APPLIED GENOMICS:

DEVELOPMENT OF BIOINFORMATICS PIPELINES FOR ANALYZING CLINICAL PEDIATRIC GENOMIC DATA

by

Erin L. Crowgey

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

Winter 2016

APPLIED GENOMICS:

DEVELOPMENT OF BIOINFORMATICS PIPELINES FOR ANALYZING CLINICAL PEDIATRIC GENOMIC DATA

by

Erin L. Crowgey

Approved: _____
Cathy H. Wu, PhD
Edward G. Jefferson Professor of Bioinformatics & Computational Biology

Approved: _____
Babatunde A. Ogunnaike, PhD
Dean of the College of Engineering

Approved: _____
Ann L. Ardis, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Cathy H. Wu, Ph.D
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Anders Kolb, M.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Carl Schmidt, Ph.D
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Huey-Jen Lee Lin, Ph.D
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Shawn Polson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Chuming Chen, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my primary advisor, Dr. Cathy H. Wu, for her unwavering support and encouragement during this process. Dr. Wu's enthusiasm and excitement for bioinformatics, combined with her leadership skills, made her an outstanding advisor for this project. She has inspired me to push my intellectual boundaries, while maintaining balance in my personal life. I cannot express what a tremendous impact that Cathy has had on my career development. Thank you Cathy, for your wisdom and support!

Second, I would like to express my deepest gratitude to Dr. Anders Kolb from Nemours Alfred I DuPont Hospital for Children. Without his support, this project would not have been possible. Dr. Kolb is an outstanding physician, mentor, and scientist. I appreciate his time and valuable insight on this project. Dr. Kolb went above and beyond for this project, and has helped me realize my long term career goals. Thank you Andy, I have tremendous respect for you as a mentor and a friend.

I would like to also extend my gratitude to my committee members who have provided excellent guidance along the way. Dr. Carl Schmidt, who was also my professor for the bioinformatics course and preliminary exam, is truly an amazing human to discuss science with. Dr. Huey-Jen Lee Lin for her excellent ideas, encouragement, and positive attitude during this process. I would like to express thank you to Dr. Shawn Polson and Dr. Chuming Chen for their helpful guidance on all things related to bioinformatics and scripting. I would like to also thank Dr. Katia Sol-Church from Nemours Children Hospital for giving me the opportunity to work on

a rare Mendelian disease dataset. She provided essential insight for the Mendelian disease analysis, and was instrumental in presenting and publishing this work.

During my time as a graduate student I formed several new friendships. Thank you to Dr. Jennifer Wyffels for the encouragement along this long process and for helping me to keep positive. I also really appreciate all of my friends outside of graduate school for reminding me about the important things in life. I would also like to thank all of the professors and past co-workers for providing me with the skills and knowledge required for this type of project. Special thank you to Dr. Jennifer Van Eyk, Ian Wright, Dr. David Fancy, and Dr. Ross Chambers, all of whom had a significant impact on my scientific career development.

Lastly, I would like to thank my family. During graduate school I decided to make the life-long commitment to my husband, Michael Kaczmarski. He has supported me unconditionally through this process, and I'm eternally blessed with his love. Thank you to my mother and father, Virginia Crowgey, and Fred Crowgey, for their life-time support and encouragement, without which none of this would be possible. I would also like to thank my sister, Adrienne Husty, for always having my back in life no matter what.

Funding for this dissertation project was provided by the Leukemia Research Foundation of Delaware, and support from the University of Delaware Center for Bioinformatics and Computational Biology Core (use of biohen compute cluster) was made possible through funding from Delaware INBRE (NIGMS GM103446), Delaware EPSCoR (NSF EPS-0814251, NSF IIA-1330446), the State of Delaware, and the Delaware Biotechnology Institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The onset and prognosis of several human diseases, such as cancer, are characterized by specific genomic alterations. The sequencing and assembly of the human genome is enabling advancements in personalized medicine, but the process of associating genetic mutations to a specific human disease and treatment is still complex. Recent advancements in DNA sequencing technologies, known as next generation sequencing (NGS), are enabling the detection of many genomic alterations at once. However, a primary limiting factor to clinical applications of genomic NGS is downstream bioinformatics analysis.

A novel approach was created for analyzing whole exome sequencing (WES) datasets (sequenced on the Illumina platform) from clinical patients diagnosed with a rare Mendelian disease. One of the datasets used to help establish the methodologies was paired-end WES from six patients, plus their family members, with a rare disorder characterized by facio-skeletal abnormalities. Robust bioinformatics pipelines were implemented for trimming, genome alignment, single nucleotide polymorphisms (SNPs) detection and annotation, and copy number variation detection. Quality control metrics were analyzed at each step of the pipeline to ensure data integrity for clinical applications. The variants are annotated with three custom modules that enable flexible filtering of the variants based upon criteria such as quality of variant, inheritance pattern (e.g. dominant, recessive, X-linked), and minor allele frequency.

Bioinformatics methodologies were also developed for analyzing NGS data generated from 19 pediatric acute myeloid leukemia patients. The bioinformatics

pipelines developed were focused on single nucleotide variant detection and annotation, combined with genomic insertion / deletion detection and annotation. A list of verified single nucleotide variants was provided with the clinical NGS dataset, and the pipeline was capable of detecting ~94% of the verified single nucleotide variants using a combination of Mutect and Shimmer.

The bioinformatics pipeline developed reported high quality single nucleotide variants that were previously not reported to the Children's Oncology Group. Furthermore, detection and analysis of an internal tandem duplication (ITD) in FLT3, a known clinically relevant mutation in pediatric AML associated with poor prognosis, was conducted using Pindel. The ITD was detected in 5 of the 6 patient's NGS data, which had previously only been detected using PCR and electrophoresis. Collectively, this dissertation project provides a unique method for prioritizing and visualizing genomic variants using functional annotations, including gene ontologies and pathway enrichment strategies.

Chapter 1

INTRODUCTION

The Precision Medicine Initiative announcement in early 2015 (Collins & Varmus, 2015) has brought to light the need to increase funding and progress in the medical field towards delivering timely and efficient individual patient care. Precision medicine is not a new concept, although there are various definitions depending on the organization defining the idea. In practice precision medicine is a paradigm that utilizes molecular analysis techniques in the tailored treatment of an individual patient.

For some diseases, personalized medicine plays an important role in appropriately classifying the pathophysiology and potential treatment strategies that have maximum efficacy and minimal side effects. An early example from 1956, is the genetic basis for the selective toxicity of fava beans and the antimalarial drug primaquine, both of which were discovered to be a deficiency in the metabolic enzyme glucose-6-phosphate dehydrogenase (Alving, Carson, Flanagan, & Ickes, 1956; Mager et al., 1965). Certainly leveraging genetic information related to adverse toxicity to a drug should be within the realm of health care today.

There are many challenges and opportunities for developing an infrastructure capable of truly delivering a personalized treatment based on real time molecular biology data. These challenges vary in scope and magnitude, ranging from electronic consent, specimen collection and management, to data generation and analysis. As technologies are quickly advancing, so is the magnitude of data associated with any single patient. Nowadays, the magnitude of data is not only related to the sheer

number of patients, or sample size, but in the vast amount of information that is being generated for each patient / sample throughout time. This type of big data is complex and traditional data processing applications are not adequate.

In 2001, Doug Laney defined data growth challenges and opportunities as being three-dimensional, increasing in volume, velocity, and variety (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf). Over a decade later, big data is now described by 5Vs: volume, velocity, variety, veracity, and value (http://www.martinhilbert.net/big-data-for-development/). The 2 additional Vs highlight important considerations around the trustworthiness (veracity) of the data, and the ability to actually translate the data into something meaningful (value). It is increasingly obvious that more data does not translate into more knowledge, as big data constraints are certainly playing a major role in developing systems capable of realizing personalized medicine (Fan, Han, & Liu, 2014).

New developments in molecular biology techniques, which are often used to gain insight into genetic disorders, are generating massive amounts of data that need processed and refined prior to being incorporated into Electronic Health Records (EHR) or clinical decision support applications (Fan et al., 2014). Although these new techniques, such as next generation sequencing, are gaining momentum in the clinical field, there are no gold standards for data analysis, interpretation, and integration.

Ultimately, there is a unique niche in the informatics field, especially in bioinformatics, to develop and deliver robust methodologies capable of analyzing the massive amount of data being generated by new technologies. Bioinformatics, an interdisciplinary field, focuses on developing methods/ algorithms and software for

unraveling complex biological data. It is a field that requires the intersection of computer science, statistics, mathematics, biology, and engineering, with the ultimate goal of being applied in the clinical field for translating large biological datasets into knowledge, clearly requiring a team science approach.

1.1 Overview Genetic Disorders

Deoxyribonucleic acid (DNA) is the hereditary molecule passed down between generations. The complete set of an individual's DNA, often referred to as the genome, is comprised of loci that encode for functional RNA molecules, like messenger RNA (mRNA), that are ultimately translated into a protein. Interestingly, more than 98% of the human genome does not encode for mRNA molecules that are translated into protein sequences, (Elgar & Vavouri, 2008), but rather for noncoding functional RNAs (ribosomal RNA, etc.), *cis*- and *trans*- regulatory elements, pseudogenes, repeat sequences, transposons and viral elements, and telomeres (Ludwig, 2002).

Genetic disorders can be inherited or caused by *de novo* mutations, and are characterized by epigenetic patterns or genetic alteration(s) within a specific region of the genome. The genetic abnormality can be small, such as single nucleotide polymorphisms (SNP) that characterize cystic fibrosis, or large rearrangements, such as the chromosomal abnormality associated with Down syndrome (trisomy 21). Most genetic disorders have been associated with genetic alterations within coding regions, or regions that affect the coding of proteins, and are often divided into three types of categories: 1. Single-gene disorder 2. Chromosomal disorders 3. Complex disorders (https://www.nlm.nih.gov/medlineplus/geneticdisorders.html). There are many

challenges associated with characterizing a genetic disorder, creating a unique niche for developing appropriate bioinformatics methodologies.

Biology is complex and understanding diversity in signaling pathways caused by genomic alterations is a challenging task. There is tremendous redundancy in protein function, and many proteins are associated with several cellular processes and location. For example, FMS-like tyrosine kinase 3 receptor (*FLT3*), when activated by ligand, undergoes dimerization followed by tyrosine kinase activation, which leads to transphosphorylation of the receptor cytoplasmic domain and recruitment of SH2-domain containing adaptor proteins. These adaptor proteins activate several downstream signal transduction pathways such as Ras, Raf, mitogen-activated protein kinase (MAPK), and phosphatidylinositide 3-kinase (PI3K), and synergize with other growth factor signaling events, which leads to proliferation and differentiation of hematopoietic stem cells (Rosen et al., 2010; Gilliland & Griffin, 2002).

Recently, there has been a major paradigm shift in the scientific community to broaden the approach of studying genetic disorders using a single variant approach, like traditional genome wide association studies (GWAS), to more sophisticated techniques that look at the potential of having many genes dysregulated (Dudbridge, 2013). This can be difficult as genomic alterations may vary between patients, because misregulation of similar pathways can result in the same overall phenotype. For example, pediatric acute myeloid leukemia (AML) is a heterogeneous disease characterized by different underlying genetic alterations in numerous genes. These genomic alterations can be large, i.e. large translocation events, or small single nucleotide variants (SNVs), but the overall clinical diagnostic is still AML (Gamis, Alonzo, Perentesis, & Meshinchi, 2013).

Another major challenge, and often over-looked, is the ability to unravel the interactions between genetics, environment, and lifestyle. Expression of genes within an organism can be influenced by the environment, at both the internal (within the body) and external level (location). Gender is one type of internal environment that can affect gene expression (Ingrid Lobo, 2008), whereas drugs, chemicals, and temperature are types of external factors that can affect gene expression.

Often times it is impossible to collect all of the appropriate environmental factors that may be present within a disease cohort, making it nearly impossible to incorporate this type of analysis with genetic data, unless the study and implementation has been robustly designed. And ultimately, well controlled twin studies are required to de-complex the interplay between environmental factors and genetics, like the famous National Aeronautics and Space Administration (NASA) twin experiment (https://www.nasa.gov/twins-study), but often times it is difficult to find such pedigrees.

Additionally, it is difficult to assess or measure genetic susceptibility, which is further complicated by ethical and legal issues. Predisposition does not necessarily predict that a patient will develop a specific disease, and understanding how lifestyle impacts disease onset is essential. Ultimately, there are numerous ethical and social ramifications of having a patient's genome analyzed, which should not be taken lightly.

1.1.1 Rare Mendelian Diseases

Gregor Mendel, a nineteenth-century Austrian monk, derived the basic principles of the laws of inheritance by studying phenotypes in garden peas that were passed down from generation to generation. There are 3 basic laws to Mendelian

genetics: the Law of Segregation, the Law of Independent Assortment, and the Law of Dominance. Segregation is a process that takes place during gamete formation that separates the alleles for each gene so that each gamete carriers only 1 allele for each gene. Genes for different traits can segregate independently during gamete formation, which is the second Law. Finally, the Law of Dominance pertains to dominant and recessive alleles, with the dominate allele determining the observed phenotype/effect in a given organism.

A study that examined >5,700 consecutive admissions into Rainbows Babies and Children's Hospital (Cleveland, OH) found an underlying genetic component for ~70% of the admitted children (McCandless, Brunger, & Cassidy, 2004), supporting that genetics plays an important role in pediatric disorders. The human genome project (HPG) has characterized over 2,500 Mendelian disorders with genetic mutation data, providing a rich resource to clinical investigators (Abecasis et al., 2010; Manwar Hussain, Khan, & Ali Mohamoud, 2014) and further highlights the interplay between genetics and disease progression. The Online Mendelian Inheritance in Man (OMIM), a catalog of human Mendelian disorders, contains 20,267 entries describing 13,606 genes from ~7,000 disorders (Manwar Hussain et al., 2014), supporting there is diversity in genes associated with human disease.

Mendelian diseases can be a single-gene disorder characterized by a single allele inherited from one parent or both. The allele can cause a single phenotypic effect or result in many phenotypic effects, termed pleiotropy. However, over the years researchers and clinicians are discovering Mendelian disorders are often associated with several genetic loci inherited from one parent or both.

For example, cystic fibrosis was considered to be a single-gene disorder, but recent data suggests that the severity and onset of the disease is related to many genetic alterations (Cutting, 2014). Williams-Beuren syndrome (MIM#194050) and DiGeorge/velocardiofacial (MIM# 188400) were also considered single gene mutations until recently, and are now linked to gene deletion or duplication events known as copy number variations (CNVs) (Manwar Hussain et al., 2014). However, these types of CNVs can occur spontaneously and are not necessarily inherited in every case.

Over the last decade new techniques in DNA sequencing have significantly impacted the study of rare Mendelian diseases (Gilissen, Hoischen, Brunner, & Veltman, 2011), by focusing, in a non-hypothesis driven approach, on genetic alterations that cause changes in protein-coding regions. These types of studies are enabled because of advancements in DNA sequencing. In 1977, Fred Sanger published a pivotal paper describing a primer-extension method for sequencing DNA molecules (Sanger, Nicklen, & Coulson, 1977). For decades, this underlying method has been the gold standard for DNA sequencing, and is often referred to as Sanger sequencing. However, during the 1990s and early 2000s new DNA sequencing technologies, termed next generation sequencing (NGS), have been developed. These techniques harness the ability to sequence many DNA molecules at once, a stark difference between the traditional Sanger sequencing method.

The transition to sequencing many DNA molecules at once has created a major burden on downstream analysis approaches, as the traditional methods for analyzing sequencing data are no longer appropriate. Although, NGS is particularly suited for clinical discovery as it collects a comprehensive set of molecular alterations. The

largest barrier to realization of NGS potential in personalized medicine is the time and computing intensive downstream bioinformatics analysis. Analyzing NGS data, whether with publicly available algorithms or commercially available software packages, requires an appropriate computational infrastructure.

There are commercially available packages that provide a suite of tools for analyzing NGS data. However, these applications can be expensive, difficult to integrate with other algorithms, and often times do not contain all of the desired algorithms for analyzing the NGS data from start to finish. The lack of flexibility is a trade-off for ease of use, but many analyses require a certain degree of optimization. Nonetheless, commercially available packages do provide a solution for the novice scientist, and applications like Ingenuity Variant Analysis and CLC Genomics Workbench have certainly benefitted the scientific community and have helped to increase our understanding of genetic alterations associated with disease phenotypes.

Publicly available algorithms for analyzing NGS data tend to focus on a single aspect of data analysis, and do not provide a comprehensive workflow from start to finish requiring the construction of a bioinformatics pipeline. While the use of public tools is free, a computer scientist with advanced training is still typically required as the data files are large and a graphical interface is not available for the majority of algorithms. Furthermore, there are no established gold standards for translating NGS into clinical knowledge, as different diseases may require different strategies for analyzing genomic data.

Fortunately, there are many groups working on standardization and best practices. For example, the Broad Institute, a collaborative group of scientists, have developed a Genome Analysis Toolkit (GATK) (DePristo et al., 2011), and have

released best practice guidelines for analyzing NGS data, such as marking duplicate reads and local realignment around known variants. There are many publicly available algorithms for supporting the key steps in NGS analysis, such as genome alignment (bwa) (H. Li, 2013) and variant calling (Haplotyper). Furthermore, there are high quality curated databases, like ensembl and UniProt KB that are publicly available for downstream analysis and integration strategies.

The genomics community has done a great job of creating standard file formats that help with interoperability between tools / algorithms and data exchange. For example, the standard file format for the alignment of NGS reads to a genome is a sam / bam file, and is agnostic to sequencing platform. A great deal of effort by the community has also resulted in a standard file format for variant data, called a variant call file (VCF), which works well for small genetic alterations and aids in high-throughput data exchange.

Illumina is one type of NGS platform that has been widely used in human projects. Illumina has several types of instruments, and provides kits and reagents required for upstream library preparation. The Illumina platform can be used to sequence DNA or RNA (cDNA). A cost-effective approach for using NGS sequencing in the clinical environment for detection of genetic variants associated with Mendelian diseases is to first perform exon-capture followed by NGS (Wang, Liu, Yang, & Gelernter, 2013). Previous studies have demonstrated that 15X average coverage is sufficient for homozygous SNP detection, and 33X for heterozygous, using Illumina short-read technology (Sims, Sudbery, Ilott, Heger, & Ponting, 2014; Bentley et al., 2008).

Exome capture is a technique for enriching regions of the genome that are responsible for coding proteins, which represents about 1-2% of the human genome (Wang et al., 2013). When using a whole exome sequencing (WES) approach coverage, both depth and breadth, of protein coding regions should be appropriate to allow accurate detection of genomic alterations within those regions. For example, targeted NGS was recently used to analyze a region of the genome connected with Noonan Syndrome and was able to validate variants previously associated with the syndrome, while also detecting new variants of interest (Lepri et al., 2014).

For this dissertation, an approach was created for analyzing whole exome sequencing datasets (sequenced on the Illumina platform) from clinical patients diagnosed with a rare Mendelian disease. The genomic variants are annotated with three custom modules that enable flexible filtering of the variants based upon criteria such as quality of variant, inheritance pattern (e.g. dominant, recessive, X-linked), and minor allele frequency. Specifically, the goal for this dissertation was to create a flexible bioinformatics pipeline for analyzing clinical NGS samples generated from rare Mendelian diseases that can integrate genomic variant data with genomic, proteomic, pathway, and patient specific inheritance data.

One of the WES datasets used to help establish the bioinformatics methodologies was generated from six patients, plus their family members, with a rare disorder characterized by facio-skeletal abnormalities. Robust bioinformatics pipelines were implemented for trimming, genome alignment, single nucleotide polymorphisms (SNPs) detection, and copy number variation detection. Quality control metrics were analyzed at each step of the pipeline to ensure data integrity for clinical applications.

1.1.2 Pediatric Cancers: Acute Myeloid Leukemia

Approximately 10,380 children in the United States under the age of 15 will be diagnosed with cancer in 2015 (American Cancer Society). The most prevalent cancers in children are different from those seen in adults (American Cancer Society), with the most common type of cancer in children being leukemia or otherwise known as cancers of the bone and blood marrow. Leukemia accounts for about 30% of all pediatric cancer cases (American Cancer Society), and currently there is no known cure. Other cancers dominant in children are brain / central nervous system tumors, neuroblastoma, sarcomas and Wilms tumor.

The most common types of leukemia cancers in children are acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML). AML is a complex disease characterized by dysregulation of signal transduction pathways in hematopoietic progenitors that ultimately results in the increase of proliferation and survival of non-complete differentiated cells (S. Meshinchi, 2003). AML is considered a disease of the genome as many genomic alterations are required for disease onset and progression. Genomic variants for AML are often described as either Type I mutations, which alter cell proliferation, or Type II mutations, which alter cell differentiation pathways (*Cancer Genomics*, 2014).

