing needs of NGS bioinformatics. In terms of the depth of coverage from NGS data, I implemented a module for comprehensive evaluation of a list of genes directly implicated in the disease or condition being studied (**Chapter 3**). The result is a gene level report on nucleotide regions not adequately covered by sequenced reads, implying scant data for efficient variant calling [2]. The turnaround time for analysis of a single genome is currently about 75 hours, prohibitively long for clinical

adaptability of NGS. At the same time, not the entire genome's evaluation is clinically actionable, and thus a vast majority of WGS variants are of low pertinence. As part of the thesis, we developed an iterative approach of WGS bioinformatics focusing on clinically relevant genomic regions and report results in less than 5 hours [3] (**Chapter 4**). The final aspect of this dissertation is exploring the interpretation and phenotypic implication of NGS DNA variants (**Chapter 5**). We sequenced 89 individuals who had comprehensive medical records and clinical diagnosis over their lifetime. Genes affected by SNV and INDEL identified for these 89 individuals were carefully curated for the potential function and the results were correlated to the individual's clinical diagnoses. This study helped us understand the challenges of phenotypic annotation of genotypes from NGS data and uncertainty surrounding clinical reporting from NGS interpreted results.

With the goal of making clinical applicability of NGS a reality, this thesis went into deeper evaluation as various hurdles were encountered. This meant building an integrated infrastructure for basic streamlined analysis of NGS data, ensuring clinical applicability with quick turn-around of prioritized results and interpretation of NGS results for clinicians, geneticists or the individual end users. As the field evolves, there are a lot of exciting avenues for future exploration. With decreasing costs and increasing availability of sequencing data, large-scale studies involving hundreds and even thousands of NGS samples have been initiated. The bioinformatics workflows are evolving to utilize information from multiple samples [166], perform local sequence

assembly around INDEL and complex variants [167] and use tumor and normal information in order to detect somatic variant calls [168-173]. Furthermore, additional applications of DNA NGS data are identification of genomic events like Copy Number and Structural Variants [174-182]. ENCODE [29] and other whole genome atlas projects have uncovered functional evidence for vast regions of the genome, way beyond the comfortable 1-2% coding exon regions. The grappling with NGS data and associated applications will determine the successes and unraveling of biological and disease mechanisms in the next decade.

At least 10% of the genes known to be linked to disease have been reported as not adequately sequenced by NGS [63]. Deciding what the results from NGS data analysis mean for the patient remains an active area of research and continues to improve. Lastly, in order to better understand the complex mechanisms of human disease, additional types of genomic data along with proteomic and metabolomic data would need to be integrated. Bioinformatics advancements would remain essential for further and more ubiquitous clinical utilization of next generation DNA sequencing data.

Bibliography

1. Asmann YW, Middha S, Hossain A, Baheti S, Li Y, Chai H-S, Sun Z, Duffy PH, Hadad AA, Nair A *et al*: **TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data**. *Bioinformatics* 2012, **28**(2):277-278.
2. Klein CJ, Middha S, Duan X, Wu Y, Litchy WJ, Gu W, Dyck PJB, Gavrilova RH, Smith DI, Kocher J-P *et al*: **Application of whole exome sequencing in undiagnosed inherited polyneuropathies**. *Journal of Neurology, Neurosurgery & Psychiatry* 2014.
3. Middha S, Baheti S, Hart SN, Kocher J-PA: **From Days to Hours: Reporting Clinically Actionable Variants from Whole Genome Sequencing**. *PloS one* 2014, **9**(2):e86803.
4. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor**. *Bioinformatics* 2010, **26**(16):2069-2070.
5. Kahvejian A, Quackenbush J, Thompson JF: **What would you do if you could sequence everything?** *Nat Biotech* 2008, **26**(10):1125-1133.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The Sequence of the Human Genome**. *Science* 2001, **291**(5507):1304-1351.
7. Human Genome Sequencing C: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.
8. **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
9. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage [phi]X174 DNA**. *Nature* 1977, **265**(5596):687-695.
10. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals**. *Nat Rev Genet* 2004, **5**(5):335-344.
11. Bentley DR: **Whole-genome re-sequencing**. *Current Opinion in Genetics & Development* 2006, **16**(6):545-552.
12. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotech* 2008, **26**(10):1135-1145.
13. Mitra RD, Church GM: **In situ localized amplification and contact replication of many individual DNA molecules**. *Nucleic Acids Research* 1999, **27**(24):e34-e39.
14. Gresham D, Dunham MJ, Botstein D: **Comparing whole genomes using DNA microarrays**. *Nat Rev Genet* 2008, **9**(4):291-302.

15. Soni GV, Meller A: **Progress toward Ultrafast DNA Sequencing Using Solid-State Nanopores**. *Clinical Chemistry* 2007, **53**(11):1996-2001.
16. Healy K: **Nanopore-based single-molecule DNA analysis**. *Nanomedicine* 2007, **2**(4):459-481.
17. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome**. *Science* 2005, **309**(5741):1728-1732.
18. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT *et al*: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**(7189):872-876.
19. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M *et al*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome**. *Nature* 2008, **456**(7218):66-72.
20. Ahn S, Kim T, Lee S, Kim D, Ghang H, Kim D, Kim B, Kim S, Kim W, Kim C *et al*: **The first Korean genome sequence and analysis: full genome sequencing for a socioethnic group**. *Genome Res* 2009.
21. Kim J-I, Ju YS, Park H, Kim S, Lee S, Yi J-H, Mudge J, Miller NA, Hong D, Bell CJ *et al*: **A highly annotated whole-genome sequence of a Korean individual**. *Nature* 2009, **460**(7258):1011-1015.
22. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J *et al*: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**(7218):60-65.
23. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N: **What can exome sequencing do for you?** *Journal of Medical Genetics* 2011.
24. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De la Vega FM, Donnelly P, Egholm M *et al*: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
25. Theis JL, Zimmermann MT, Larsen BT, Rybakova IN, Long PA, Evans JM, Middha S, De Andrade M, Moss RL, Wieben ED *et al*: **TNNI3K mutation in familial syndrome of conduction system disease, atrial tachyarrhythmia and dilated cardiomyopathy**. *Human Molecular Genetics* 2014.
26. **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353):609-615.
27. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T *et al*: **Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants**. *Nat Genet* 2010, **42**(11):969-972.
28. **Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA** [http://evs.gs.washington.edu/EVS/]
29. The EPC: **A User's Guide to the Encyclopedia of DNA Elements (ENCODE)**. *PLoS Biol* 2011, **9**(4):e1001046.

30. Robertson JA: **The $1000 Genome: Ethical and Legal Issues in Whole Genome Sequencing of Individuals**. *The American Journal of Bioethics* 2003, **3**(3):35-42.
31. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics**. *Journal of Genetics and Genomics* 2011, **38**(3):95-109.
32. Rizzo JM, Buck MJ: **Key Principles and Clinical Applications of ‚ÄúNext-Generation‚Äù DNA Sequencing**. *Cancer Prevention Research* 2012, **5**(7):887-900.
33. Kuhlenbäumer G, Hullmann J, Appenzeller S: **Novel genomic techniques open new avenues in the analysis of monogenic disorders**. *Human Mutation* 2011, **32**(2):144-151.
34. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC *et al*: **Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome**. *Nat Genet* 2010, **42**(9):790-793.
35. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA *et al*: **Exome sequencing identifies the cause of a mendelian disorder**. *Nat Genet* 2010, **42**(1):30-35.
36. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG *et al*: **Exome Sequencing Identifies WDR35 Variants Involved in Sensenbrenner Syndrome**. *The American Journal of Human Genetics* 2010, **87**(3):418-423.
37. Hoischen A, van Bon BWM, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G *et al*: **De novo mutations of SETBP1 cause Schinzel-Giedion syndrome**. *Nat Genet* 2010, **42**(6):483-485.
38. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkalo*f*ülu Aô, √ñzen S, Sanjad S *et al*: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing**. *Proceedings of the National Academy of Sciences* 2009, **106**(45):19096-19101.
39. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I: **Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders**. *J Hum Genet* 2012, **57**(10):621-632.
40. Ku C-S, Naidoo N, Pawitan Y: **Revisiting Mendelian disorders through exome sequencing**. *Hum Genet* 2011, **129**(4):351-370.
41. Lupski J, Gonzaga-Jauregui C, Yang Y, Bainbridge M, Jhangiani S, Buhay C, Kovar C, Wang M, Hawes A, Reid J *et al*: **Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy**. *Genome Medicine* 2013, **5**(6):57.
42. Teer JK, Mullikin JC: **Exome sequencing: the sweet spot before whole genomes**. *Human Molecular Genetics* 2010, **19**(R2):R145-R151.
43. Nishiguchi KM, Tearle RG, Liu YP, Oh EC, Miyake N, Benaglio P, Harper S, Koskiniemi-Kuendig H, Venturini G, Sharon D *et al*: **Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural**