Pediatric AML is a heterogeneous disease and can be divided into several sub-classifications. The first morphologic-histochemical classification system for AML was developed by the French-American-British (FAB) working group and consisted of 10 major classifications. However, there are major differences, in both the diagnostic criteria and disease management, between pediatric and adult AML.

Pediatric AML is a rare disease with only ~500 children a year diagnosed (cancer.net) and prognosis has improved over the decades with overall survival rates

of 50% to 60% (Zwaan et al., 2003). The highest rate of incidence in pediatric AML is 1.6 per 100,000 in the first year of life, and decreases with age until about the forth decade of life (Schuback, Arceci, & Meshinchi, 2013). However, relapse is a major concern in pediatric AML and accounts for more than half of the deaths in pediatric leukemia cases (S. Meshinchi, 2003).

Mutations associated with AML are found in several genes that regulate hematopoiesis including *FLT3*, *NPM1*, *CEBPA*, *RAS*, *c-KIT*, and *WT1*. However, the majority of genomic alterations do not have a prognostic value and a specific target or a distinct pathway has not been identified for therapeutic intervention (Schuback et al., 2013). Currently, subsets of genomic alterations associated with pediatric AML are under investigation for predicting response and outcome to treatment. For example co-occurring mutations, such as an internal tandem duplication (ITD) in the *FLT3* gene accompanied by *NUSP98/NSD1* are associated with disease severity and treatment response (Ostronoff et al., 2014; Longo, Döhner, Weisdorf, & Bloomfield, 2015).

Cytogenetic markers have played a major role in the diagnosis and classification of AML. The majority of pediatric AML cases can be divided into distinct cytogenetic categories; 25% have either a translocation of t(8;21) or an inversion Inv(16); 12% have a distinct translocation t(15; 17); 20% have rearrangements involving the *MLL* gene; and 20% do not have a clear karyotype abnormality (Schuback et al., 2013). Small genetic alterations, including medium sized insertions and deletions, are not typically detected by traditional cytogenetic studies and their association with AML is an area of current research that requires other types of molecular biology techniques.

Recent advancements in DNA sequencing technology have aided in our ability to detect numerous genomic alterations, ranging from a single variant to large chromosomal translocations, from a single genome. These advancements have helped clinicians gain a better understanding of genetic alterations associated with pediatric AML, as *a prior* knowledge is not required. However, different bioinformatics methodologies are required because there is not a single algorithm that can detect all of the different types of genomic variants simultaneously. Therefore, developing a pipeline for pediatric AML is more complex compared to developing a pipeline for adult AML, which is largely characterized by single nucleotide variants (SNVs).

The goals for this dissertation were to translate large genomic data sets generated from pediatric AML, into biological knowledge by integrating SNVs with structural variant information. A bioinformatics pipeline was developed for WES, TES, and WGS from pediatric AML patients. The pipeline combines publicly available algorithms and custom scripts to detect and prioritize genomic variants associated with pediatric AML. Specifically, the pipeline was developed using Illumina sequencing data from the Children's Oncology Group available through the database of Genotypes and Phenotypes (dbGap).

A novel computational module was created for prioritizing somatic variants, by integrating different types of genomic alterations and their functional annotations, including looking at the difference in read proportions for an allele between the cancer and remission state. Bioinformatics approaches developed for the pediatric AML project are in direct response to the needs of clinical collaborators, and will address current limitations of the processed data previously provided to them.

1.2 Application of Work

Precision medicine is driven by advancements in technology and relies on the ability to convert big data into knowledge. With improvements in DNA sequencing techniques over the last decade, the concept of utilizing a person's genome for diagnostic and treatment purposes is becoming a reality. Recently, the FDA-cleared the MiSeqDx Cystic fibrosis clinical sequencing assay from Illumina, which is an vitro diagnostic (IVD) next-generation sequencing test for the CFTR gene (http://www.illumina.com/products/miseqdx-cystic-fibrosis-clinical-sequencing-assay.html).

This dissertation project seeks to address the current gaps in using NGS in the clinical field, with the ultimate goal of creating a diagnostic kit. A team science approach is necessary for this type of project, as expertize in several fields is required. It is essential to establish a robust collaboration in order to bridge the gap between clinical knowledge and molecular biology data. Additionally, an extensive computational infrastructure is required for data processing, along with key personnel that have the skill sets to maintain and develop the computational environment.

The work with the pediatric AML NGS data analysis will be continued beyond completion of this dissertation with clinical collaborators and primary advisors. One of the future applications of this project is to transition to using error correct (EC) sequencing generated from the Illumina platform (Young et al., 2015). Cancer samples are heterogeneous and it is a complex process to unravel allelic data, and low-level underlying mutations that eventual under-go clonal expansion into a cancerous state are of interest. Minimal residual disease (MRD), small number of leukemic cells that persist after treatment, in pediatric AML is a major concern, and the ability to detect these variants at low level in the remission state will be essential. Detecting

variants associated with treatment options, like FLT3/ITD, and drug resistance (*TET*2) is also of importance and will hopefully be enabled by future application of this work.

The long term goal is to further develop the EC sequencing application to include structural variants (SVs), and currently preliminary data is being generated for this project. Collectively, the concepts used in this project can apply to other diseases associated with SNVs and SVs. A major lesson learned from this project is that not all variants are equally detectable, and optimization of bioinformatics methodologies is a critical step in data analysis.

For rare Mendelian disease, a flexible pipeline was created for analyzing WES datasets (sequenced on the Illumina platform) from clinical patients diagnosed with a rare Mendelian disease. The application of the workflow was broaden to other rare Mendelian diseases, and the fundamental concepts were applied to other datasets which consisted of microarray and genome wide association (GWAS) data (manuscript in preparation).

The incorporation of genetic data into the decision making processes is complex. With the ability to analyze large genomic datasets, comes the task to prioritize the data into meaningful knowledge. Currently, there are no commercial pipelines that systematically integrates EHR with genetic or genomic data (Kannry & Williams, 2013). The National Human Genome Research Institute sponsored Electronic Medical Records and Genomics (eMERGE) Network is a federally funded consortium that publishes guidance for integrating genomic data into EHR and will be utilized in future studies.

One of the key considerations defined for using genomic information in the clinic is the ability to store structured genetic information, as the data needs to be read

by a computer in order to be used in clinical decision-support engines (Kannry & Williams, 2013). Numerous steps are required to process NGS into a state that is able to be structured in a meaningful way. Ultimately, pipelines and methodologies, like the ones developed in this dissertation project, will be an essential component for integrating genomics data into EHR.

Chapter 2

EXPERIMENTAL PROCEDURES AND QUALITY CONTROL ANALYSIS

2.1 Library Quality Analysis

All of the NGS datasets were generated prior to this analysis. The platform used for sequencing, for both the Rare Mendelian Disease and pediatric AML, was Illumina using standard insert sizes (~300 bp) and read lengths (2x100bp). There are no exact guidelines for applying WES in the analysis of human diseases in regards to per base sequence content / quality and per read sequence content / quality. Therefore, one of the goals for this dissertation was to understand compositional ranges of human WES libraries, with the goal of being able to flag libraries outside of appropriate ranges.

The Rare Mendelian disease data was obtained post trimming, whereas the pediatric AML dataset required processing of the NGS reads. Cutadapt (http://journal.embnet.org/index.php/embnetjournal/article/view/200/479) was implemented to trim (default parameters) the libraries based on common Illumina primers/adaptors and base quality scores on 3` end. On average the WES library size for the pediatric AML project was 88 million reads prior to trimming, and 85 million after trimming (Figure 2.1 panel A). The average number of base pairs (bps) prior to trimming was 17 billion, and after trimming was 15 billion bps (Figure 2.1 panel B). The total number of reads and base pairs analyzed in the downstream steps are ~5 billion and ~900 billion, respectively.

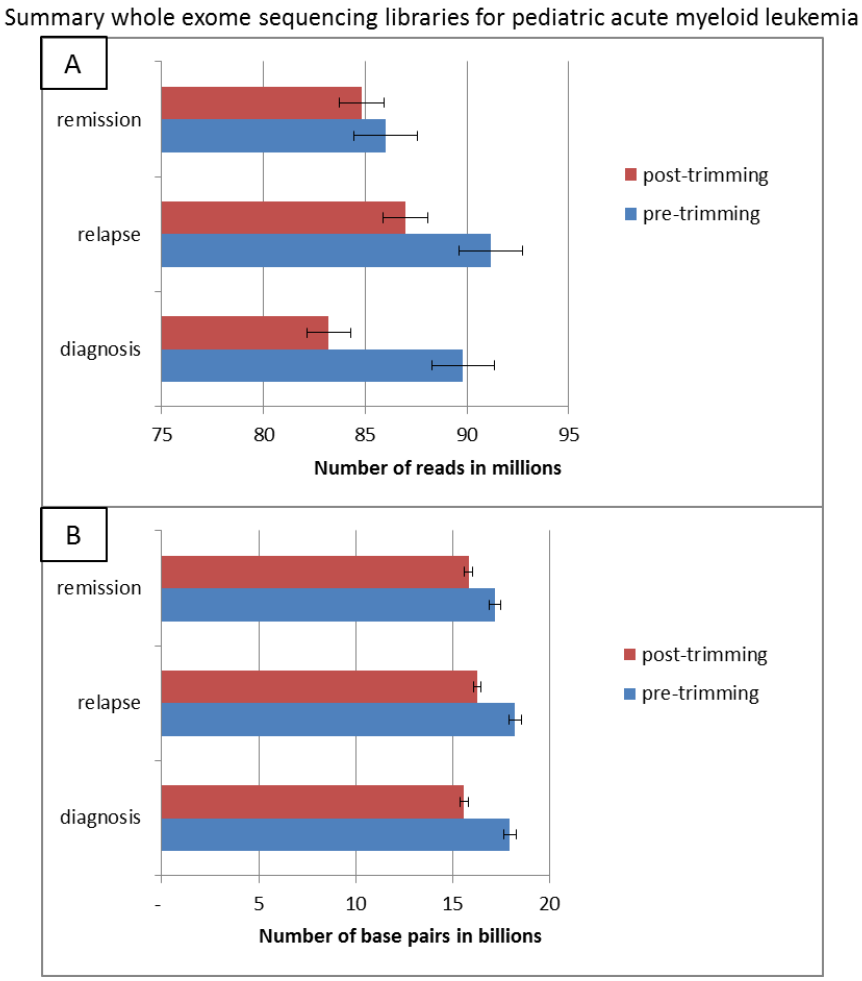

Figure 2.1. Summary pediatric acute myeloid leukemia whole exome sequencing libraries. Panel A: Average library size in reads prior to trimming (blue) and post trimming (red). Panel B: Average library size base pairs prior to trimming (blue) and post trimming (red).

The average library size for the rare Mendelian disease was 42 million reads. These libraries were received post trimming, and did not require any further processing at this stage for downstream analysis. Table 2.1 summarizes the number of

reads per each whole exome sequencing library for each individual analyzed in the rare Mendelian disease cohort.

Table 2.1. Summary of whole exome sequencing libraries for the rare Mendelian disease cohort

Family	Pedigree	Number of reads
Family 1	mother	52,872,147
	father	36,826,844
	proband	32,326,640
Family 2	proband	38,152,736
Family 3	proband	42,007,611
	mother	46,306,409
	father	43,073,956
	sibling	49,996,203
	half-sibling	42,713,741
Family 4	proband	41,062,797
	mother	36,846,007
Family 5	proband	36,528,981
	sibling	42,641,664
	mother	51,745,035

For the pediatric AML project, each patient had 3 library files: diagnosis, relapse, and remission (Table 2.2).

Table 2.2 Summary of whole exome sequencing libraries for pediatric acute myeloid leukemia

	Number of Reads		
Sample	Diagnosis	Relapse	Remission
1	82,135,479	87,921,387	82,999,123
2	87,864,267	91,147,349	94,877,166
3	88,129,462	92,469,296	83,203,827
4	93,641,247	84,158,585	91,720,751
5	92,393,296	95,951,606	96,111,734
6	89,580,286	83,815,457	91,722,272
7	41,789,418	45,576,513	85,826,923
8	43,864,766	63,815,907	101,882,273
9	83,810,691	102,359,394	90,531,798
10	96,678,375	96,122,527	83,909,272
11	83,019,468	71,835,671	79,884,291
12	85,299,361	89,112,722	82,193,333
13	90,298,552	101,174,790	76,692,428
14	71,117,271	76,696,115	82,472,367
15	99,868,442	103,242,370	72,050,714
16	85,789,382	97,059,746	83,235,923
17	91,686,513	99,140,096	99,592,862
18	87,332,231	85,999,543	67,723,824
19	97,260,933	102,133,993	90,300,444

The US Centers for Disease Control and Prevention (CDC) assembled a national working group, termed Next-generation sequencing: Standardization of Clinical Testing or Nex-StoCT, to lead an initiative for defining platform-independent guidelines for using NGS in the clinical field (Gargis et al., 2012). The Supplementary Guidelines published by Nex-StoCT highlight key quality metrics that should be considered when establishing and validating a clinical NGS workflow. All of the metrics published as being necessary for evaluating the analytical performance of NGS (Table 2.3) were taken into consideration and algorithms/modules for measuring the quality metrics were incorporated into the pipeline.

Table 2.3. Summary of quality control metrics for NGS data outlined by Nex-StoCT working group

Quality Metric	Description
Depth of Coverage	The depth of coverage characteristic of a particular region should be established
Uniformity of Coverage	The coverage across the targeted regions that must be achieved to produce reliable sequencing results should be established
GC bias	GC bias in all parts of the genome included in the assay should be determined
Transition/Transversion Ratio	The ratio of transitions to transversions should be comparable to published values
Base Call Quality Score	Informatics filters should be established to eliminate any reads with raw base calls lower than the established quality score
Mapping Quality	Informatics filters should be established to eliminate any reads that map to non-targeted
Removal of Duplicate Reads	Informatics filters should be established to eliminate duplicate reads resulting in clonal amplification
First base read success	The number of reads that pass the established quality filters should be established during assay validation
Decline in Signal Intensity	During assay validation the expected signal intensity across a read should be evaluated to establish the normal performance ranges and expected decline in signal intensity

Fastqc, a platform independent NGS quality tool, was selected for the quality analysis algorithm for the pipeline as it assesses the quality of the sequencing run and starting library material (Babraham Institute). Fastqc can import data from alignment files or raw NGS data, and reports an overview of quality statistics that may indicate problems or biases in the NGS data. There are 10 modules in the fastqc pipeline that report a value of pass, warning, or fail for the NGS library. Fastqc also outputs the raw data generated for the different metrics, making it flexible and easy to customize.

The basic statistics module outputs the total number of sequences and the maximum and minimum sequence length. The per sequence quality module helps to determine if the library has a portion of reads with low quality values due to poor imaging during the sequencing step. Poor quality sequences should be a very small percentage of the total sequences. Occasionally the 3` end of NGS reads can be of poor quality due to longer exposure to damaging chemicals, and should be trimmed to help with accurate genome alignment and variant detection (Figure 2.2).

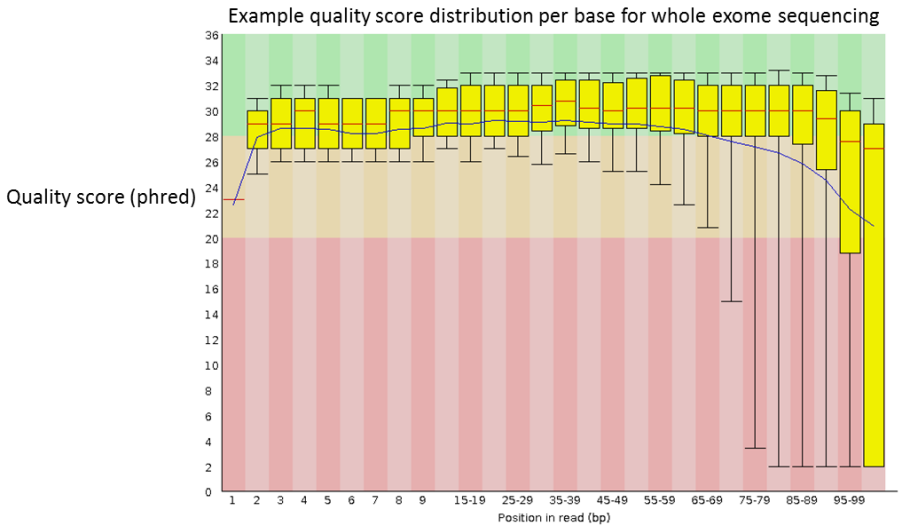

Figure 2.2. Example of the distribution of sequence quality scores for a whole exome sequencing library. Plot was generated using fastqc (Babraham Institute). The x-axis is position on the read, and the y-axis is the quality score (phred). A box plot is displayed at each position in the read. The green shading represents good quality scores, orange shading represents medium quality scores, and the red shading represents poor quality scores.

Per base sequence content module plots the proportion of each base position in a file for each of the four DNA bases (Babraham Institute). Many of the WES files received a warning for per base sequence content prior to trimming. This fastqc module issues a warning if the difference between A, T, G, or C is greater than 10% in any position. The datasets analyzed were exome-captured, and compositional biases are known to exist for exonic regions (Kozlowski, de Mezer, & Krzyzosiak, 2010). Furthermore, nucleotide composition is known to vary within the genome, and at the gene level (Louie, Ott, & Majewski, 2003).

Different segments of genes, have specific functional requirements, and exons have preferences for codon usage (Louie et al., 2003). Exons also undergo splicing, which also influences nucleotide content, as these signals are encoded in very specific nucleotide sequences. Previously it has been shown that the first exons within a gene are elevated in C+G, however, this trend shifts for exons further in the gene and exon edges are A+T rich (Louie et al., 2003). Figure 2.3 represents sequence biases within the gene structure, highlighting the G/C and A/T rich regions.

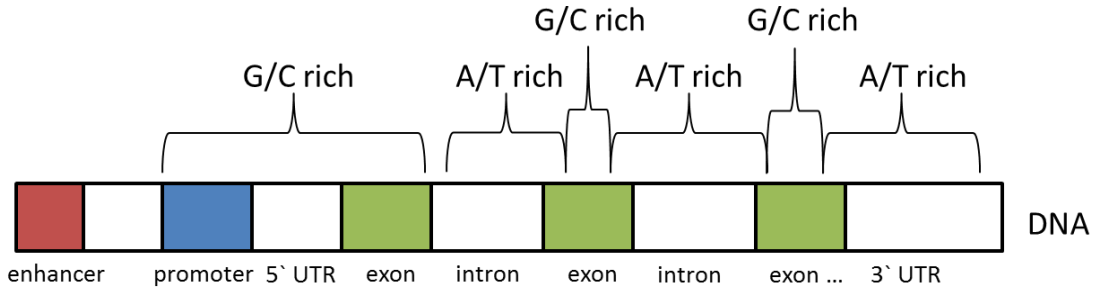

Figure 2.3. Gene structure and nucleotide composition. Red = enhancer. Blue = promoter. White = 3/5' UTR and introns. Green = exons. The brackets above indicate regions that are either G/C or A/T rich.

For the pediatric AML dataset, several of the libraries, both prior and post trimming had a slight enrichment for A+T bases compared to G+C. Figure 2.4 shows an example of per base sequence content for one of the WES libraries. Since known compositional biases are known for exonic regions, and all libraries after trimming had an A+T enrichment at a given base less than 30%. The libraries were deemed of appropriate quality and used in the downstream analysis.

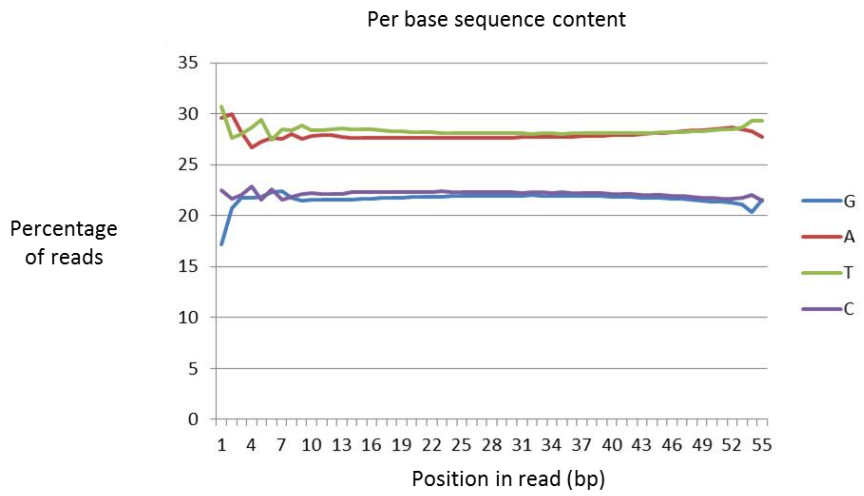

Figure 2.4. Example per base sequence content for whole exome sequencing dataset. The x-axis represents the position on the read, and the y-axis is the percentage of reads with a particular base. G = guanine (blue line). A = adenine (red line). T = thymine (green line). C = cytosine (purple line).