changes and NEK2 as a new disease gene. *Proceedings of the National Academy of Sciences* 2013.
44. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y-Q: **Evaluation of next-generation sequencing software in mapping and assembly**. *J Hum Genet* 2011, **56**(6):406-414.
45. Scholz MB, Lo C-C, Chain PSG: **Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis**. *Current Opinion in Biotechnology* 2012, **23**(1):9-15.
46. Fonseca NA, Rung J, Brazma A, Marioni JC: **Tools for mapping high-throughput sequencing data**. *Bioinformatics* 2012, **28**(24):3169-3177.
47. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly**. *Nat Meth* 2009, **6**(11s):S6-S12.
48. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nat Rev Genet* 2011, **12**(6):443-451.
49. Mardis E: **The $1,000 genome, the $100,000 analysis?** *Genome Medicine* 2010, **2**(11):84.
50. McPherson JD: **Next-generation gap**. *Nat Meth* 2009, **6**(11s):S2-S5.
51. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP *et al*: **SVA: software for annotating and visualizing sequenced human genomes**. *Bioinformatics* 2011, **27**(14):1998-2000.
52. Grant JR, Arantes AS, Liao X, Stothard P: **In-depth annotation of SNPs arising from resequencing projects using NGS-SNP**. *Bioinformatics* 2011, **27**(16):2300-2301.
53. Nix D, Di Sera T, Dalley B, Milash B, Cundick R, Quinn K, Courdy S: **Next generation tools for genomic data generation, distribution, and visualization**. *BMC Bioinformatics* 2010, **11**(1):1-12.
54. Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S: **GAMES identifies and annotates mutations in next-generation sequencing projects**. *Bioinformatics* 2011, **27**(1):9-13.
55. Shetty A, Athri P, Mondal K, Horner V, Steinberg K, Patel V, Caspary T, Cutler D, Zwick M: **SeqAnt: A web service to rapidly identify and annotate DNA sequence variations**. *BMC Bioinformatics* 2010, **11**(1):471.
56. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Research* 2010, **38**(16):e164.
57. Sifrim A, Van Houdt J, Tranchevent L-C, Nowakowska B, Sakai R, Pavlopoulos G, Devriendt K, Vermeesch J, Moreau Y, Aerts J: **Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease**. *Genome Medicine* 2012, **4**(9):73.
58. Cheng Y-C, Hsiao F-C, Yeh E-C, Lin W-J, Tang C-YL, Tseng H-C, Wu H-T, Liu C-K, Chen C-C, Chen Y-T *et al*: **VarioWatch: providing large-scale and comprehensive annotations on human genomic variants in the next generation sequencing era**. *Nucleic Acids Research* 2012, **40**(W1):W76-W81.

59. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S: **AnnTools: a comprehensive and versatile annotation toolkit for genomic variants**. *Bioinformatics* 2012, **28**(5):724-725.
60. Le SQ, Durbin R: **SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples**. *Genome Research* 2010.
61. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses**. *Nat Rev Genet* 2014, **15**(2):121-132.
62. Wang W, Wei Z, Lam T-W, Wang J: **Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions**. *Sci Rep* 2011, **1**.
63. Dewey FE, Grove ME, Pan C, et al.: **CLinical interpretation and implications of whole-genome sequencing**. *JAMA* 2014, **311**(10):1035-1045.
64. Ajay SS, Parker SCJ, Ozel Abaan H, Fuentes Fajardo KV, Margulies EH: **Accurate and comprehensive sequencing of personal genomes**. *Genome Research* 2011, **21**(9):1498-1505.
65. Metzker ML: **Sequencing technologies [mdash] the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
66. Puckelwartz MJ, Pesce LL, Nelakuditi V, Dellefave-Castillo L, Golbus JR, Day SM, Cappola TP, Dorn GW, Foster IT, McNally EM: **Supercomputing for the parallelization of whole genome analysis**. *Bioinformatics* 2014, **30**(11):1508-1513.
67. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Kallberg M, Kumar SA, Liao A *et al*: **Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms**. *Bioinformatics* 2013.
68. Tarczy-Hornoch P, Amendola L, Aronson SJ, Garraway L, Gray S, Grundmeier RW, Hindorff LA, Jarvik G, Karavite D, Lebo M *et al*: **A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record**. *Genet Med* 2013, **15**(10):824-832.
69. Domchek S, Weber BL: **Genetic Variants of Uncertain Significance: Flies in the Ointment**. *Journal of Clinical Oncology* 2008, **26**(1):16-17.
70. Altshuler DM, Gibbs RA, Consortium TIH: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**(7311):52-58.
71. Berg JS, Khoury MJ, Evans JP: **Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time**. *Genet Med* 2011, **13**(6):499-504.
72. Nekrutenko A, Taylor J: **Next-generation sequencing data interpretation: enhancing reproducibility and accessibility**. *Nat Rev Genet* 2012, **13**(9):667-672.
73. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant**

analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* 2014, **15**(2):256-278.
74. Garcia Castro A, Thoraval S, Garcia L, Ragan M: **Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator**. *BMC Bioinformatics* 2005, **6**(1):87.
75. Schuster SC: **Next-generation sequencing transforms today's biology**. *Nature methods* 2008, **5**(1):16-18.
76. Ansorge WJ: **Next-generation DNA sequencing techniques**. *New biotechnology* 2009, **25**(4):195-203.
77. Metzker ML: **Sequencing technologies - the next generation**. *Nature reviews*, **11**(1):31-46.
78. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA *et al*: **Exome sequencing identifies the cause of a mendelian disorder**. *Nature genetics*, **42**(1):30-35.
79. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE *et al*: **Targeted capture and massively parallel sequencing of 12 human exomes**. *Nature* 2009, **461**(7261):272-276.
80. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA *et al*: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy**. *The New England journal of medicine*, **362**(13):1181-1191.
81. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG *et al*: **Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome**. *American journal of human genetics*, **87**(3):418-423.
82. Bonnefond A, Durand E, Sand O, De Graeve F, Gallina S, Busiah K, Lobbens S, Simon A, Bellanne-Chantelot C, Letourneau L *et al*: **Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome**. *PloS one*, **5**(10):e13630.
83. Harbour JW, Onken MD, Roberson ED, Duan S, Cao L, Worley LA, Council ML, Matatall KA, Helms C, Bowcock AM: **Frequent Mutation of BAP1 in Metastasizing Uveal Melanomas**. *Science (New York, NY*.
84. Richter BG, Sexton DP: **Managing and Analyzing Next-Generation Sequence Data**. *PLoS Comput Biol* 2009, **5**(6):e1000369.
85. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research*, **38**(16):e164.
86. Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S: **GAMES identifies and annotates mutations in next-generation sequencing projects**. *Bioinformatics (Oxford, England)*.
87. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, Caspary T, Cutler DJ, Zwick ME: **SeqAnt: a web service to rapidly identify and annotate DNA sequence variations**. *BMC bioinformatics*, **11**:471.

88. Nix DA, Di Sera TL, Dalley BK, Milash BA, Cundick RM, Quinn KS, Courdy SJ: **Next generation tools for genomic data generation, distribution, and visualization**. *BMC bioinformatics*, **11**:455.
89. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
90. McPherson JD: **Next-generation gap**. *Nature methods* 2009, **6**(11 Suppl):S2-5.
91. **FASTQC: A quality control tool for high throughput sequence data** [http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc]
92. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics (Oxford, England)* 2009, **25**(16):2078-2079.
93. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, **43**(5):491-+.
94. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.
95. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M *et al*: **SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors**. *Bioinformatics (Oxford, England)*, **26**(6):730-736.
96. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nature protocols* 2009, **4**(7):1073-1081.
97. **SeattleSeq annotation tool** [http://gvs.gs.washington.edu/SeattleSeqAnnotation/HelpAbout.jsp]
98. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research* 2001, **29**(1):308-311.
99. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acids Research* 2003, **31**(13):3812-3814.
100. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Research* 2014, **42**(D1):D199-D205.
101. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000, **28**(1):27-30.
102. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**(1):207-210.
103. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive**