The per base GC content module calculates the GC content across the length of each sequence and compares it to a modelled normal distribution of GC content. The majority of the samples analyzed passed this module, and a few had a warning. Figure 2.5 shows an example that had a slight variation of the expected GC content. For all libraries used in the downstream analysis, the sum of the deviations from the normal distribution represented < 30% of the reads.

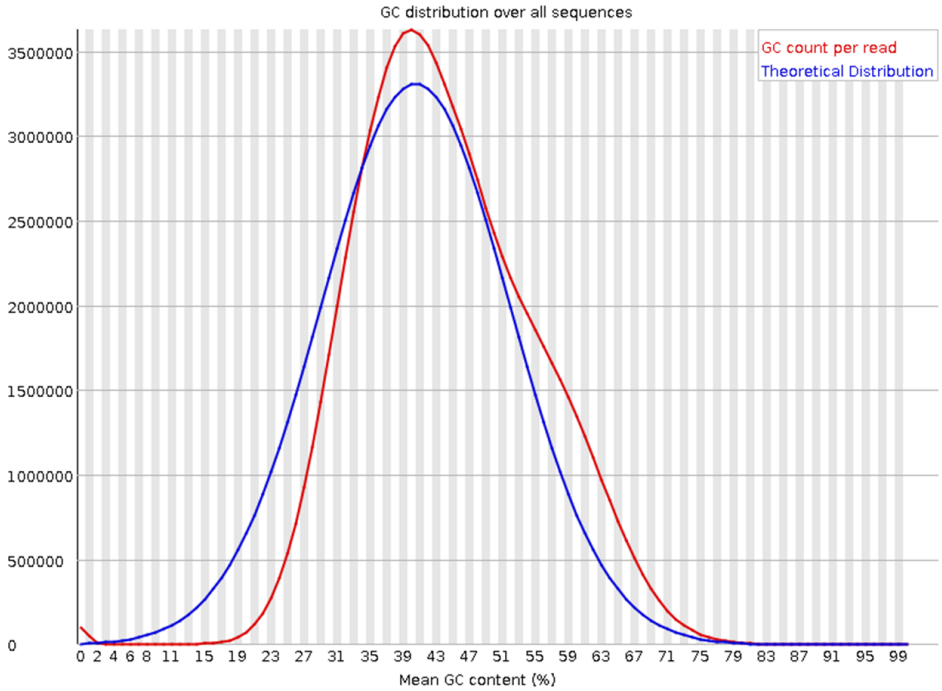

Figure 2.5. Example of GC distribution over all sequences within an exome sequencing file (read line) and the theoretical distribution (blue line). The x-axis is the mean GC content (%), and the y-axis is the frequency.

The per base N content module plots the percentage of base calls at each position for which N, no base call, appears. Only 2 libraries analyzed had an issue with per base N content, and these samples were excluded from the downstream analysis. The sequence duplication level module counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication.

This is important as libraries with high sequence duplication levels indicate a potential enrichment bias during library preparation. However, certain types of library preparations, such as RNA-seq, are known to have some bias that can lead to high duplication levels. But for WES the number of duplicated reads should be minimal.

For the pediatric AML study non-unique sequences for any given library were less than 20% of the total. Duplicate reads can be (and were) marked in downstream analysis steps and were therefore not trimmed from the libraries.

The fastqc module called overrepresented sequences calculates any sequences that may be overrepresented in the NGS reads, i.e. adaptors. All of the libraries analyzed passed the overrepresented sequence module after trimming with cutadpt, as no residual primers were detected or any other type of over-represented sequence.

The Kmer module measures the number of each 7-mer at each position in the library. It then uses a binomial test to look for significant deviations from an even coverage at all positions (Babraham Institute). Any kmers with a positional biased enrichment are reported. This module will report a warning if any kmer is imbalanced with a binomial p-value < 10^-5. Libraries which derive from random priming can show kmer bias at the start of the library due to an incomplete sampling of the possible random primers (Babraham Institute).

2.2 Summary WES Libraries

All of the libraries analyzed for the rare Mendelian disease passed all of the fastqc modules, as expected since they were trimmed prior to this analysis (Table 2.4). Although, all of the libraries had a warning for sequence duplication levels, a certain degree of duplication is possible due to targeted-capture biases (Table 2.4). Since the duplication levels were less than <25% for any given library they were deemed of appropriate quality and duplicate reads were marked in downstream analysis steps. A few of the libraries had warnings for the per base sequence content, but were deemed of appropriate quality for downstream analysis as the difference between A and T, or

G and C was less than < 20% in any position and compositional bias for WES are known to exist.

Table 2.4. Summary fastqc analysis per paired-end Illumina whole exome sequencing for rare Mendelian disease dataset. The subject ID refers to the patient, and the read is for either pair 1 or pair 2 of the library.

Subject ID	1		2		3		4		5		6		7		8		9		10		11		12		13		14	
Read	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Basic Statistics	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Per Base Sequence Quality	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Per sequence quality score	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Per base sequence content	P	P	W	W	W	W	P	P	W	W	W	W	P	W	W	W	P	P	P	P	W	W	P	W	W	P	W	P
Per base GC content	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Per base N content	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Sequence Length Distribution	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Sequence Duplication Levels	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
Overrepresented sequences	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Kmer Content	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	W	P	P	P	P	P	P	P	P	P	P

For the pediatric AML data set, 57 WES library files were analyzed. All of pair 1 reads prior to trimming either passed or had a warning for the following fastqc modules: basic statistic, per sequence quality content, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, over-represented sequences. For two of the libraries the pair-1 reads failed kmer content, and 33 failed per base sequence quality; however, after trimming all of the libraries passed per base sequence quality.

The pair-2 reads from the NGS libraries all passed or had a warning for the following fastqc modules: basic statistic, per base sequence content, per base GC content, and sequence length distribution. Forty-one libraries failed the following fastqc modules: per base sequence quality, per sequence quality, per sequence GC content, per base N content, sequence duplication levels, overrepresented sequences,

and kmer content. In general the pair-2 reads from the NGS libraries were of poorer quality compared to the pair 1 reads from the same NGS library.

Table 2.5 summarizes the criteria and thresholds implemented for the pediatric AML data set. Briefly, all reads were trimmed from the 5 and 3` end to remove any bases with < 20 phred score. A read had to have a minimum observed mean quality score > 27 to be used in the downstream analysis. The per base sequence content could not be > 30%, and the sum of deviation from the normal distribution for GC content had to be < 15%. Adaptor sequences were trimmed (Illumina), and an imbalance with a binomial p-value < ~10-5 was used for kmer analysis. All of the libraries had less than < 20% non-unique reads from the total.

Table 2.5. Summary of criteria applied for analyzing whole exome sequencing libraries from pediatric patients with acute myeloid leukemia

Module	criteria / Ranges applied
per base sequence quality	> 20 phred trimming
per sequence quality score	observed mean quality > 27 (0.2% error rate)
per base sequence content	< 30% in any position
per base GC content	sum of deviation from the normal distribution < 15%
Overrepresented sequences	trimmed adaptor sequences
kmer content	imbalanced with a binomial p-value < ~10^-5
duplication levels	< 20% of the total

2.3 Genome Alignment, Processing, and Quality Analysis

High quality genome alignment of NGS reads is essential for accurate detection of genomic variants. The tool selected for genome alignment was Burrows-Wheeler Aligner (bwa) (H. Li, 2013). Specifically, the NGS reads were mapped to the

human reference genome hg19 using bwa-mem version bwa-0.7.4. Bwa-mem was a recently developed algorithm added to the bwa alignment tool suite, and has similar features as bwa-sw. Bwa-mem is recommended for high-quality queries because it is faster and more accurate compared with bwa-sw, and has better performance than bwa-backtrack for 70-100bp Illumina reads. All of the samples had >95% of their reads mapped to the human reference genome hg19.

The Broad Institute has developed a publicly available software package, Genome Analysis Toolkit (GATK), that provides a suite of algorithms required for analyzing WES data including data quality assurance. Following GATK best practices (DePristo et al., 2011), the alignment files were processed using Picard Tools Version 1.67 (http://picard.sourceforge.net/) and GATK v3.1-1. The first step in processing the alignment files is to mark duplicate reads because they potentially represent a clonal amplification rather than a randomly sheared DNA fragment (Figure 2.6).

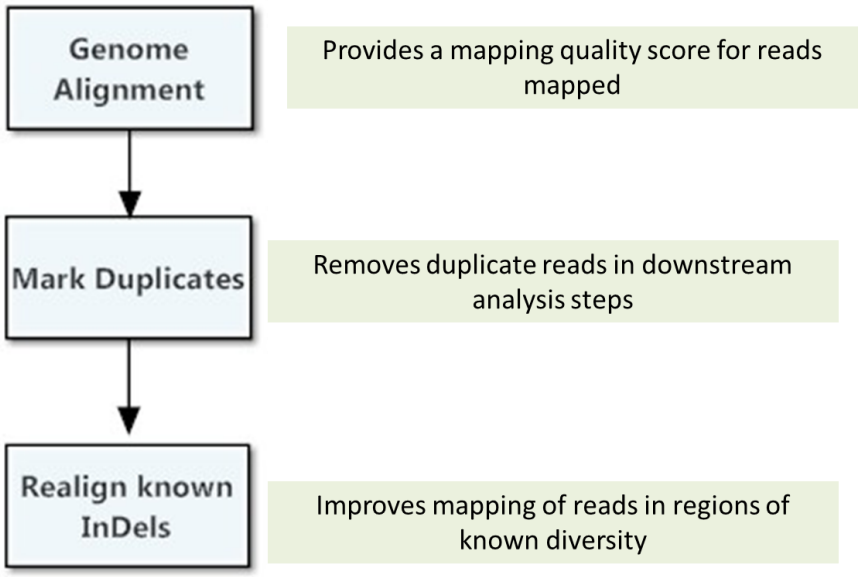

Figure 2.6. Workflow for processing genome alignment files. The first step is to align the next generation sequencing reads to a reference genome. Duplicate reads are marked, and re-aligned around known variants. InDels = insertions and deletions.

Next the alignment files are analyzed for potential intervals that need re-alignment based on known genetic alterations (Figure 2.6). These regions undergo a re-alignment step to optimize the alignment of the NGS reads to the reference sequence. Finally, the alignment file undergoes a recalibration of base quality scores based on the adjustments from the proceeding steps. After recalibration the quality scores, in theory, should be more accurate because the new score is closer to the actual probability of mismatching to the reference genome.

After the alignment files are appropriately processed it is important to assess the quality of the genome alignment. There are several different criteria that can be used, such as percentage of reads mapped to the genome, but it is important to consider the depth and breadth of coverage. Coverage calculations are especially

important when trying to understand variant calling capabilities and limitations. GATK provides an algorithm, Diagnose Target, that measures the depth and breadth of coverage for a defined list of genome coordinates (Figure 2.7). The output results from DiagnoseTarget are analyzed with a custom R script that allows visualization of the depth and breadth of coverage and enables the pipeline to output summary statistics for the exact exons captured during library preparation. The depth of coverage is how many times a nucleotide was sequenced, whereas the breadth of coverage is the average coverage per base per sequence interval.

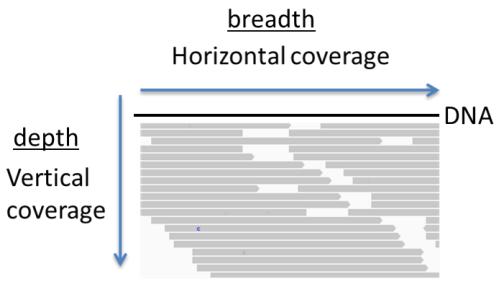

Figure 2.7. Depth and breadth of coverage of NGS reads (grey lines) aligned to reference sequence (black line).

For the rare Mendelian dataset, the alignment files were annotated with the kit used to capture the exons (NimbleGene). The depth and breadth of coverage was calculated for each individual using the NimbleGene annotation file. The average depth of coverage per exon per sample was ~75 reads (Figure 2.8); however, there were several extreme outliers, exons with excessively high and low coverage, for each sample. Previous studies have demonstrated that 15X average coverage is sufficient for homozygous SNP detection, and 33X for heterozygous, using Illumina short-read

technology (Sims et al., 2014) and (Bentley et al., 2008). Using the probe information provided by the manufacturer we can estimate the predicted depth of coverage to be 45 ((20,000,000 reads * 100bp) / 44,100,000 bp). The breadth of coverage is the horizontal coverage. For example, if an exon is 100 bp and 70 of the bp have at least 1 read or more, the breadth of coverage would be 70% for the exon. For the 14 patients analyzed < 3.5% of the exons had 100% coverage.

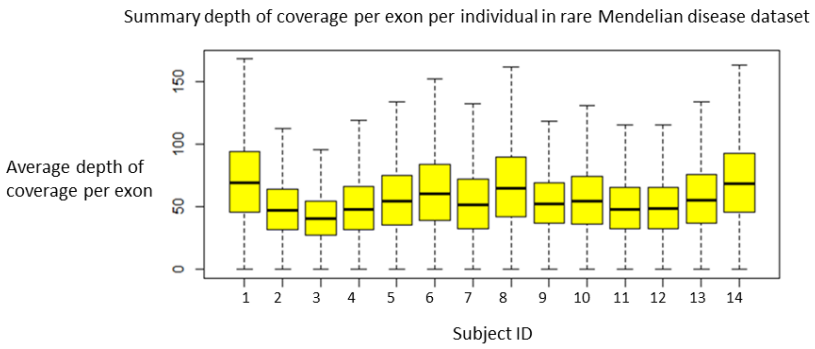

Figure 2.8. Boxplots of the depth of coverage per exon for WES generated from the rare Mendelian disease cohort. The x-axis is the subject ID, and the y-axis is the average depth of coverage per exon.

For the pediatric AML dataset, particular coding regions were of interest as our clinical collaborators are focused on therapeutically relevant oncogenes. The first gene inspected was *FLT3*, a known pediatric AML candidate gene. One of the major goals of this dissertation project is to determine the feasibility of using NGS to analyze known AML cancer genes, and *FLT3* is a top candidate. Figure 2.9 represents the depth of coverage calculations for all 24 exons annotated for *FLT3*.

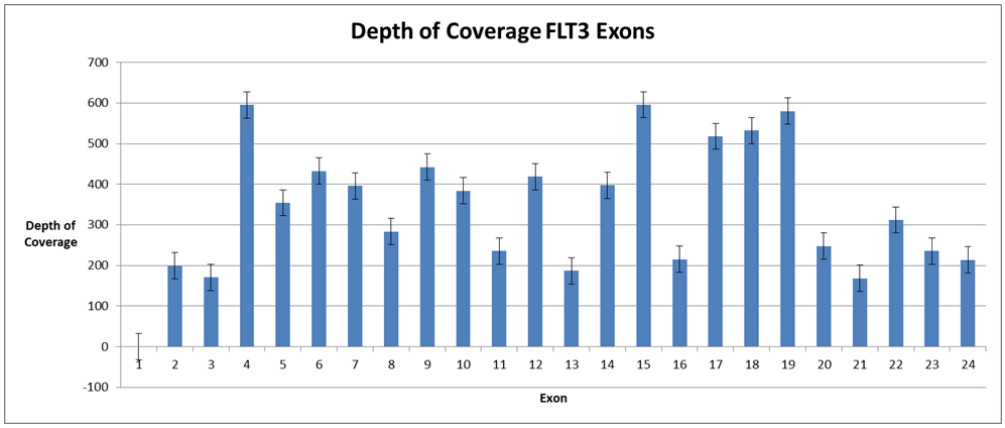

Figure 2.9. Depth of coverage *FLT3* exons. The x-axis represents the exon number in FLT3, and the y-axis is the average depth of coverage for the 19 samples analyzed.

Cancer samples often represent a heterogenic population of tumor cells along with health cells, and adequate depth of coverage for calling genomic variants may vary between samples. The heterogeneity of these samples, and the lack of understanding number of genome molecules captured, complicates the ability to accurately calculate allelic ratios and rare variants.

2.4 Rare Mendelian Disease Workflow

A bioinformatics workflow was created for analyzing NGS data generate from rare Mendelian disease cohort(s) (Figure 2.10). The pipeline starts with raw NGS reads and produces an annotated output that can easily be filtered depending on user input requirements. Quality control steps were implemented at each step (discussed in more detail below), and the overall workflow leverages publicly available algorithms and databases. The workflow also consists of custom modules, including annotation

of inheritance pattern, and was developed in collaboration with clinical researchers. The goal of the workflow is to translate big data into knowledge.

SNPs naturally occur between healthy individuals with estimates ranging from 1 in 1,000 to 1 in 1,500 nucleotides (Jorde & Wooding, 2004; Sachidanandam et al., 2001; Schneider et al., 2003). Collectively, SNPs result in ~3 million nucleotide differences using the estimated genome size of ~3 billion nucleotides (haploid). SNPs can alter the expression of genes, cause changes in translated amino acid sequences, and affect the binding of miRNAs. SNPs can occur in linkage disequilibrium (LD) with other SNPs located within the same genomic region. LD is the non-random association of alleles at two or more loci that have been inherited from the same chromosome. LD is influenced by several factors such as DNA recombination, mutation rate, and genetic drift.

An important consideration for building the bioinformatics workflow is the ability to distinguish between variants that are common versus those that are relevant to the disease phenotype. To aid in this process, population level data (minor allele frequencies) and functional annotations, such as gene ontologies and pathway information, are provided for downstream filtering strategies.

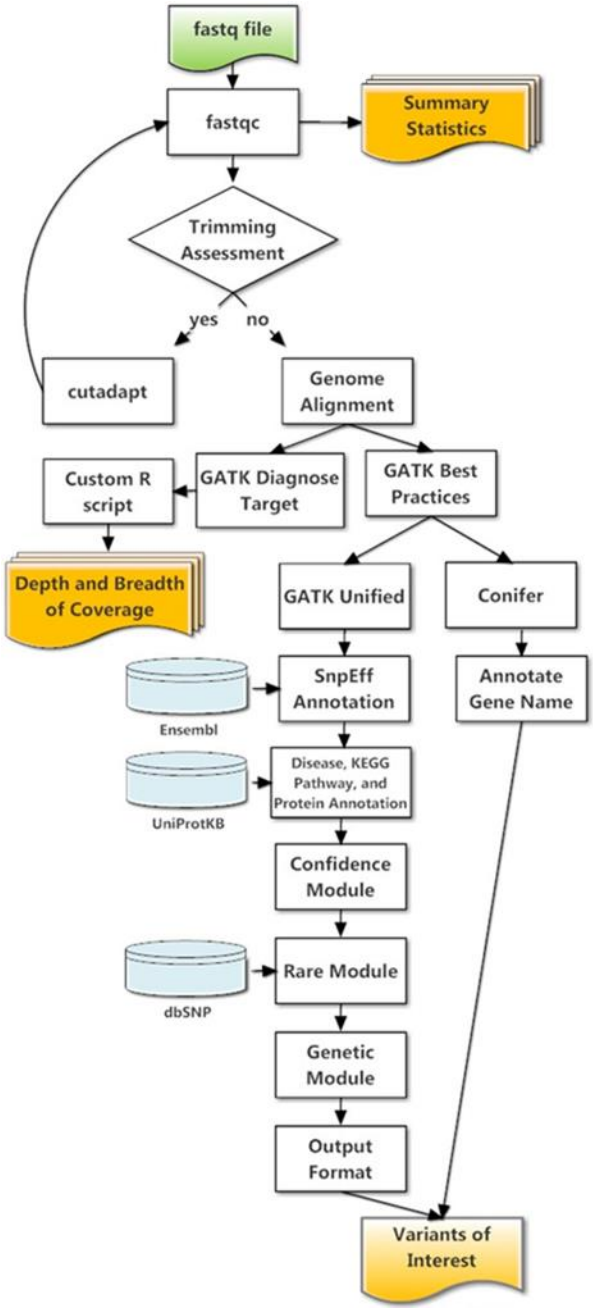

Figure 2.10. Overview of bioinformatics methodologies and custom pipeline for annotating and filtering single nucleotide polymorphisms and insertions/ deletions. Green box is the starting input fastq file. Square boxes represent processes in the workflow. Blue cylinder represents databases. Orange box represent output files from the pipeline.

Algorithms for detecting SNPs in genome reference alignment files have been extensively developed over the past decade, and are well established in the literature. For the dataset presented GATK Unified was selected for detecting SNP and InDel using the multi-sample option. GATK Unified was selected as the preferred algorithm as it is well-maintained and supported. Also, initial comparisons with Samtools and GATK Haplotyper indicated that this program was more consistent with detecting previously established SNPs. Although, since that comparison GATK Haplotyper has transitioned out of beta development and Unified has been retired.

GATK Unified v3.1-1 was used for SNP and small InDel detection using the default parameters (Mckenna et al., 2010). GATK Unified is cited as being sensitive for detecting InDels that are < 20 bp (Narzisi et al., 2014). The minimum base quality score (mbq) used during SNP detection was 17, and the minimum phred-scaled confidence threshold at which variants should be called was 30. Quality checks were implemented in the workflow for SNP and small INDEL detection. VCF tools is a publicly available package of programs designed to work with VCF files (Danecek et al., 2011). A combination of VCF tools and PSEQ (http://pngu.mgh.harvard.edu/~purcell/plink/) algorithms were used to extract metrics on SNPs and InDel detected.

Quality checks for SNP and InDel detection are an important factor for establishing bioinformatics methodologies in clinical applications. Therefore, quality checks were implemented in the workflow leveraging VCF tools (Danecek et al., 2011) and PLINK/SEQ (http://pngu.mgh.harvard.edu/~purcell/plink/). Variant call files (VCF) are the standard output files generated from genomic variant callers, such as GATK Unified, Haplotyper, and samtools. PLINK/SEQ is an open-source C++

library specific for analyzing human genetic variation data, with a specific focus to provide a platform for analytic tool development for variation data from large-scale resequencing and genotyping project, specifically WES and WGS studies.

Seven metrics were analyzed to help assess the quality of the raw and processed SNP data: number of non-reference genotypes, number of genotypes with a minor allele, number of heterozygous genotypes for an individual, number of total variants, genotyping rate for an individual, number of singletons (MAC == 1), and transition / transversion ratio. These metrics are described in more detail / context in Chapter 3.

The VCF files are annotated with SnpEff (Cingolani et al., 2012) and custom scripts. SnpEff annotates the variants with numerous characteristics such as gene name, Ensemble biotype, effect of variants, impact of variants, and transcript ID. The VCF are then annotated with 3 custom scripts, and then formatted into a user-friendly output (described in more detail in Chapter 3).

2.5 Pediatric AML NGS Workflow

A bioinformatics workflow was created for analyzing NGS datasets generated from cancer samples (Figure 2.11). The workflow starts with raw NGS data and generates knowledge maps that help to highlight important protein-protein interactions and mis-regulated pathways. The workflow enables detection of multiple types of genomic variants, and leverage publicly available algorithms and databases. It also includes custom modules for prioritizing somatic variants and annotating relapse specific mutations.

Somatic mutations are difficult to detect because they occur at low frequency in the genome and might only be present in a small fraction of the DNA molecules (Cibulskis et al., 2013). Often tools used for detecting germline SNPs are not recommended for detecting single nucleotide variants. The sensitivity and specificity of an algorithm to detect a somatic mutation is dependent on several characteristics such as sequencing depth, local sequencing error rate, and allelic fraction. Using a collaborative approach, The Broad Institute developed a somatic SNP detection algorithm called Mutect (Cibulskis et al., 2013).

Mutect analyzes both a normal and a cancer alignment file, from the same patient, simultaneously and consists of four key steps: 1) removal of low-quality sequence data, 2) variant detection in the tumor sample using a Bayesian classifier, 3) filtering to remove false positives resulting from correlated sequencing artifacts that are not captured by the error model, 4) designation of the variants as somatic or germline by a second Bayesian classifier.

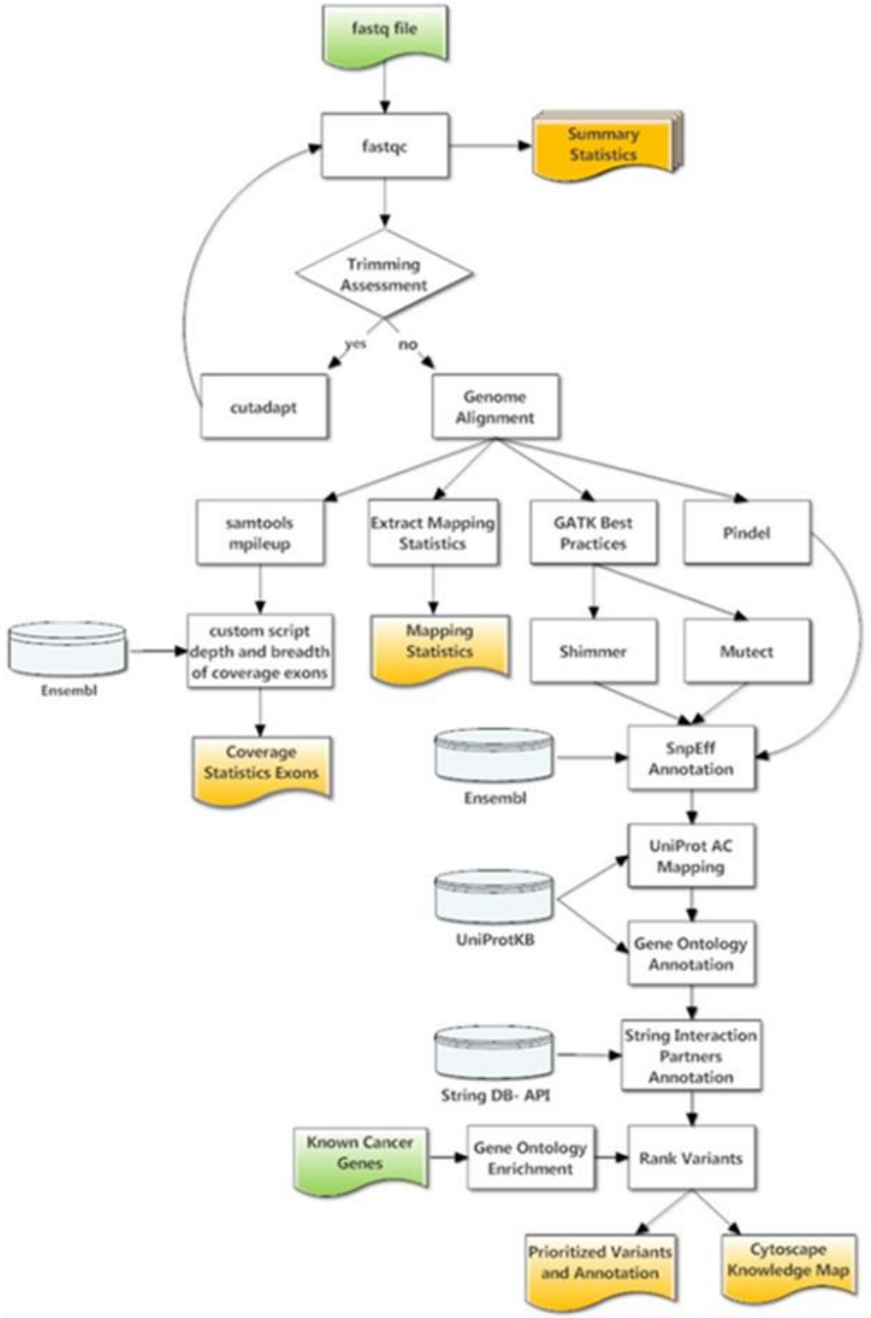

Figure 2.11. Overview of bioinformatics methodologies and custom pipeline for annotating and filtering single nucleotide variants and insertions / deletions. Green box is the starting input fastq file. Square boxes represent processes in the workflow. Blue cylinder represents databases. Orange box represent output files from the pipeline.

There are multiple approaches for detecting SV using NGS data. For example, *de novo* assembly, with either the complete dataset or un-mapped reads, is one strategy for detecting large SV (Y. Li et al., 2011). One limitation to this approach is that it can only detect homozygous SV because detecting heterozygous SV requires assembly of haplotype sequences, which is a complex problem that is not fully resolved.

Reference mapping strategies are another approach, and include concepts around split pair-end read mapping, read coverage depth analysis, or analysis of inconsistent insert size of paired-end reads. These approaches first require the NGS reads to be mapped to a reference genome, and then the alignment files are analyzed for genomic variants. The detection of SV using NGS data requires accurate prediction of copy, content, and structure. Often algorithms developed for detecting SVs are specific for a class of SVs, making it a necessity to incorporate multiple SV algorithms into the workflow.

Several different SV algorithms were tested, including pindel, breakdancer, and VarScan2. Pindel can detect breakpoints of large deletions, medium sized insertions, inversions, and tandem duplications by leveraging a pattern growth approach (Ye, Schulz, Long, Apweiler, & Ning, 2009). Previously Pindel has been cited for detecting an internal tandem duplication (ITD) in the FLT3 gene by using a pattern growth approach to analyze NGS data misaligned to the reference genome due to biological differences. This SV is a major focus of this dissertation as it is linked to treatment strategies and various sub-types of pediatric AML are classified by SV.

BreakDancer predicts 5 types of structural variants: insertions, deletions, inversions, inter- and intra-chromosomal translocations, and the results from

BreakDancer can be directly feed into Pindel to help enhance the analysis as a whole. VarScan is another package capable of detecting SVs, including copy number variations (CNVs) and InDels (Koboldt et al., 2012). Ultimately, the analysis pipeline requires the integration of all of these variant callers, because pediatric AML is a heterogenic disease that is not characterized by a few variants. For this dissertation project, the analysis focused on integrating Mutect (SNV) and Pindel (InDels).

The VCFs were first annotated with SnpEff, and then processed with a custom script for annotating gene ontologies and protein-protein interactions. The pipeline leverages publicly available resources provided by the UniProt consortium, plus additional resources such as STRING and cytoscape. The annotation steps are applied in the downstream prioritization of the variant data, which is described in further detail in Chapter 4.

Copy number variations (CNVs) were detected using the CoNIFER algorithm. A unique aspect of CoNIFER is the ability to detect rare CNVs in a small dataset, as most CNV algorithms require large datasets. CoNIFER uses a singular value decomposition method to normalize copy number variation (Krumm et al., 2012). The CNVs were annotated with gene information by comparing the CNV coordinates with ensembles gene annotation coordinate file.

Chapter 3

RARE MENDELIAN DISEASES

The test data set included 14 samples (6 probands and their unaffected parents and siblings) of Illumina paired-end whole exome sequencing, and all prior validated results were kept blinded to the bioinformaticians. Figure 3.1 represents the pedigree information for the dataset that was used to aid in the development of the bioinformatics methodologies. Ideally, when analyzing Mendelian diseases having complete trio datasets is preferred; however, it is not always possible to obtain genomic sequencing data from all individuals, which is a common challenge in the clinical setting. Therefore, the focus was on developing pipelines such that the downstream analysis would be robust and flexible for complete and incomplete trio data sets.

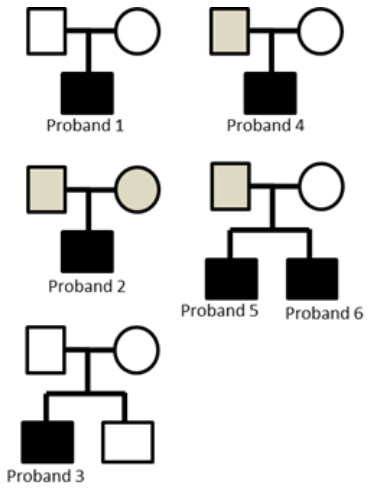

Figure 3.1. Pedigree overview test dataset rare Mendelian disease. Black fill = proband, white = unaffected plus next generation sequencing data available, grey = unaffected and next generation sequencing data was not available. Circles = female. Squares = Male.

SNPs were simultaneously called per individual using GATK Unified based on NGS reads mapped to reference genome hg19 (bwa). Table 3.1 summarizes the number of alternative alleles called along with other quality metrics generated by VCF tools and PLINK/PSEQ. Note the table displays un-filtered SNPs results.

Table 3.1. Summary of single nucleotide variants detected in the rare Mendelian disease data set.

Family	Individual	total variants	genotypes minor allele	genotyping rate	singletons	transition / transversion ratio
1	proband	421,972	119,559	0.44	753	2.1
	mother	511,566	147,356	0.54	5567	2.0
	father	445,033	126,871	0.47	3983	2.1
2	proband	455,220	132,903	0.48	7249	2.1
3	proband	460,964	127,515	0.48	1109	2.0
	mother	497,875	138,342	0.52	3048	2.0
	father	515,782	147,660	0.54	4039	2.0
	brother	487,733	135,717	0.51	1561	2.1
	half-sister	470,183	135,397	0.5	5086	2.1
4	proband	440,493	127,615	0.46	4216	2.1
	mother	431,266	124,383	0.45	3974	2.1
5	proband	427,036	146,209	0.45	4837	2.1
	proband	457,682	157,787	0.48	5220	2.1
	mother	487,126	172,677	0.5	9091	2.1

Typically a genotyping rate of >95 % is expected for high quality data (Anderson et al., 2010), and the raw un-filtered data was ~50%. Genotyping rate is the ratio between the number of genotypes called and the total number of genotypes, and was closer to ~99% for all samples once a read depth requirement of 2 was required (Figure 3.2), highlighting the importance of post-processing steps.

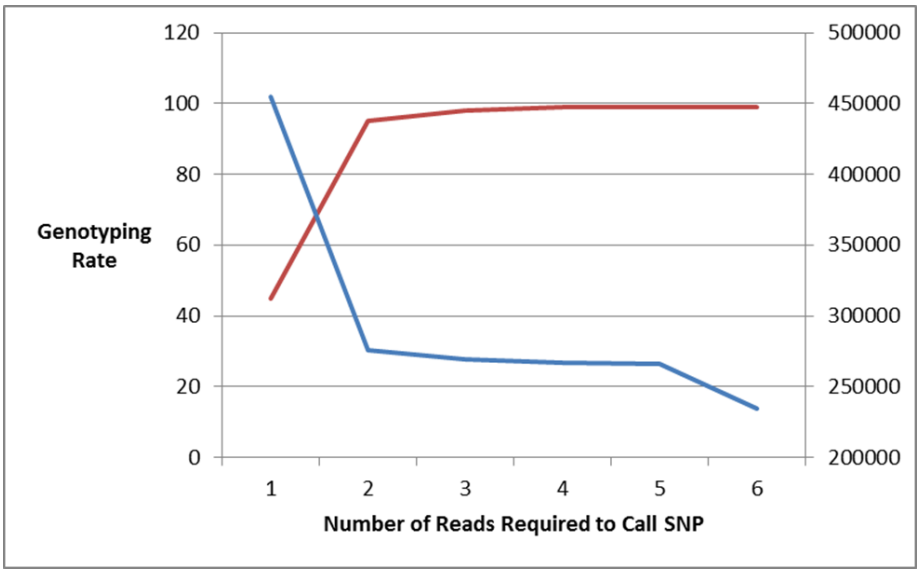

Figure 3.2. Number of single nucleotide variants detected (blue line, right y-axis) at various read depths (x-axis), and the genotyping rate of single nucleotide variants (red line, left y-axis) at various read depths

Transition to transversion ratio is expected to be ≥ 2 for exon-capture datasets (Keller, Bensasson, & Nichols, 2007), and the data was consistent with this trend. For the presented dataset the VCF was annotated with SnpEff Version 3.3a (Cingolani et al., 2012) using the package GRCh37.75 annotation recommended by SnpEff at the time. The majority of the SNPs were annotated with protein coding (Figure 3.3), and these variants were the main focus of the downstream analysis. This approach has its limitations as variants located in miRNA regions could be of interest, but require different analysis strategies.

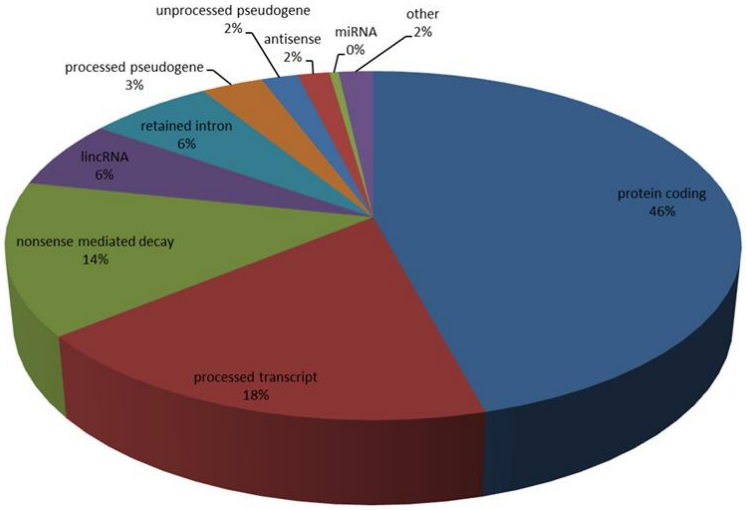

Figure 3.3. Distribution of ensembles biotype for regions of the genome that single nucleotide polymorphism was detected (un-filtered).

It is important to note that applying a confidence of at least 5 reads being required to call a variant, increases the percentage of variants in protein coding regions to ~80%. The kit used for NGS, SeqCap EZ Human Exome Library V2.0, covers ~20,000 in the human genome. The probes are designed using annotation files from NCBI and miRBase and in total cover 44.1 Mb of the genome with 36.5 Mb specific to protein coding regions (83%). The regions captured include genomic segments that flank regions of the genome that are annotated to be transcribed.

The first custom script generated, Confidence Module, uses both the genotype quality score (GQ), calculated by GATK Unified, and read depth as filterable annotations. The Confidence Module annotates the VCF (Variant Call Format; v4.2) with either PASS (GQ > 30 and read depth >= 5) or FAIL. The Confidence Module was developed with flexibility to allow the user to specify whether both alleles need to have the same read depth.

This is an important consideration as allelic imbalance can arise from mosaic samples. Mosaicism could indicate a mutation that happened early in development and is therefore not present in every cell of the body (Graham & Hennekam, 2014). However, such apparent allelic imbalance in the data may also arise artificially as a result of errors during sequencing.

The next custom script generated, Rare Module, first annotates the variant with a global minor allele frequency using dbSNP data (Sherry et al., 2001), and then annotates the variant as PASS or FAIL based on the user specific requirements. For rare inherited disorders, a common paradigm is that a true causative SNP will have an extremely low minor allele frequency and/or may not be present in any databases, such as dbSNP. With the vast amount of NGS data being produced and analyzed, and the incorporation of disease and non-disease results into public databases, the pipeline did not prioritize variants that were considered novel (i.e. no dbSNP RS identifier). Instead the SNPs were annotated with a global minor allele frequency, using dbSNP, and allowed the user to specify the minor allele frequencies of interest.

For the dataset presented a minor allele frequency of $\leq 4\%$ was used. A quality check implemented for this module was the creation of a distribution plot of the annotated minor allele frequencies. During the initial development of this module it was noted that several of the SNPs had minor allele frequencies reported by dbSNP greater than 50%, which is a contradictory metric indicating that for these alleles the reference genome used for alignment (hg19) represents the minor allele and not the major allele (Figure 3.4). This is a very important point, which can be easily overlooked in analysis resulting in true minor alleles being filtered because the reported frequency was for instance in the 96 to 100% range rather than the expected 0

to 4%. Therefore, the Rare Module was enhanced to take into consideration alleles that were identified at positions in which the reference genome does not represent the major allele.

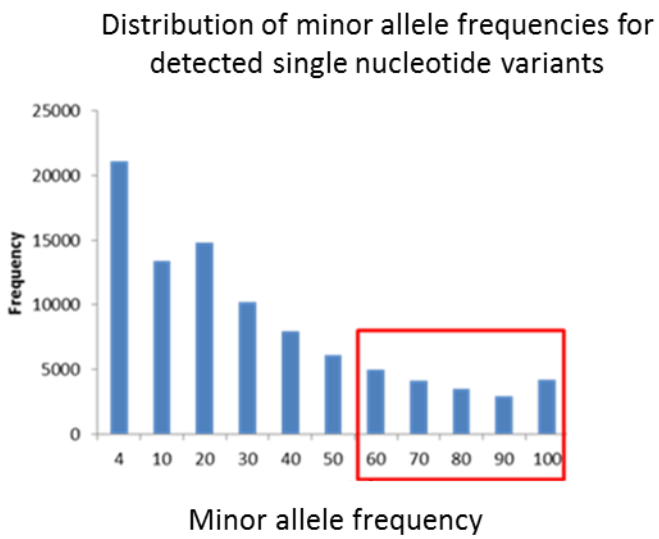

Figure 3.4. Representation of the distribution of minor allele frequencies detected from a single whole exome sequencing library. The x-axis is the minor allele frequency, and the y-axis is the frequency at which the minor allele occurs. The red box represents alleles that are the major allele.