for functional genomics data sets,Äîupdate. *Nucleic Acids Research* 2013, **41**(D1):D991-D995.
104. Consortium TIG: **Expression Project for Oncology (expO)**.
105. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Res* 2003, **31**(4):e15.
106. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat Biotech* 2011, **29**(1):24-26.
107. Thorvaldsd√≥ttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Briefings in Bioinformatics* 2013, **14**(2):178-192.
108. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Research* 2004, **32**(suppl 1):D493-D496.
109. Team RDC: **R: A Language and Environment for Statistical Computing**. 2012.
110. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Meth* 2010, **7**(4):248-249.
111. Goya R, Sun MGF, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M *et al*: **SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors**. *Bioinformatics* 2010, **26**(6):730-736.
112. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE *et al*: **Targeted capture and massively parallel sequencing of 12 human exomes**. *Nature* 2009, **461**(7261):272-276.
113. Anastasio N, Ben-Omran T, Teebi A, Ha KCH, Lalonde E, Ali R, Almureikhi M, Der Kaloustian VM, Liu J, Rosenblatt DS *et al*: **Mutations in SCARF2 Are Responsible for Van Den Ende-Gupta Syndrome**. *The American Journal of Human Genetics*, **87**(4):553-559.
114. Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**(3):R25.
115. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**(5):713-714.
116. Gregg EW, Sorlie P, Paulose-Ram R, Gu Q, Eberhardt MS, Wolz M, Burt V, Curtin L, Engelgau M, Geiss L: **Prevalence of lower-extremity disease in the US adult population >=40 years of age with and without diabetes: 1999-2000 national health and nutrition examination survey**. *Diabetes Care* 2004, **27**(7):1591-1597.
117. Callaghan B, McCammon R, Kerber K, Xu X, Langa KM, Feldman E: **Tests and expenditures in the initial evaluation of peripheral neuropathy**. *Archives of internal medicine* 2012, **172**(2):127-132.
118. Chia L, Fernandez A, Lacroix C, Adams D, Plante V, Said G: **Contribution of nerve biopsy findings to the diagnosis fo disabling neuropathy in the elderly: A retrospective revie of 100 consecutive patients**. *Brain* 1996, **119**:1091-1098.

119. Myers MI, Peltier AC: **Uses of Skin Biopsy for Sensory and Autonomic Nerve Assessment**. *Curr Neurol Neurosci Rep* 2013, **13**:323.
120. Smith AG: **Diagnosis of neuropathy: comment on "tests and expenditures in the initial evaluation of peripheral neuropathy"**. *Archives of internal medicine* 2012, **172**(2):132-133.
121. Dyck PJ, Oviatt KF, Lambert EH: **Intensive evaluation of referred unclassified neuropathies yields improved diagnosis**. *Ann Neurol* 1981, **10**(3):222-226.
122. **Online Mendelian Inheritance in Man, OMIM®**. 2011.
123. Saporta AS, Sottile SL, Miller LJ, Feely SM, Siskind CE, Shy ME: **Charcot-Marie-Tooth disease subtypes and genetic testing strategies**. *Ann Neurol* 2011, **69**(1):22-33.
124. Murphy SaM, Davidson GL, Brandner S, Houlden H, Reilly MM: **Mutation in FAM134B causing severe hereditary sensory neuropathy**. *Journal of Neurology, Neurosurgery & Psychiatry* 2012, **83**(1):119-120.
125. Amato AA, Reilly MM: **The death panel for Charcot-Marie-Tooth panels**. *Annals of neurology* 2011, **69**(1):1-4.
126. Latour P, Vial C: **[Molecular diagnosis of axonal forms of Charcot-Marie-Tooth disease]**. *Rev Neurol (Paris)* 2009, **165**(12):1122-1126.
127. Klein CJ, Duan X, Shy ME: **The Inherited Neuropathies: Clinical Overview and Update**. *Muscle & Nerve* 2013, **(In Press)**.
128. Foo JN, Liu JJ, Tan EK: **Whole-genome and whole-exome sequencing in neurological diseases**. *Nature reviews Neurology* 2012, **8**(9):508-517.
129. Klein CJ, Botuyan M-V, Wu Y, Ward CJ, Nicholson GA, Hammans S, Hojo K, Yamanishi H, Karpf AR, Wallace DC: **Mutations in DNMT1 cause hereditary sensory neuropathy with dementia and hearing loss**. *Nat Genet* 2011, **43**(6):595-600.
130. Landour√© G, Sullivan JM, Johnson JO, Munns CH, Shi Y, Diallo O, Gibbs JR, Gaudet R, Ludlow CL, Fischbeck KH *et al*: **Exome sequencing identifies a novel TRPV4 mutation in a CMT2C family**. *Neurology* 2012, **79**(2):192-194.
131. Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, Hulme W, Siskind C, Vance J, Shy M *et al*: **Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family**. *Annals of neurology* 2011, **69**(3):464-470.
132. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA *et al*: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy**. *The New England journal of medicine* 2010, **362**(13):1181-1191.
133. Foo J-N, Liu J-J, Tan E-K: **Whole-genome and whole-exome sequencing in neurological diseases**. *Nat Rev Neurol* 2012, **8**(9):508-517.
134. Ivanova N, Claeys KG, Deconinck T, et al.: **Hereditary spastic paraplegia 3a associated with axonal neuropathy**. *Archives of Neurology* 2007, **64**(5):706-713.
135. Guelly C, Zhu P-P, Leonardis L, Papi*f*á L, Zidar J, Schabh√ºttl M, Strohmaier H, Weis J, Strom TM, Baets J *et al*: **Targeted High-Throughput Sequencing**