VCF files were then annotated with the Genetic Module, to prioritize potential variants based upon observed heritability patterns. The Genetic Module algorithm can annotate the alleles with the following heritability patterns: recessive, dominant, *de novo*, X-linked, or Y-linked. For the dataset presented a recessive inheritance pattern (Table 3.2) was used based on the pedigree information and the expert knowledge of our research collaborators. It is important to note that during the filtering process, no

candidate variants are truly removed from the VCF. They are simply annotated with PASS/FAIL indications for given filters allowing for flexibility to easily re-filter the data.

Table 3.2. Allele annotations for recessive and dominant genetic analysis. A 0 indicates the reference allele, and a 1 indicates an alternative allele.

	Unaffected Parent	Affect Parent	Proband
Recessive	0/0 or 0/1	1/1	1/1
Dominant	0/0	0/1 or 1/1	0/1 or 1/1

An allele was considered a candidate for recessive inheritance if the parents were 0/1 (heterozygous) and the child was 1/1 (homozygous alternative). Likewise, a variant shared on an X-linked gene was considered a candidate when shared uniquely between the proband and his heterozygous mother. In previous validation studies of NGS data, our clinical collaborators had seen evidence of allelic imbalance resulting from tissue specific somatic mosaicism (i.e. blood vs buccal cells).

Therefore, we also annotated alleles that either of the parents were 0/1 or 0/0 (homozygous reference) and the child was 1/1 as potential candidates that would need further validation for a recessive pattern of inheritance. Any variant that either of the parents were 1/1 and the child was 1/1 were not considered as relevant as none of the parents were affected. All of the probands had ~200-500 variants identified after the custom modules (Confidence, Genetic, and Rare) were run on the data (Table 3.3).

Table 3.3. Summary of SNPs for confidence, rare, and genetic analysis (clean up top title line- center with under-line = total number of variants; and then start, post confidence, etc.

		Number of single nucleotide variants				
Family	Individual	Start	Confidence	Minor allele frequency	Genetic analysis	Controls substracted
1	proband	421,972	185,033	70,753	385	202
	mother	511,566	212,023	82,205		
	father	445,033	195,864	74,950		
2	proband	455,220	201,467	76,843	NA	506
3	proband	460,964	203,705	77,384	481	417
	mother	497,875	200,679	77,855		
	father	515,782	204,281	78,865		
	brother	487,733	231,720	81,722		
	half-sister	470,183	209,129	79,959		
4	proband	440,493	198,252	75,692	833	318
	mother	431,266	192,376	73,843		
5	proband	427,036	186,047	69,891	543	306
	proband	457,682	202,122	76,133	543	306
	mother	487,126	209,695	79,244		

Depending on the proband, ~15% of the prioritized variants caused a non-synonymous change (Figure 3.5). The majority of the SNPs were located in intron regions near exon boarders and downstream of annotated genes.

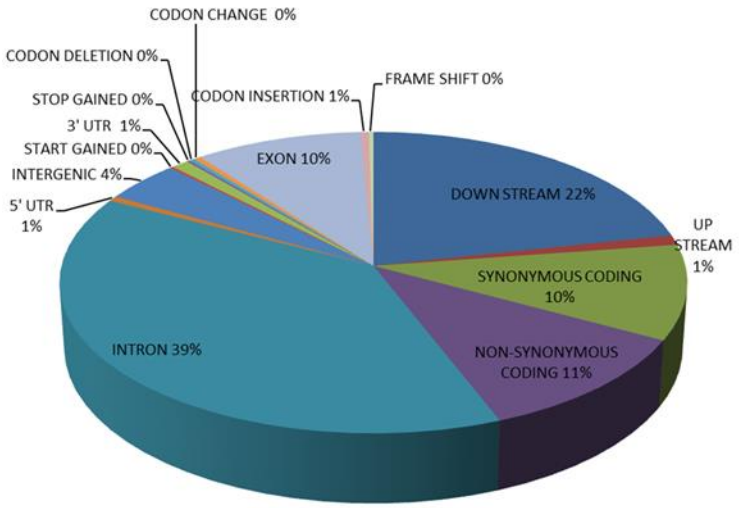

Figure 3.5. Distribution of the effects of single nucleotide polymorphisms detected for an individual whole exome sequencing dataset.

After the robust annotation and filtering of the variants, the VCF is then decomposed into focused lists of genomic variants of interest for each individual. A user-friendly Variant Output File (Table 3.4) is generated to allow the clinician to easily filter on desired characteristics such as minor allele frequency or effect of variant. A single variant is represented per line (row) and the filterable columns can be divided into 5 categories: genomic, protein, pathway, disease, and patient specific. Table 3.4 lists all of the filterable characteristics and provides a single row example from the Variant Output File (Example Variant Output). By providing the user-friendly Variant Output File, the pipeline connects numerous resources and enables a flexible filtering strategy.

Table 3.4. Filterable characteristics of the variant output file generated from the bioinformatics workflow.

Characterisitc	Feature	Resource	Example Row Variant Output
Genome	Chromosome	hg19	chr10
	Position of Variant	hg19	17171656
	Variant ID	dbSNP	rs76788243
	Reference Allele	hg19	T
	Alternative Allele	GATK Unified	G
	Quality of Variant	GATK Unified	9485.04
	Gene Name	HUGO	CUBN
	Effect of Variant	SnpEff	non synonymous change
	Minor Allele Frequency	dbSNP	0.034
	Rare Module	Custom	pass
Protein	Biotype	Ensembl	protine coding
	Amino Acid Change	SnpEff	I37L
	Protein Feature	UniProt	mature (processed) chain
	UniProt Accession	UniProt	O60494
Pathway	KEGG Pathway	KEGG	Organismal Systems
Disease	Gene Associated with Disease	UniProt	Recessive hereditary megaloblastic anemia 1 MIM:261100
	Genetic Inhertance Pattern	Custom	Recessive
Patient Specific	Genotype Probands	GATK Unified	G/G
	Confidence Module	Custom	pass
	Allele Depth	GATK Unified	101

To further enhance the development of the pipeline, text and data mining modules were developed. The unique gene lists for each sample were combined into one large list and batch up-loaded into the STRING interface to view protein-protein interactions between the translated gene candidates. The network text file generated by STRING was downloaded and viewed inside Cytoscape. The STRING analysis yielded several protein-protein interactions between the genes of interest for each of the probands and provided a potential system for connecting genes with genomic variants between different probands. However, a single mutation or gene was not discovered that could link all of the probands together.

Variants that passed all three custom modules and had one of the following SnpEff categories, codon change and insertion, codon insertion, frame shift, non-synonymous coding, and start gained, were mapped to a UniProt Accession and

loaded into the custom iProXpress web interface (Huang, Hu, Arighi, & Wu, 2007; McGarvey, Zhang, Natale, Wu, & Huang, 2011). The web-based iProXpress provides tools for functional profiling, such as pathway and GO enrichment analysis, and allows for custom display of selected fields from more than 150 underlying databases, such as OMIM disease information (Figure 3.6).

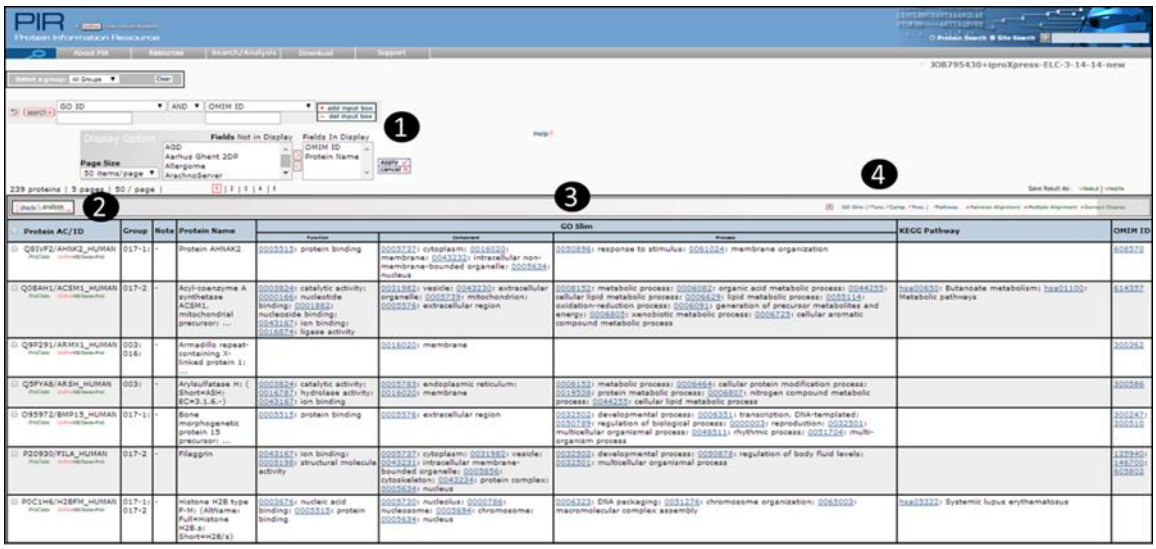

Figure 3.6. Integrating Genomic variant data with protein and disease rich information. The iProXpress web interface (1) User can customize the fields displayed on webpage. (2) User has quick access to protein rich information resources through UniProt resource. (3) Gene Ontology information displayed and enables functional enrichment analysis. (4) Pathway annotation and enables pathway enrichment analysis.

SNPs and InDels are not the only type of genomic variant that can influence protein coding regions of the genome. Copy number variations (CNVs) in gene-coding regions may also influence gene expression levels (Zhao, Wang, Wang, Jia, &

Zhao, 2013), and it has been previously reported that about half of identified CNVs overlap with protein-coding regions (Sebat et al., 2004). Numerous studies have linked serious developmental and malformation disorders, such as DiGeorge and Prader-Willi syndrome, to genome deletions events (Conrad, Andrews, Carter, Hurles, & Pritchard, 2006).

Recent studies support that CNVs account for about ~12% of the genomic differences between human (Redon et al., 2006). CNVs, such as deletions, that are associated with rare diseases should also be rare in the general population. Interestingly, it has been suggested that SNPs in region that are hemizygous for a deletion are generally miscalled as homozygous for the allele that is present (Conrad et al., 2006). Therefore, the SNPs will violate the rules of Mendelian transmission, and linking SNP data with CNV can lead to a better understanding of gene deletion events (Conrad et al., 2006).

Computational algorithms for determining structural variants, such as CNVs, from NGS reads aligned to a reference genome have various strategies. For example, some tools use paired-end and/or longer read mapping to identify "breakpoints" in the DNA coverage, while others use differences in read depth. Germline CNV covers anywhere from 3-12% of the genome and usually occur in regions of the genome that overlap between people (Liu et al., 2013). A CNV associated with a rare Mendelian disease should also be rare in the population. CoNIFER, an algorithm specific for CNV, is capable of detecting rare CNVs in small datasets by using a singular value decomposition method to normalize copy number variation (Krumm et al., 2012).

All 14 samples, plus three control samples chosen from the 1000 Genome Project (NA18517, ERR034551; NA18507, SRR764745; NA18956, SRR766028),

were analyzed with CoNIFER (Version 0.2.2) simultaneously. Using the inflection point in the singular values plot, it was determined that a single value decomposition of 3 was appropriate for the analysis. Table 3.5 is a summary for the copy number variations detected.

Table 3.5. Summary CoNIFER results

	Total	Duplications	Deletions	chromosomes
Proband 1	8	2	6	16
Mom	5	5	0	13, 14, 16
Dad	11	3	8	14, 16, 19, 5
Proband 2	6	5	1	13, 14, 4, 5, 6, 8
Proband 3	3	1	2	19
Mom	4	3	1	19, 7, 9
Dad	2	2	0	1, 13
Half-sister	7	6	1	1, 16, 17, Y, 5, 7
Proband 4	1	0	1	1
Mother	2	2	0	15
Proband 5	10	5	5	1, 16, 19, 6, 7
Proband 6	8	1	7	1, 16, 19, 4
Mom	4	2	2	1, 17, Y

The genomic coordinates of the CNVs were cross-referenced with the Database of Genomic Variants and Decipher (http://dgv.tcag.ca/dgv/app/home). Several of the genomic coordinates of the CNVs over-lapped with genomic coordinates in Decipher, which is a database of chromosomal imbalances and related phenotype found in humans (Firth et al., 2009). Three of the probands had deletions and duplications on chromosome 16 in a similar region. Two hundred and six consented Decipher patients were reported with duplication and deletions in this

region. Two syndromes were reported for this region related to neurocognitive disorders (Table 3.6).

Table 3.6. Decipher syndromes with copy number variations in region of interest

Syndrome	Chromosome	Start Position (bp)	End Position (bp)
16p13.11 recurrent microdeletion neurocognitive disorder susceptibility locus	16	14986684	16486684
16p13.11 recurrent microduplication neurocognitive disorder susceptibility locus	16	14986684	16486684

The CNV results were also visualized using the plot call scripts provided by CoNIFER (Figure 3.7). This region was selected for graphical display because three of the probands from two unrelated families had a similar deletion pattern, and there have been recent publications (Tropeano et al., 2013) linking this region of the human genome with phenotypes closely related to this example. There are multiple genes within the genomic region that the deletions were detected on chromosome 16, and of interest to our clinical collaborators was *XYLT1*. This gene encodes for the xylosyltransferase 1 protein, and is involved in the biosynthesis of glycosaminoglycan. Upon cross referencing these results with the SNP data, our clinical collaborators were able to link 2 additional probands to this region of interest.

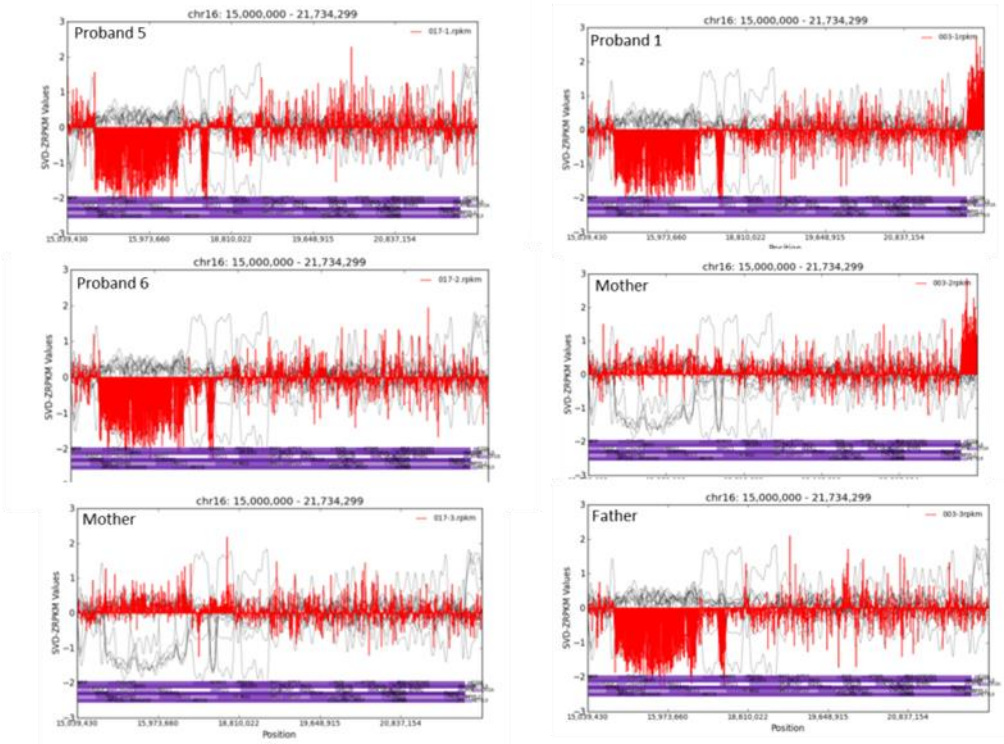

Figure 3.7. Copy Number Variation Detection. Graphical display of copy number variation detected by CoNIFER. X-axis = SVD-ZRPKM values for each exon calculated by CoNIFER. Red lines = SVD-ZRPKM values for each probe from the sample of interest. Purple bars = genes. Grey lines = smoothed SVD-RPKM values for each probe for a given sample.

The clinical collaborators for this research project were able to verify the CoNIFER results using a different technology / platform. CytoScan HD Array can reliably detect 25-50 kb copy number changes across the genome at high specificity with SNP (allelic) call corroboration. Recently, the FDA has approved the use of the Cytoscan Dx technology for detection of chromosomal variations in patients with various disorders. The Cytoscan results indicated a loss of copy number in the same region as the CoNIFER analysis (Figure 3.8).

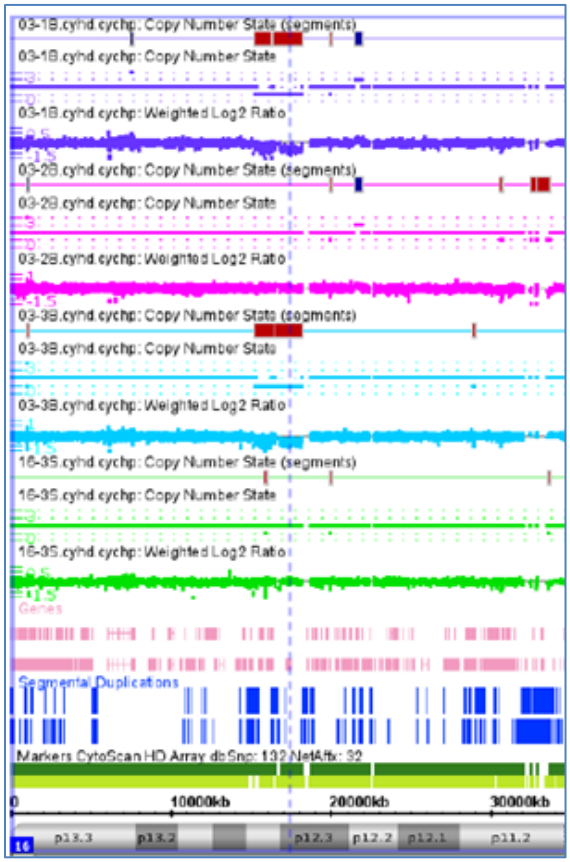

Figure 3.8. Cytoscan Results. Clinical collaborate Deborah Stabley from Nemours Alfred I. DuPont Hospital for Children provided the cytoscan results. The results indicate a copy number deletion on chromosome 16 in a similar location as the CoNIFER results.

A comprehensive analysis of the Mendelian disease presented required integration of SNPs and CNVs, as a single allele change in all probands was not discovered. By combining the SNP results with the copy number variation results a candidate gene / region of the genome was identified as being a potential clinically relevant gene of interest for this rare disorder. This gene is still currently under investigation by the clinical collaborators as it lies in a region for which the reference genome is incomplete and hence a region that was poorly captured from the WES.

A main gap/issue with analyzing genomic NGS data from a Mendelian disorder is translating the raw data into a focused list of variants that are relevant to the diseased phenotype. One algorithm developed to help address this issue is Phevor, a method that combines multiple biomedical ontologies together to prioritize and rank a list of variants (Singleton et al., 2014).

One of the limitations of Phevor is that it requires the researcher to first identify a focused list of variants of interest from the raw data, as it does not process the NGS data from start to finish. The methodologies developed for this dissertation project address these types of limitations, and provides a pipeline for analyzing the data from start to finish. It uses standard file formats, allowing the researcher to analyze the data at various points with other tools such as Phevor.

Specifically, this dissertation project addresses the need to develop modules that allow the researcher to create a focused list of genomic variants. Scripts were developed that can annotate the genomic variants with the appropriate inheritance pattern, confidence, and minor allele frequency. These custom modules allow a researcher to bin variants of interest together and prioritize downstream validation efforts. Furthermore, custom modules were created to link the variants to disease specific information (OMIN) and protein specific information (protein function, domain information, etc.).

The bioinformatics methodologies presented incorporated several different types of genomic alteration detection methods and allowed for a comprehensive understanding of the genomic architecture for the rare diseased analyzed. By applying a system of annotations, prioritizations, inheritance filters, and functional profiling and analysis, a unique methodology for further filtering of disease relevant variants that

impact protein coding genes was developed. The methodology creates a focused list of potential clinically relevant variants.

There are some limitations to the pipeline developed as it focuses on protein coding regions of the genome, and genomic variants located in other regions, such as miRNAs, are not thoroughly analyzed. There are also technical limitations of WES, like probe-capture bias. Ultimately it would be interesting to analyze other types of NGS data from this same patient cohort, i.e. RNA-seq and/or whole genome sequencing, as each dataset can complement the other providing a more comprehensive analysis of how particular SNPs and/or CNVs are contributing to the genetic disorder.