Identifies Mutations in atlastin-1 as a Cause of Hereditary Sensory Neuropathy Type I. *The American Journal of Human Genetics* 2011, **88**(1):99-105.
136. Klein CJ, Duan X, Shy ME: **Inherited neuropathies: Clinical overview and update**. *Muscle & Nerve* 2013, **48**(4):604-622.
137. **GeneReviews® [Internet]** [http://www.ncbi.nlm.nih.gov/books/NBK1116/]
138. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN: **Human Gene Mutation Database (HGMD®): 2003 update**. *Human Mutation* 2003, **21**(6):577-581.
139. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA *et al*: **Identification of genetic variants using bar-coded multiplexed sequencing**. *Nat Meth* 2008, **5**(10):887-893.
140. **Novoalign** [www.novocraft.com]
141. Farrell CM, O,ÄôLeary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, Diekhans M, Barrell D, Searle SMJ, Aken B *et al*: **Current status and new features of the Consensus Coding Sequence database**. *Nucleic Acids Research* 2014, **42**(D1):D865-D872.
142. Kohane IS: **The incidentalome: A threat to genomic medicine (vol 296, pg 212, 2006)**. *Jama-J Am Med Assoc* 2006, **296**(12):1466-1466.
143. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.
144. Solomon BD, Nguyen A-D, Bear KA, Wolfsberg TG: **Clinical Genomic Database**. *Proceedings of the National Academy of Sciences* 2013, **110**(24):9851-9855.
145. Church GM: **The Personal Genome Project**. *Molecular Systems Biology* 2005, **1**(1).
146. Biesecker LG: **Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project**. *Genet Med* 2012, **14**(4):393-398.
147. Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF *et al*: **The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine**. *Genome Res* 2009, **19**(9):1665-1674.
148. Fabsitz RR, McGuire A, Sharp RR, Puggal M, Beskow LM, Biesecker LG, Bookman E, Burke W, Burchard EG, Church G *et al*: **Ethical and Practical Guidelines for Reporting Genetic Research Results to Study Participants: Updated Guidelines From a National Heart, Lung, and Blood Institute Working Group**. *Circulation: Cardiovascular Genetics* 2010, **3**(6):574-580.
149. Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD: **Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility**. *Genome Res* 2012, **22**(3):421-428.
150. Olson JE, Ryu E, Johnson KJ, Koenig BA, Maschke KJ, Morrisette JA, Liebow M, Takahashi PY, Fredericksen ZS, Sharma RG *et al*: **The Mayo Clinic**

Biobank: A Building Block for Individualized Medicine. *Mayo Clinic proceedings Mayo Clinic* 2013, **88**(9):952-962.
151. **Online Mendelian Inheritance in Man** [http://omim.org/]
152. **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422):56-65.
153. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.
154. Fu W, O/'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA *et al*: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants**. *Nature* 2013, **493**(7431):216-220.
155. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O/'Daniel JM, Ormond KE *et al*: **ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing**. *Genet Med* 2013, **15**(7):565-574.
156. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
157. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E: **ACMG clinical laboratory standards for next-generation sequencing**. *Genet Med* 2013, **15**(9):733-747.
158. Tavtigian SV, Greenblatt MS, Goldgar DE, Boffetta P: **Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group**. *Human Mutation* 2008, **29**(11):1261-1264.
159. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FBL, Hoogerbrugge N, Spurdle AB, Tavtigian SV: **Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results**. *Human Mutation* 2008, **29**(11):1282-1291.
160. **Olmsted County Public Health Service records** [http://www.co.olmsted.mn.us/OCPHS/reports/Documents/chna3.pdf]
161. Gauldie J, Kolb M, Ask K, Martin G, Bonniaud P, Warburton D: **Smad3 Signaling Involved in Pulmonary Fibrosis and Emphysema**. *Proceedings of the American Thoracic Society* 2006, **3**(8):696-702.
162. Warburton D, Shi W, Xu B: **TGF-β-Smad3 signaling in emphysema and pulmonary fibrosis: an epigenetic aberration of normal development?** In., vol. 304; 2013: L83-L85.
163. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA *et al*: **Guidelines for investigating causality of sequence variants in human disease**. *Nature* 2014, **508**(7497):469-476.
164. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, Jiang Q, Allen AS, Zhu M: **An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data**. *Hum Mutation* 2014.
165. Dorschner Michael O, Amendola Laura M, Turner Emily H, Robertson Peggy D, Shirts Brian H, Gallego Carlos J, Bennett Robin L, Jones Kelly L, Tokita Mari J,

Bennett James T *et al*: **Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes**. *The American Journal of Human Genetics* 2013, **93**(4):631-640.
166. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z: **Variant Callers for Next-Generation Sequencing Data: A Comparison Study**. *PLoS One* 2013, **8**(9):e75619.
167. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS: **ABRA: improved coding indel detection via assembly based re-alignment**. *Bioinformatics* 2014.
168. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A *et al*: **JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data**. *Bioinformatics* 2012, **28**(7):907-913.
169. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L: **SomaticSniper: identification of somatic point mutations in whole genome sequencing data**. *Bioinformatics* 2012, **28**(3):311-317.
170. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK: **Strelka: accurate somatic small-variant calling from sequenced tumor,Äìnormal sample pairs**. *Bioinformatics* 2012, **28**(14):1811-1817.
171. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome Research* 2012, **22**(3):568-576.
172. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. *Nat Biotech* 2013, **31**(3):213-219.
173. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y *et al*: **An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data**. *Nucleic Acids Research* 2013, **41**(7):e89.
174. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L *et al*: **Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome**. *Science* 2007, **318**(5849):420-426.
175. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: **High-resolution mapping of copy-number alterations with massively parallel sequencing**. *Nat Meth* 2009, **6**(1):99-103.
176. Xie C, Tammi M: **CNV-seq, a new method to detect copy number variation using high-throughput sequencing**. *BMC Bioinformatics* 2009, **10**(1):80.
177. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M: **Detecting copy number variation with mated short reads**. *Genome Research* 2010, **20**(11):1613-1622.

178. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavar√© S: **CNAseg,Äîa novel framework for identification of copy number changes in cancer from second-generation sequencing data**. *Bioinformatics* 2010, **26**(24):3051-3058.
179. Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E: **Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization**. *Bioinformatics* 2011, **27**(2):268-269.
180. Miller CA, Hampton O, Coarfa C, Milosavljevic A: **ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads**. *PLoS One* 2011, **6**(1):e16327.
181. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing**. *Genome Research* 2011, **21**(6):974-984.
182. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP *et al*: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation**. *Nat Meth* 2009, **6**(9):677-681.

Appendix A.

Permissions

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jun 04, 2014

This is a License Agreement between Sumit Middha ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3402190032569
License date	Jun 04, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data:
Licensed content author	Yan W. Asmann, Sumit Middha, Asif Hossain, Saurabh Baheti, Ying Li, High-Seng Chai, Zhifu Sun, Patrick H. Duffy, Ahmed A. Hadad, Asha Nair, Xiaoyu Liu, Yuji Zhang, Eric W. Klee, Krishna R. Kalari, Jean-Pierre A. Kocher
Licensed content date	01/15/2012
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Publisher of your work	n/a
Expected publication date	Aug 2014
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

Figure A.1: Permission to reproduce Bioinformatics journal manuscript for TREAT

BMJ PUBLISHING GROUP LTD. LICENSE TERMS AND CONDITIONS

Jun 04, 2014

This is a License Agreement between Sumit Middha ("You") and BMJ Publishing Group Ltd. ("BMJ Publishing Group Ltd.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by BMJ Publishing Group Ltd., and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3395380287816
License date	May 24, 2014
Licensed content publisher	BMJ Publishing Group Ltd.
Licensed content publication	Journal of Neurology, Neurosurgery and Psychiatry
Licensed content title	Application of whole exome sequencing in undiagnosed inherited polyneuropathies
Licensed content author	Christopher J Klein, Sumit Middha, Xiaohui Duan, Yanhong Wu, William J Litchy, Weihong Gu, P James B Dyck, Ralitza H Gavrilova, David I Smith, Jean-Pierre Kocher, Peter J Dyck
Licensed content date	Mar 6, 2014
Type of Use	Thesis/Dissertation
Requestor type	Author of this article
Format	Print and electronic
Portion	Figure/table/extract
Number of figure/table/extracts	5
Will you be translating?	No
Circulation/distribution	100000
Title of your thesis / dissertation	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Expected completion date	Aug 2014
Estimated size(pages)	100
BMJ VAT number	674738491
Billing Type	Invoice
Billing address	200 1st St SW ROCHESTER, MN 55901 United States
Permissions Cost	0.00 USD
VAT (0.00%)	0.00 USD
Total	0.00 USD
Terms and Conditions	