Whole exome sequencing generates massive amounts of data that even when processed can be difficult for clinicians and biomedical scientists to analyze and apply appropriately in the medical health field. Rigorous bioinformatics methodologies are required to analyze the data with appropriate statistical methods that will ultimately link the genetic data to the disease phenotype. In total 14 exome NGS samples were analyzed for the presented data set, and the innovative annotation of variants allowed a thorough analysis of the genomic alterations by using a system biology approach for data analysis.

After applying filters for high impact variants and recessive variants inherited from parents, a focused list of potential clinically relevant variants was generated for our clinical collaborators. They were able to compare the results of our analysis with other pipelines, which had already enabled the identification of the *XYLT1* as being clinically relevant for this Mendelian disorder. Our pipeline developed did detect

similar alterations compared with these other methodologies, although there were additional genomic alterations prioritized in our data analysis.

Chapter 4

PEDIATRIC ACUTE MYELOID LEUKEMIA

4.1 Overview Project

The database of Genotypes and Phenotypes (dbGaP) was developed by the NCBI to help distribute and archive the results of studies that investigate the interactions between genotypes and phenotypes. One of the many datasets deposited in dbGaP is from a National Cancer Institute (NCI) funded project called Therapeutically Applicable Research to Generate Effective Treatments (TARGET). The goals of TARGET are to determine the genetic changes that drive the initiation and progression of childhood cancers with the aim of identifying therapeutic targets and prognostic markers that will lead to more effective treatment strategies (https://ocg.cancer.gov/programs/target/overview).

TARGET is a collaborative consortium of extramural and NCI investigators, where the majority of the team members are from the Children's Oncology Group, otherwise known as COG. As a NCI funded clinical trials group, COG consists of more than 9,000 experts in childhood cancer at more than 200 leading children's hospitals (https://www.childrensoncologygroup.org/index.php/about-us). It is the largest organization focused exclusively on childhood and adolescent cancer research.

More than 90% of children diagnosed with cancer each year in the United States are cared for at a COG's member institute. The mission of COG is to cure all children and adolescents with cancer, by reducing the short and long-term complications of cancer treatments. They focus on research strategies that can help

determine the cause of childhood cancers, which will ultimately help guide efforts to prevent disease onset.

The TARGET program first focused on two pilot projects, high-risk subtypes of acute lymphoblastic leukemia (ALL) and neuroblastoma (NBL). The success of these projects lead to the expansion of TARGET's efforts, and currently they work on acute myeloid leukemia (AML), osteosarcoma, select kidney tumors, ALL, and NBL. TARGET researchers work collaboratively to generate, analyze, integrate, and interpret high quality genomics data.

The TARGET datasets are available to the research community through a data request process controlled by dbGaP. Furthermore, the consortium has a dedicated webpage (https://ocg.cancer.gov/programs/target), which provides the research community with descriptions of available datasets and the corresponding clinical trial information (Figure 4.1). Not all of the data generated by TARGET is currently available to the research community, but the long term goal is to make it accessible.

Figure 4.1. Therapeutically Applicable Research to Generate Effective Treatments webpage (https://ocg.cancer.gov/programs/target).

The type of datasets generated by TARGET range in analyte type (DNA / RNA) and NGS applications (Figure 4.2), enabling the development of informatics capable of integrating omics data. For this dissertation project, the AML data was requested through dbGaP and an appropriate IRB protocol was submitted to the University of Delaware (Appendix C). The experimental design for the project was to collect 3 samples per patient: diagnosis, remission, and relapse (Figure 4.2). The remission sample was designated as the healthy control, and in theory can be used to screen out germline mutations from somatic.

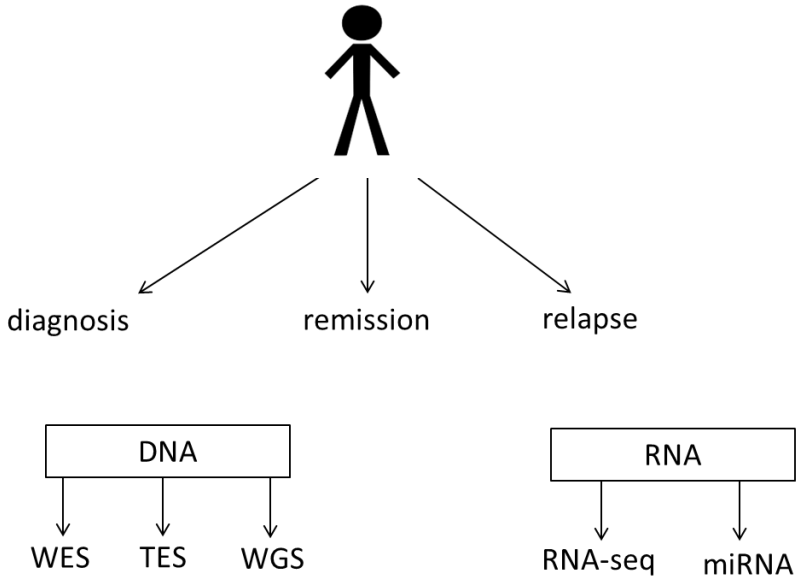

Figure 4.2. Experimental design for pediatric AML dataset. For each individual patient a bone marrow or peripheral blood samples was taken at diagnosis, remission, and relapse. DNA/RNA was extracted from the samples and prepared for NGS. WES= whole exome sequencing. TES = targeted exome sequencing. WGS = whole genome sequencing. RNA-seq = transcriptome sequencing. miRNA= mircoRNA.

In total, using the Illumina platform, there were 20 patients with WES, 30 patients with WGS, 800 patients with TES, 200 patients with RNA-seq, and 300 patients with miRNA. The first samples received for this project were the WES files, and will be the main focus for the work presented. The majority of the samples analyzed were bone marrow, and it is important to consider the biological characteristics for this type of sample as it influences interpretation of the data.

There are two different types of bone marrow, red and yellow. Red marrow is considered the hematopoietic tissue, whereas yellow marrow is fat cells, cartilage, and bone. Hematopoietic stem cells (HSCs) have self-renewal capabilities combined with the ability to progressively differentiate into common myeloid progenitor or common lymphoid progenitor cells (Figure 4.3) (Kosan & Godmann, 2016). Myeloid progenitors give rise to megakaryocytes, eosinophils, basophils, erythrocytes, monocytes, and neutrophils.

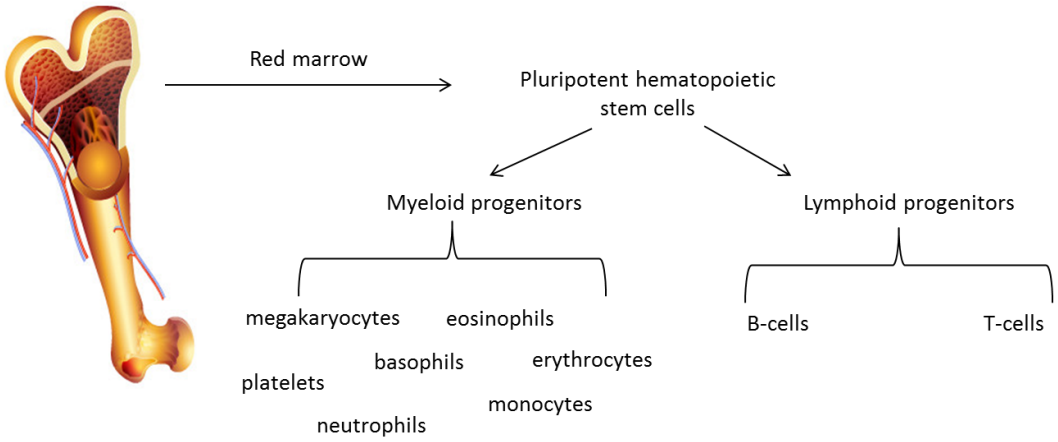

Figure 4.3. Overview of red bone marrow and pluripotent hematopoietic stem cell differentiation. Pluripotent hematopoietic stem cells differentiate into myeloid progenitors (megakaryocytes, eosinophils, monocytes, platelets) or lymphoid progenitors (B and T cells).

The progenitor cells and differentiated cells within the red marrow have distinct cell surface markers. Myeloid progenitor cells are often characterized by CD34 and CD127, and more differentiated cells, like megakaryoblasts, have other important cell surface markers such as CD38 (Novelli, Ramirez, & Civin, 1998). Techniques developed in flow cytometry can help with binning / sorting cells prior to DNA extraction; however, in general the cellular composition of red marrow is heterogenic, which is further complicated in leukemia as there is a mix of healthy and diseased cells. This is important consideration as it impacts the interpretation of allelic ratio / proportions.

4.2 Single Nucleotide Variant Analysis

For patients with WES data, COG provided a list of 254 verified variants, from either the relapse or diagnosis sample, for the 20 patients that were analyzed. Algorithms used for detecting single nucleotide variants (SNVs) have different strategies for dealing with sensitivity and specificity issues, which can alter false positive and negative rates. Recently a comparison of two bioinformatics pipelines (SomaticSniper and Mutect) on the same genomic data set generated from 133 AML patients, highlighted that the tools had inconsistent results (Bodini et al., 2015).

It is important to keep in mind that cancer samples are heterogeneous, and represent a complex mixture of leukemic cells (and possible contamination with health cells). Additionally, the remission sample can be contaminated with residual leukemic cells, which can result in the detection of somatic mutations that might be flagged as germline. Two different algorithms were used to call single nucleotide variants from the WES, and the results were compared to the provided list of verified variants.

Mutect (Version 1.1.4) detected 228 of the variants (90%), and of those variants detected 210 (83%) passed as high quality somatic mutations. All of the diagnosis samples had > 83% of the verified somatic mutations detected by Mutect, with an average detection of ~94% (Figure 4.4). For eight of the diagnosis samples, 100% of the verified variants were detected. The relapse samples had a slightly lower average of detection for the verified variants (91%) compared to the diagnosis samples. Ten of the relapse samples had 100% of the verified variants detected.

A second SNV algorithm, Shimmer, was also used for analyzing the leukemic and normal WES datasets, and the results were compared to the list of verified variants. If the number of reads containing a non-reference allele is greater than a minimum threshold, a Fisher's exact test is performed to test the null hypothesis that the variant allele is distributed randomly between the tumor and control sample.

As thousands and millions of sites are tested for a patient's leukemic and control genome, Shimmer performs a multiple testing correction on the Fisher exact P-values and reports alleles that have a false discovery rate below a desired maximum q. This FDR provides a conservative estimate of the proportion of predicted variants that are not true somatic variants, but are instead the consequence of random variation for which the normal sample has a predicted genotype of homozygous reference (Cantarel et al.).

Shimmer (Version 5.8.8) detected 197 of the 254 verified variants (78%). The average detection rate for the diagnosis sample was 78%, and for the relapse sample it was ~80% (Figure 4.4). For five of the diagnosis samples 100% of the verified variants were detected, and for 4 of the relapse samples 100% of the verified variants were detected.

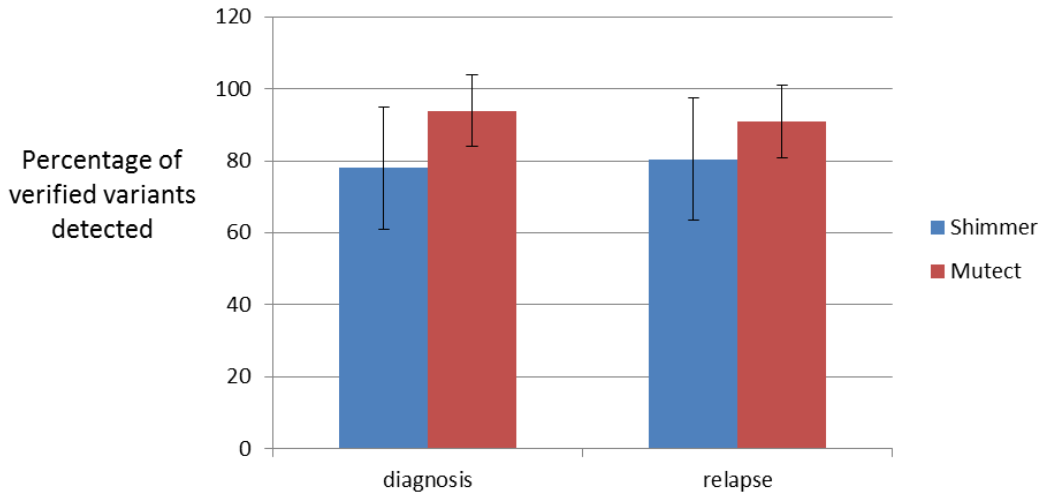

Figure 4.4. Comparison between Mutect and Shimmer for single nucleotide variant detection in 19 whole exome sequencing from patients with acute myeloid leukemia. The x-axis represents the sample type (diagnosis or relapse), and the y-axis represents the percentage of verified variants detected. Blue = Shimmer results. Red = Mutect results.

Mutect detected 32 verified variants that were not reported by Shimmer, whereas Shimmer only detected 1 verified variant that was not reported by Mutect. Therefore, for the downstream analysis only the variants reported by Mutect were used at this time. However, in the future an algorithm such as Baysic could be used to incorporate multiple SNV algorithms (Cantarel et al., 2014). Baysic uses a Bayesian statistical method based on latent class analysis to combine variant sets produced by different bioinformatics packages into a high confidence set of genomic variants (Cantarel et al., 2014).

The medium number of SNVs detected by Mutect for the diagnosis and relapse samples is 300 and 326, respectively (Figure 4.5). A limitation to the current analysis is the determination of false positives from the results as additional lab bench techniques are needed. The long term goal of this project is to continue working with our clinical collaborators on validating high quality novel targets. It is difficult to compare the numbers to other studies as pediatric cancers are thought to have a lower frequency of SNVs compared to adults. Recently, St Jude's published the results from MLL-R ALL study, which reported on average > 100 SNVs per pediatric patient (Andersson et al., 2015).

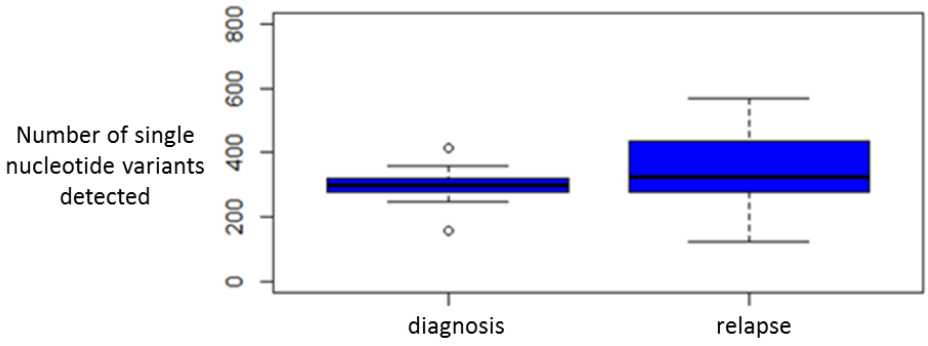

Figure 4.5. Distribution of number of single nucleotide variants detected at the diagnosis and relapse state. The x-axis represents the sample type (diagnosis or relapse), and the y-axis represents the number of single nucleotide variants detected.

For each of the SNVs detected, an allelic proportion (# of reads support alternative allele / total # of reads) was calculated. For high quality germline mutations the allelic proportion is typically centered around 0.5 for a diploid

heterozygote. However, for cancer / somatic mutations the allelic proportion is not centered around 0.5 due to the heterogeneity of the sample and contamination of healthy cells. For example, originally in pancreatic ductal adenocarcinoma (PDAC) it was thought that ~70-80% of patients have a mutation in *KRAS* (Lennerz & Stenzinger, 2015). However, a new study suggests the frequency is closer to 93% when applying a broader range for allelic proportions (Lennerz & Stenzinger, 2015).

Allelic ratios can be used to help guide treatment strategies for pediatric AML, as seen with the FLT3 / ITD allelic-ratio. For the pediatric AML samples analyzed, the majority of the SNVs for any given individual had an allelic proportion of 0.1 to 0.2 (Figure 4.6). There were some somatic alleles greater than 0.5, although these were the minority and a significant gene candidate / pathway was not identified by focusing on these somatic alterations.

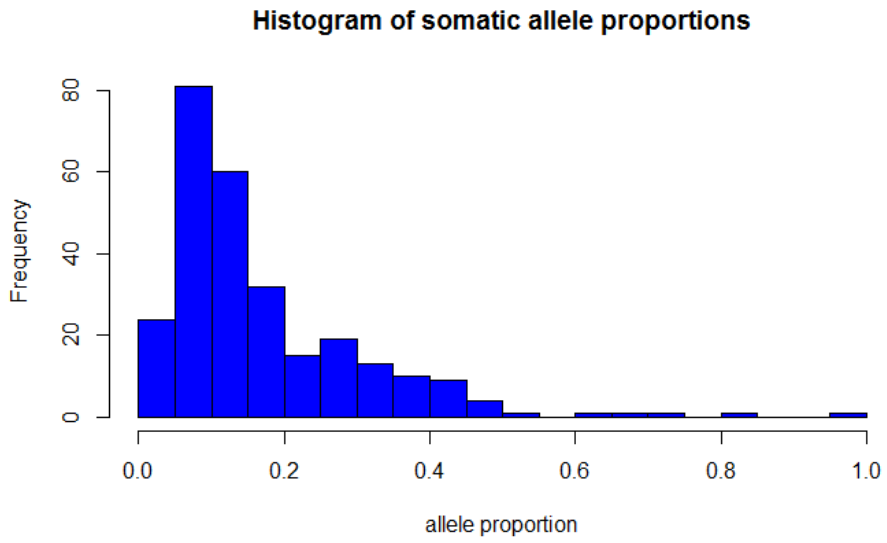

Figure 4.6. Example of pediatric AML distribution of somatic allele proportions calculated from whole exome sequencing. The x-axis represents the allele proportion, and the y-axis represents the frequency at which an allele proportion is measured.

The somatic variants detected by Mutect were annotated with SnpEff (Version 3.3a) using the package GRCh37.75. Using SnpEff annotated transcript ID, variants in the VCFs were mapped to UniProt Accession Numbers and Gene Ontology information using a custom script. Protein-protein interactions for the protein coding genes were determined using reactomeFI and/or the STRING API (Jensen et al., 2009), a database of known and predicted protein interactions derived from: genomic context, high-throughput experiments, co-expression, and previous knowledge. The annotated VCFs were used in downstream analysis. Of interest are also the germline mutations detected by Mutect, but these variants require a different type of analysis compared to the somatic changes and were not the main focus of this project.

4.2.1 Prioritization and Visualization of SNVs

A risk-adapted stratification system is a relatively new concept that has been implemented to help with the diagnosis and treatment of pediatric AML and is based on a finer resolution of the genetic alterations present. In particular, COG has established classifications for risk-directed therapy based on cytogenetic, molecular, and minimal residual disease (MRD) information (Pui, Carroll, Meshinchi, & Arceci, 2011). There are numerous types of genetic alterations used in this risk stratification system, such as SNVs in the activation loop of FLT3 and translocation events in core-binding and transcription factors (Testa & Pelosi, 2013).

Unraveling the functional consequences of potential somatic alterations associated with pediatric AML is a complex process. A strategy for ranking the somatic variants detected from Mutect was created by using a scoring/ filtering system that consisted of the following characteristics: confidence of genomic variant, gene ontology annotations, effect of variant, ensemble biotype, protein-protein interactions, and pathway and gene ontology enrichment analysis.

To begin with, an interactome was first created using the verified variants provided by the COG, including linker genes. The interactome was considered to be the base knowledge of the project, and the goal of this project was to further expand upon these findings. Approximately 180 genes were up-loaded into cytoscape using reactome FI app (Wu, Feng, & Stein, 2010), and further clustered based on connectivity and sub-networks (Figure 4.7).