Figure A.2: Permission to reproduce Journal of Neurology, Neurosurgery and Psychiatry manuscript

OPEN ACCESS Freely available online

PLOS ONE

From Days to Hours: Reporting Clinically Actionable Variants from Whole Genome Sequencing

Sumit Middha, Saurabh Baheti, Steven N. Hart, Jean-Pierre A. Kocher*

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America

Abstract

As the cost of whole genome sequencing (WGS) decreases, clinical laboratories will be looking at broadly adopting this technology to screen for variants of clinical significance. To fully leverage this technology in a clinical setting, results need to be reported quickly, as the turnaround rate could potentially impact patient care. The latest sequencers can sequence a whole human genome in about 24 hours. However, depending on the computing infrastructure available, the processing of data can take several days, with the majority of computing time devoted to aligning reads to genomics regions that are to date not clinically interpretable. In an attempt to accelerate the reporting of clinically actionable variants, we have investigated the utility of a multi-step alignment algorithm focused on aligning reads and calling variants in genomic regions of clinical relevance prior to processing the remaining reads on the whole genome. This iterative workflow significantly accelerates the reporting of clinically actionable variants with no loss of accuracy when compared to genotypes obtained with the OMNI SNP platform or to variants detected with a standard workflow that combines Novoalign and GATK.

Citation: Middha S, Baheti S, Hart SN, Kocher J-PA (2014) From Days to Hours: Reporting Clinically Actionable Variants from Whole Genome Sequencing. PLoS ONE 9(2): e86803. doi:10.1371/journal.pone.0086803

Editor: Charles Y. Chiu, University of California, San Francisco, United States of America

Received September 4, 2013; **Accepted** December 13, 2013; **Published** February 5, 2014

Funding: The funding was provided by the Center for Individualized Medecine at Mayo Clinic. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Figure A.3: Permission to reproduce PLOS ONE manuscript

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Jun 08, 2014

This is a License Agreement between Sumit Middha ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3402191219508
License date	Jun 04, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Biotechnology
Licensed content title	Next-generation DNA sequencing
Licensed content author	Jay Shendure and Hanlee Ji
Licensed content date	Oct 9, 2008
Volume number	26
Issue number	10
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Expected completion date	Aug 2014
Estimated size (number of pages)	100
Total	0.00 USD
Terms and Conditions	

Terms and Conditions for Permissions

Figure A.4: Permission to reproduce Figure 1 [10] for illustration

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 08, 2014

This is a License Agreement between Sumit Middha ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Sumit Middha
Customer address	200 1st St SW ROCHESTER, MN 55901
License number	3402200280888
License date	Jun 04, 2014
Licensed content publisher	Elsevier
Licensed content publication	Current Opinion in Biotechnology
Licensed content title	Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis
Licensed content author	Matthew B. Scholz,Chien-Chi Lo,Patrick SG Chain
Licensed content date	February 2012
Licensed content volume number	23
Licensed content issue number	1
Number of pages	7
Start Page	9
End Page	15
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	electronic

Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Expected completion date	Aug 2014
Estimated size (number of pages)	100
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD
Terms and Conditions	

Figure A.5: Permission to reproduce Table 1.1 [45] for illustration

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Aug 25, 2014

This is a License Agreement between Sumit Middha ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3456090877091
License date	Aug 25, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	Fast and accurate short read alignment with Burrows–Wheeler transform:
Licensed content author	Heng Li, Richard Durbin
Licensed content date	07/15/2009
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Publisher of your work	n/a
Expected publication date	Aug 2014
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

Figure A.6: Permission to reproduce Tables 2.1 and 2.2 [89] for illustration

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Aug 25, 2014

This is a License Agreement between Sumit Middha ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3456081145491
License date	Aug 25, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors:
Licensed content author	Rodrigo Goya, Mark G.F. Sun, Ryan D. Morin, Gillian Leung, Gavin Ha, Kimberley C. Wiegand, Janine Senz, Anamaria Crisan, Marco A. Marra, Martin Hirst, David Huntsman, Kevin P. Murphy, Sam Aparicio, Sohrab P. Shah
Licensed content date	03/15/2010
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	BIOINFORMATICS SOLUTION FOR CLINICAL UTILIZATION OF NEXT GENERATION DNA SEQUENCING
Publisher of your work	n/a
Expected publication date	Aug 2014
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

Figure A.7: Permission to reproduce Figure 2.2 [95] for illustration