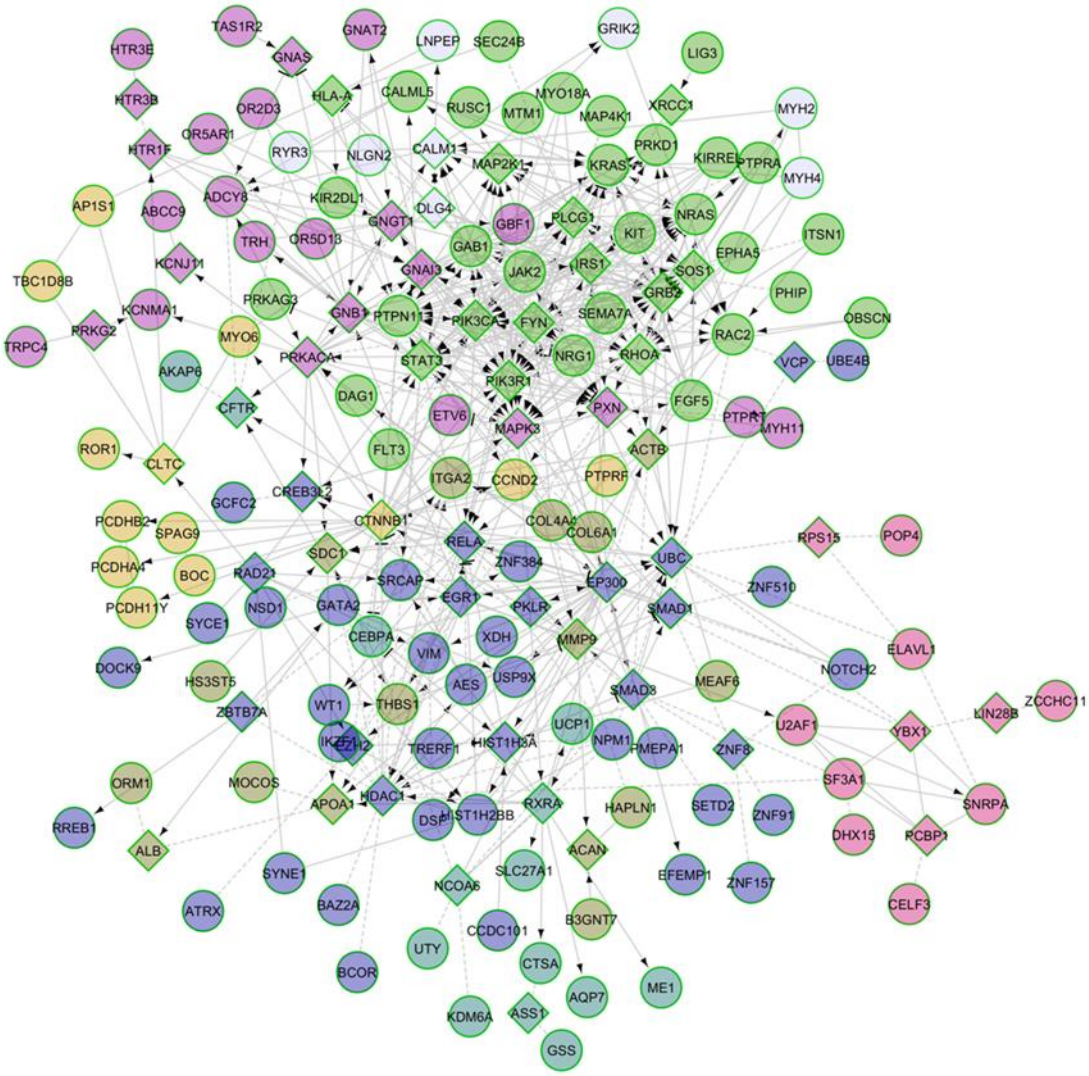

Figure 4.7. Interactome based on verified single nucleotide variants from 19 pediatric acute myeloid leukemia patients. Nodes are colored based on connectivity. Circle = gene with somatic mutation. Triangle = linker gene. Number of modules = 8.

Reactome FI top ranking sub-networks, included mast/stem cell growth factor Kit signaling, beta 1 integrin cell surface interactions, peroxisome proliferator-activated receptor (PPAR) signaling pathway, spliceosome, and tyrosine-protein

phosphatase non-receptor type II (SHP2) signaling (Figure 4.7). Mast/stem cell growth factor Kit signaling is a known pathway involved in pediatric AML, although patients with a mutation in *KIT* are rare (~4%) there are several genes with mutations that are connected to *KIT* through protein-protein interactions (Figure 4.7) and cell signaling cascades with KIT.

Mast/stem cell growth factor receptor Kit (which is the translated protein from the *KIT* gene) is a tyrosine-protein kinase on the cell surface of hematopoietic stem cells. It is the receptor for cytokine KITLG /SCF and is involved in many cellular functions such as the regulation of cell survival and proliferation, hematopoiesis, stem cell maintenance, gametogenesis, and mast cell development (Lennartsson & Ronnstrand, 2012). Upon binding of KITLG / SCF, the receptor dimerizes and activates auto-phosphorylation on tyrosine residues leading to signaling cascades involved in many pathways. The activity of KIT is naturally down-regulated by PRKCA-mediated phosphorylation of serine residues. Gleevec (imatinib), a compound produced by Novartis, also inhibits KIT signaling and is used to treat various types of cancers (Shu & Yang, 2012).

ClusterONE, a method for detecting potentially overlapping protein complexes from protein-protein interaction data (Nepusz, Yu, & Paccanaro, 2012), analysis of the interactome fully supported that candidates such as KIT, NRAS, MAPK are highly connected nodes within the pediatric AML interactome. The ClusterONE algorithm is based on the concept of a cohesiveness score and uses a greedy growth process to find groups in a protein-protein interaction (PPI) network that are likely to correspond to protein complexes. Proteins may have multiple functions, and therefore the corresponding nodes may belong with more than one cluster.

There were 2 main clusters from the analysis that differed by only a few nodes suggesting that the shared nodes have the potential to have different biological functions based on PPIs: Cluster 1= 31 nodes, density = 0.449, Quality = 0.637, p-value = 1.205e-7; Cluster 2 = 26 nodes, density = 0.502, Quality = 0.570, p-value = 1.26e-5 (Figure 4.8). The majority of the nodes was shared between the two clusters, and included key oncogenes such as KIT, NRAS, JAK2, FLT3, and PTPN11. There were 3 unique candidates in cluster 1, CALML5, ITSN1, and FLT3, and there were also 3 unique candidates in cluster 2, MAP4K1, DAG1, MTM1.

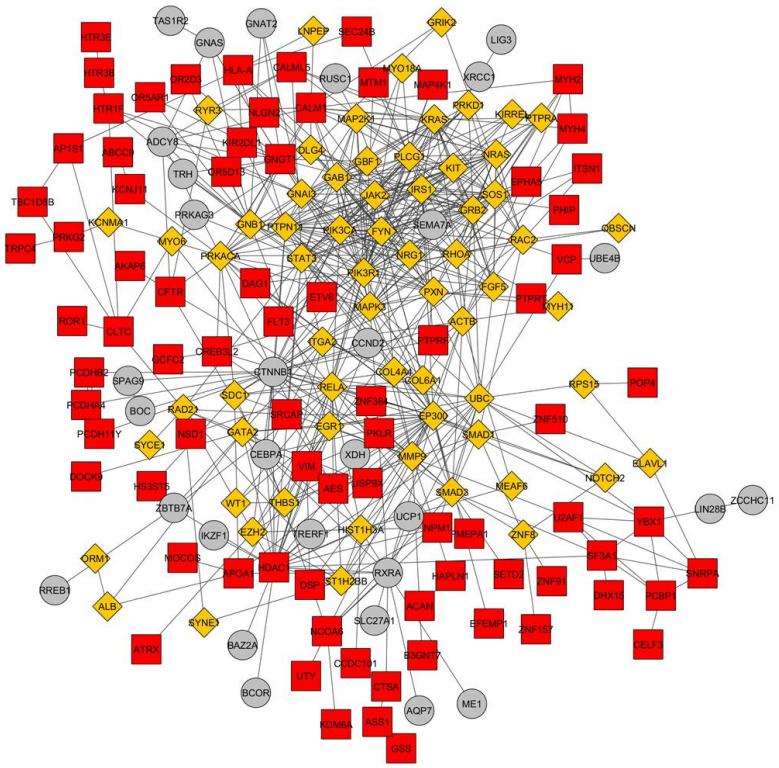

Figure 4.8. ClusterOne analysis indicates key oncogenes including *KIT*, *NRAS*, *KRAS*, *GATA2* (yellow triangles) are significantly connected with many of the mis-regulated subnetworks of pediatric acute myeloid leukemia interactome. Grey circle = outlier. Red square = clustered. Orange triangle = highly connected.

To further enhance the data analysis, a prioritization system was created for the SNVs detected with the bioinformatics pipeline. Table 4.1 highlights the 5 major categories and tools for prioritizing SNVs. It is important to note that Mutect detects both germline and somatic mutations, and the majority of the work presented below focuses on somatic mutations. A confidence score of somatic and pass as annotated by Mutect were required for any variant to be ranked for this part of the dissertation project (Table 4.1). The biotype for the variant also needed to be protein coding (Table 4.1); however, variants located in other regions of the genome are maintained in the VCF and are under-consideration for future work that focuses on integrating genomic variant data other omics data, such as miRNA.

Table 4.1. Characteristics of prioritizing SNVs.

Prioritization	Tool	Value	Score
Confidence	Variant Detection Algorithm	Somatic; PASS	Necessary
Biotype	SnpEff	protein coding	Necessary
Effect Variant	SnpEff	Alters protein sequence or splice site	+
Protein Protein Interactions	String	PPI with known cancer gene	+
Gene Ontology	Custom	GO matches Bingo Enrichment Analysis	+

The gene ontology score/filtering is based on a custom analysis. First, a list of known pediatric AML cancer genes were analyzed with BiNGO, a gene ontology enrichment tool for Cystoscape (Maere, Heymans, & Kuiper, 2005). There is flexibility in the pipeline to use different gene lists at this step. Using a list of ~50 pediatric AML oncogenes, two hundred and eighty gene ontologies with a significant p-value were extracted, and used to prioritize somatic variants (Figure 4.9). A genomic variant that was located within a gene, that was also annotated with a gene

ontology (biological process) enriched from the BiNGO analysis, was given a positive score for this portion of the prioritization. The top biological processes enriched were: protein amino acid phosphorylation, transmembrane receptor protein tyrosine kinase signaling, phosphate metabolic process, phosphorylation, negative regulation of programmed cell death, and regulation of cell proliferation (Figure 4.9).

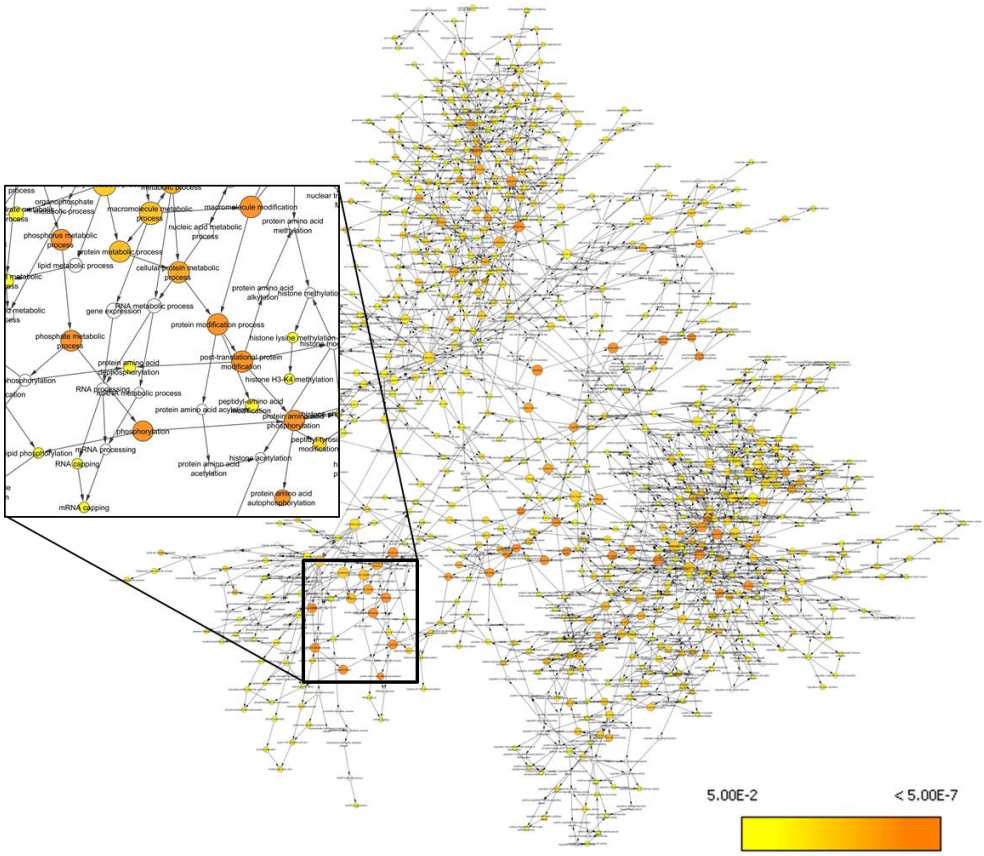

Figure 4.9. Biological process enrichment analysis for pediatric acute myeloid leukemia oncogenes using BiNGO in cytoscape. Gene ontology structures are hierarchical and graphical display of connections helps to highlight enriched terms (dark orange). Highlighted box is zoomed in to illustrate complex relationships between gene ontologies.

The prioritization module based on the effect of the genomic variant leverages SnpEff's annotation (Cingolani et al., 2012). Table 4.2 summarizes the SnpEff annotations, their impact, and the score given to corresponding variant, with the highest score given to variants that disrupt splice site acceptor / donor sites, cause a pre-mature stop codon, or cause the deletion of an exon. Moderate variants are classified as causing non-synonymous coding or a disruption of the 5` / 3` UTR, whereas low impact variants are located in introns or are synonymous coding events (Table 4.2).

Table 4.2. Prioritization SNV effect and impact

Impact	Annotation	Score
Low or Modifier	splice site region, downstream, synonymous coding intron, 3' UTR, and 5' UTR, upstream	0.25
Moderate	codon change/insertion/deletion, non-synonymous coding, splice site branch, 3' UTR deleted, 5' UTR deleted	0.5
High	exon deleted, frame shift, rare amino acid, splice site acceptor, splice site donor, stop lost, start gain/loss, stop gain/loss	1

The protein-protein interaction (PPI) score is based on the pipeline's annotation of PPIs that is derived from using the STRING API (Jensen et al., 2009), a database of known and predicted protein interactions derived from: genomic context, high-throughput experiments, co-expression, and previous knowledge. SNVs that were located in an annotated gene region and mapped to a Uniprot AC number were analyzed via STRING, and PPI information was generated for each gene candidate that had a SNV. If one of the predicted PPIs was with a known pediatric AML oncogene, the candidate /gene (variant) was given a positive score for this category.

There is flexibility in the methodology to alter the known gene list to aid in the broader application of the pipeline.

The filtering strategy leveraged the gene ontology analysis, PPIs analysis, and the effect / impact of the genomic variant to help prioritize newly identified SNVs. Collectively, the pipeline identified and ranked new variants that were not previously provided to us from the COG (Table 4.3). For example, the pipeline identified 43 new somatic variants for patient's PATABB diagnosis sample, including a mutation in transforming protein N-Ras (*NRAS*), a known oncogene in pediatric AML.

Table 4.3. Summary of ranked single nucleotide variants identified in pediatric acute myeloid leukemia

ID	Sample	Ranked Variants
PATABB	diagnosis	46
	relapse	55
PASFEW	diagnosis	30
	relapse	39
PARVUA	diagnosis	47
	relapse	42
PARUNX	diagnosis	100
	relapse	44
PARUCB	diagnosis	71
	relapse	50
PARIEG	diagnosis	68
	relapse	63
PARGVC	diagnosis	17
	relapse	53
PANVGP	diagnosis	67
	relapse	84
PANLIZ	diagnosis	85
	relapse	83
PANGTF	diagnosis	50
	relapse	59
PAMYAS	diagnosis	52
	relapse	79
PAERAH	diagnosis	56
	relapse	31

The pipeline developed also detected a SNV (rs121913529) in GTPase KRas (*KRAS*) for patient PATABB. The mutation causes a translational shift, G12D, and is predicted to be pathogenic. Interestingly, ClinVar reports pathogenic G12A in non-small cell lung cancer, G12V in juvenile myelomonocytic leukemia, and G12D in juvenile myelomonocytic (http://www.ncbi.nlm.nih.gov/clinvar/?term=rs121913529).

KRAS was connected to *NRAS* in the analysis through several gene ontologies, positive regulation of cell proliferation, ras protein signal transduction, and activation of MAPKK activity. *KRAS* and *NRAS* were also connected to another gene with a somatic mutation, nucleus accumbens-associated protein 2 (*NACC2*), through the gene ontology positive regulation of cell proliferation. *NACC2* is repressor of gene transcription, and has not been associated with pediatric AML yet.

On average the pipeline developed reported ~20-40 additional SNVs per patient compared to the list provided by COG; however, these variants require extensive analysis prior to being considered valid as somatic hits. It is always important to check the alignment files and cross reference with other data sources. Furthermore, allelic proportions are an important consideration, along with frequency in the general population. The interactome for the diagnosis state was analyzed by using the SNVs ranked with the pipeline developed. The ranked genes were uploaded into cytoscape and reactomeFI was used to clustered the genes based on connectivity and sub-networks (Figure 4.10).

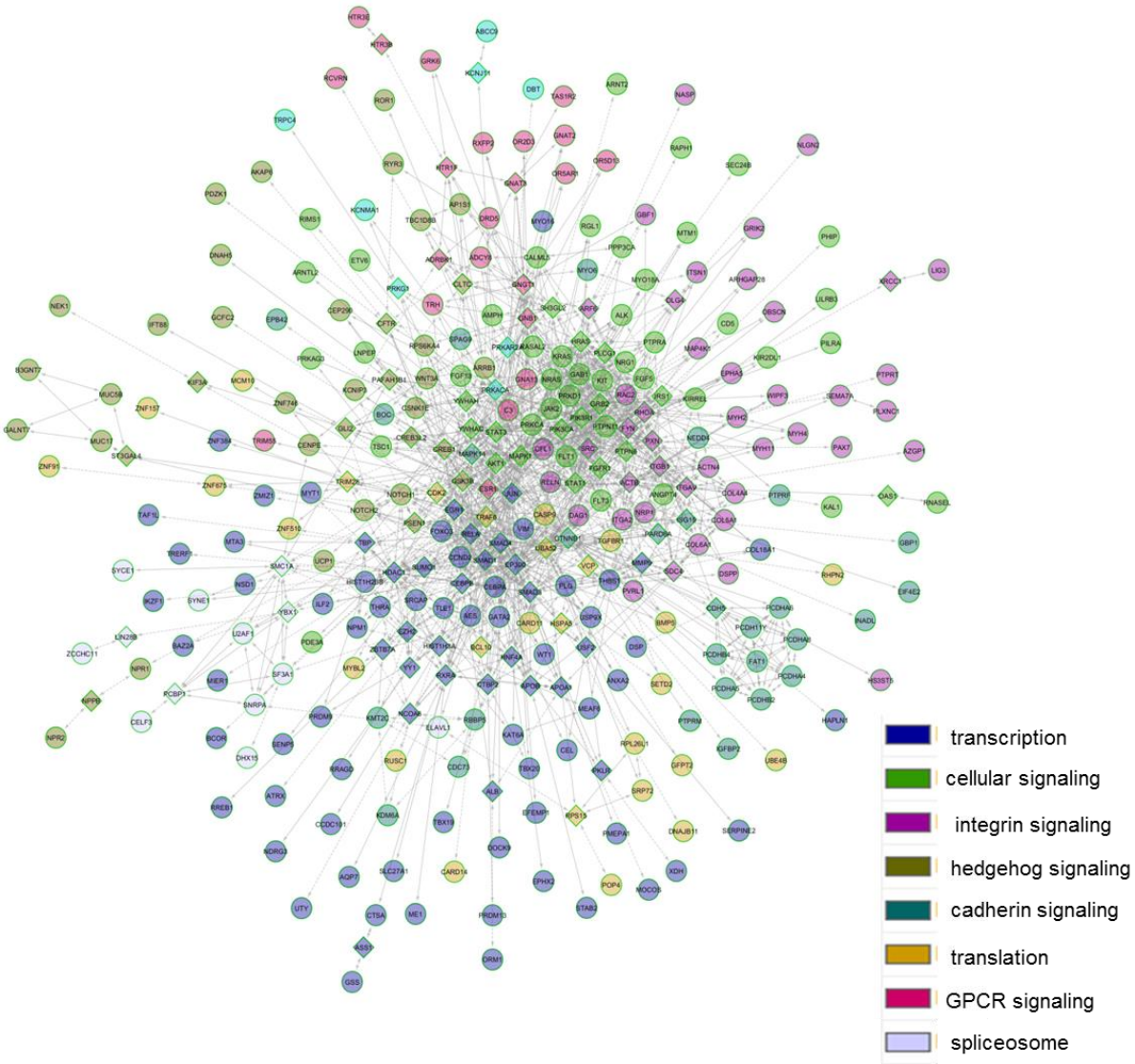

Figure 4.10. Interactome based on single nucleotide variants from 20 pediatric AML patients. Circle = gene with somatic mutation. Triangle = linker gene. Number of modules = 8. Nodes are colored based on connectivity and further annotated with function.

Eight modules were clustered and based on the subnetwork pathway analysis were divided into the following categories: transcription, cellular signaling, integrin signaling, hedgehog signaling, cadherin signaling, translation, GPCR signaling, and

splicesome (Figure 4.10).A subnetwork that was enriched in the new interactome was cadherin signaling (Figure 4.11), which plays a role in calcium-ion-dependent adhesion. Cadherins are involved in many biological processes such as development, neurogenesis, cell adhesion, and inflammation. It was noted that 1 patient had several SNVs with low allelic proportions in several of the genes clustered in the cadherin subnetwork, supporting that an alternative data analysis, such as the one presented, maybe required to help reveal important genomic alterations that are not at a high frequency.

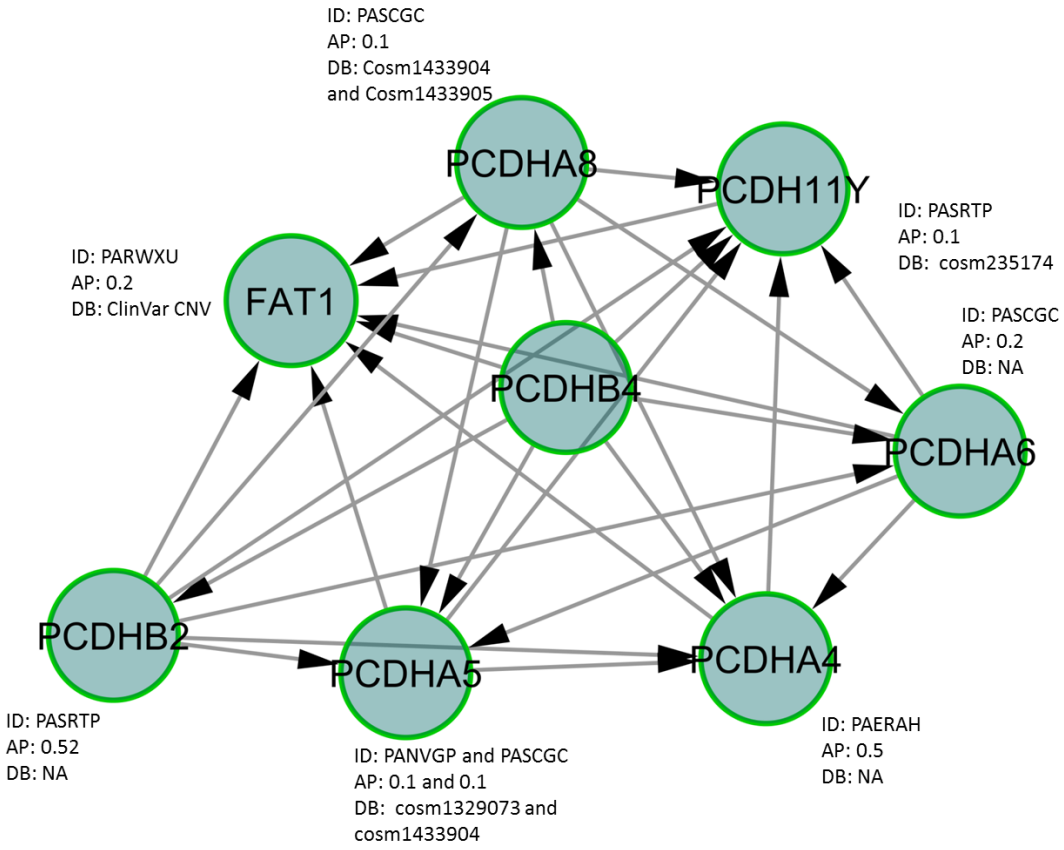

Figure 4.11. Cadherin subnetwork clustered within the pediatric acute myeloid leukemia interaction. ID = patient identifier. AP = allelic proportion. DB = annotation in databases. NA= not applicable.

For the relapse samples analyzed several of the known mutations were detected in therapeutically relevant genes such as *NSD1* (cosm235174), *TET2*, and *FIGN* (A369D). The interactome was re-generated with the new SNVs at the relapse state, and then compared with the original analysis. One of the top enriched subnetworks (Figure 4.12) was signaling in platelet derived growth factor (PDGF) (p-value < 0.05, FDR 1.0e-4, reactomeFI analysis).

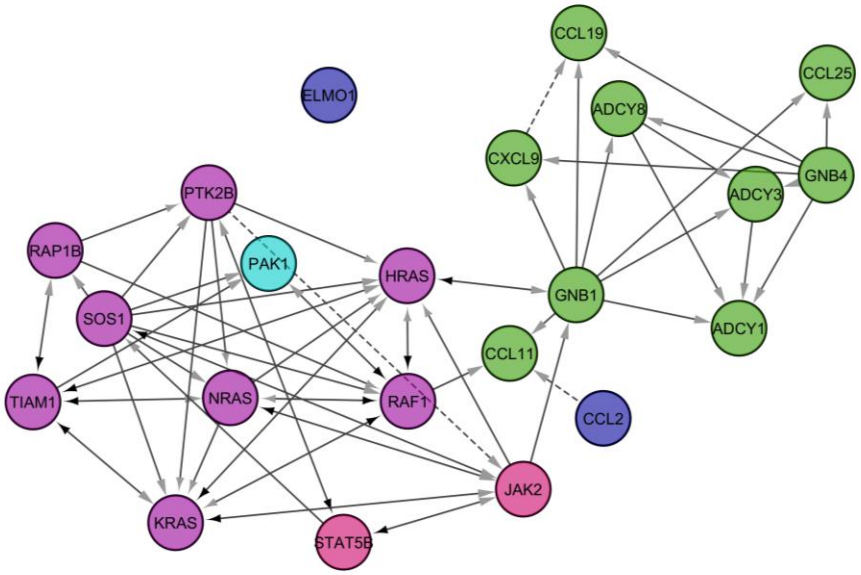

Figure 4.12. Reactome analysis for genes with single nucleotide variants that were clustered in the platelet derived growth factor signaling subnetwork.

Furthermore, in the relapse state a SNV was detected in adenylate cyclase type 8 (*ADCY8*), and predicted to cause a translational shift of R362H. ADCY8 encodes for protein that is membrane-bound and catalyzes the formation of cyclic AMP from ATP. A mutation in tyrosine-protein kinase JAK2 (*JAK2*) was also detected, and

causes a translational shift from V561F. JAK2 is a non-receptor tyrosine kinase involved in cell growth, development, and differentiation. The V561F mutation has been reported previously in myeloproliferative disorders (McLornan, Percy, & McMullin, 2006)

A SNV was also detected in nuclear pore complex protein Nup-98 (*NUP98*) that is predicted to cause a pre-mature stop codon at amino acid position 56. *NUP98* plays an important role in the nuclear pore complex assembly and maintenance. Cryptic translocations involving *NUP98* have been reported in pediatric AML (Ostronoff et al., 2014); however, this is the first stop-gained mutation to have been detected in pediatric AML.

Interestingly, only a few SNVs were detected in histone methyltransferases, however, histone acetyltransferase p300 (*EP300*) is one of the most connected proteins within the interactome generated from the relapse specific genes (Figure 4.13). *EP300* encodes for a protein that functions as a histone acetyltransferase and regulates transcription via chromatin remodeling, and has been associated with other cancers. There were several subnetworks within the interactome (Figure 4.13) that contained EP300 including factors involved in megakaryocyte development and platelet production (FDR $3.46*10^{-3}$), retinoic acid recpetors-mediated signaling (FDR $1.117*10^{-1}$), and regulation of nuclear beta catenin signaling and target gene transcription (FDR $1.31 *10^{-1}$).

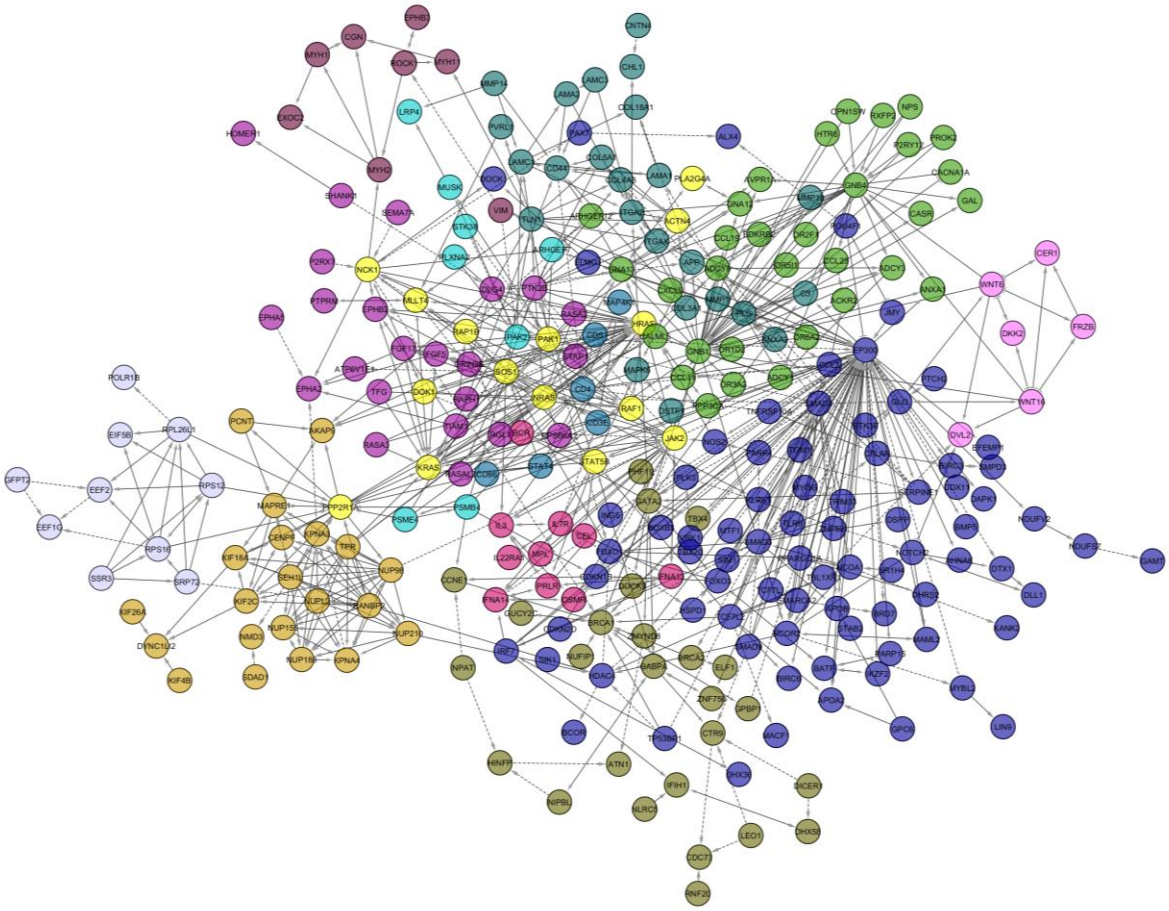

Figure 4.13. Interactome of proteins with single nucleotide variants at the relapse state in pediatric acute myeloid leukemia. Genes with a single nucleotide variant at the relapse state were up-loaded into cytoscape and analyzed with reactomeFI. Nodes are colored based on connectivity.

Furthermore, several of the patients had mutations in genes that encode for proteins involved in retinoic acid signaling, such as PRAMEF1 and PRAMEF13. Retinoids (vitamin A) can induce the differentiation of various types of stem cell (Connolly, Nguyen, & Sukumar, 2013). Retinoids have been approved by the FDA for use in the treatment for some solid tumor cancers, such as head and neck cancers,

and the ability of all-trans retinoic acid (ATRA) to initiate differentiation of promyelocytic leukemic cells to granulocytes is a promising indication of the use of retinoid based therapies in cancer (Connolly et al., 2013).

4.3 FLT3/ITD Detection and InDel Analysis

Recent studies have highlighted that genomic differences between humans arise more from structural variants (SV) compared to SNPs (Alkan, Coe, & Eichler, 2011). SVs were originally defined as genetic alterations in DNA that are ~1 kb or larger. However, advancements in DNA sequencing technology have caused a shift in this paradigm with several scientists now classifying SV as >50bp in length, and variants that are <50bp are classified as small insertions and deletions, commonly referred to as InDels (Alkan et al., 2011). There are numerous classes of SVs, such as copy number variations, large insertions / deletions, tandem duplications, inversions, and chromosomal translocations.

One of the most commonly studied SV in pediatric AML is an internal tandem duplication (ITD) located in the fms-related tyrosine kinase 3 gene (*FLT3*). The ITD is an in-frame insertion in the coding region of *FLT3* that causes ligand-independent activation of the FLT3 receptor. Interestingly, scientists have also discovered a copy-neutral loss of heterozygosity in pediatric patients with FLT3/ITD (Stirewalt, Pogosova-Agadjanyan, Tsuchiya, Joaquin, & Meshinchi, 2014). The aim of this portion of the dissertation project is to develop bioinformatics methodologies for detecting and analyzing insertions and deletions, referred to collectively as InDels, using genomic NGS data generated from pediatric AML patients.

Activating mutations in *FLT3* are the most common somatic mutations found in AML (Soheil Meshinchi et al., 2006). *FLT3* is a member of the class III tyrosine

kinase receptor family, and is expressed on hematopoietic stem and progenitor cells. Experimental evidence suggests it is essential for stem cell development and differentiation (Hayakawa et al., 2000) through a ligand dependent activation process (Soheil Meshinchi & Appelbaum, 2009). The FLT3 ligand causes dimerization of the receptor which leads to auto-phosphorylation. Once FLT3 is phosphorylated it can activate several signaling pathways that ultimately lead to cell proliferation and differentiation (Figure 4.14; generated with ProteinLounge and template from SABiosciences).

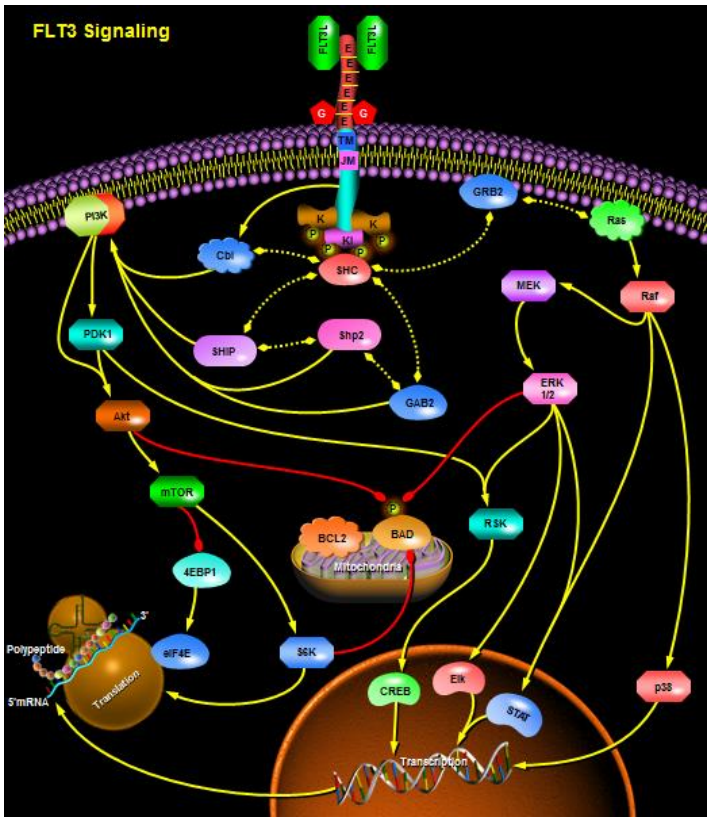

Figure 4.14. FMS-like tyrosine kinase 3 receptor signaling pathway generated with ProteinLounge and pathway template is from SABiosciences. FLT3 activation induces cellular signaling events involved in transcription and translation.

There are two major types of somatic mutations that occur in the *FLT3* gene in pediatric AML cases. SNVs causing a missense in the activation loop of the translated *FLT3* gene occur in about 7% of pediatric AML cases, and an internal duplication (ITD) in the juxtamembrane domain occur in about 15% of pediatric AML cases (Soheil Meshinchi et al., 2006). The FLT3/ITD is an in-frame insertion in exon 14 or 15 that changes the amino acid sequence in the juxtamembrane domain (Figure 4.15).

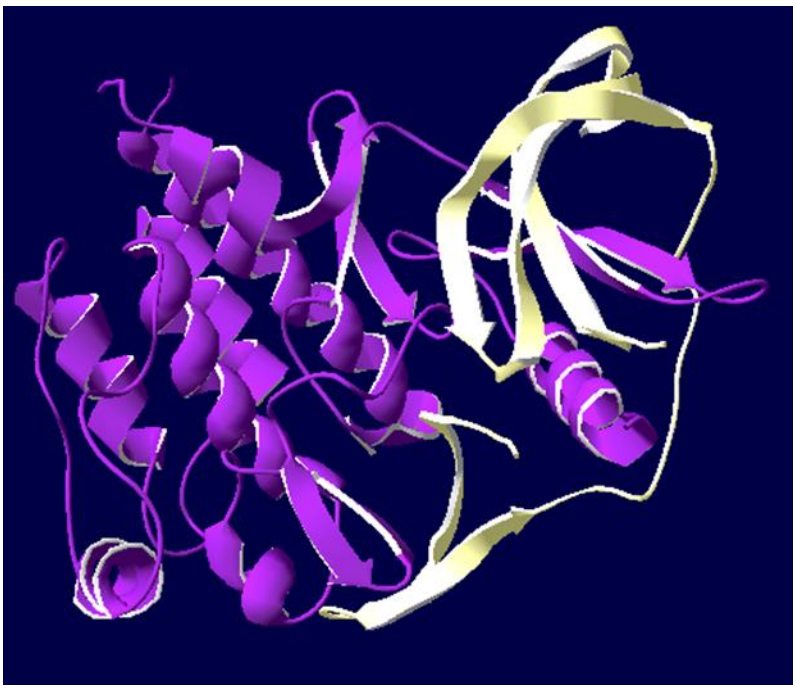

Figure 4.15. Crystal structure FMS-like tyrosine kinase 3 receptor. PDB 1RJB file was downloaded and analyzed via Swiss PDViewer. Yellow = translated exons 14 and 15. Purple = all other translated exons.

The presence of an in-frame ITD in the FLT3 protein leads to ligand-independent auto-phosphorylation of FLT3 and constitutive activation of the cell (Soheil Meshinchi & Appelbaum, 2009). A group of scientists transformed 2 different

hematopoietic progenitor cell lines with a FLT3/ITD mutant gene, which caused autonomous cell growth and activation of the MAP kinase and STAT5 pathways (Hayakawa et al., 2000). Furthermore, increased activation of MAP kinase and STAT5 have been reported for clinical AML blasts with the FLT3/ITD, and the ITD is found almost exclusively in AML (Zwaan et al., 2003).

In the clinical setting the presence of the FLT3/ITD mutations has been associated with a statistically increased risk for relapse compared to FLT3/ITD negative patients (Thiede, 2002). It is hypothesized that this risk might be due to an enhanced regrowth potential of residual diseased cells, leading to an enhanced relapse state (Zwaan et al., 2003). Furthermore, varying allelic ratios between the mutant *FLT3* and wild type seem to carry prognostic significance (S Meshinchi et al., 2001). Clinicians have tried to establish a reliable methodology to determine the allelic ratio between the ITD mutant *FLT* and wild type, which is often referred to as ITD allelic ratio or ITD-AR.

ITD-AR is calculated using Applied Biosystems GeneScan Analysis Software using a PCR based method (Thiede, 2002). The ITD-AR is calculated by dividing the peak height of the ITD product by that of the normal WT product (Soheil Meshinchi et al., 2006). Additional testing is often required to further analyze the sample for other potential genomic mutations. NGS offers an opportunity to analyze a single sample for multiple genomic alterations at once.

Detecting an ITD is a difficult task and requires special algorithms. Furthermore, a standard methodology for calculating allelic rations using NGS data is not established because read depth per allele is not necessarily proportional to the allelic ratio. Six FLT3/ITD positive pediatric AML samples, with varying FLT3/ITD

allelic ratios, were analyzed using the Pindel algorithm (Ye et al., 2009). Pindel was selected as the preferred algorithm because recently Spencer et al. (Spencer et al., 2013) compared several algorithms for the detection of FLT3/ITD and published that Pindel (Ye et al., 2009), a pattern growth approach, was the superior algorithm. The output files were converted to VCF files using the pindel2vcf script provided with the Pindel package. Only insertions located in exon 14 or 15 in the FLT3 gene were analyzed as potential ITDs.

For the 6 patients analyzed, 3 samples per patient, Pindel detected an insert in 5 of the 6 patient's diagnosis sample (Table 4.4). Pindel also detected an insert in 4 of the relapse samples and 1 of the remission samples (Table 4.4). A benefit to using Pindel is that it provides a better resolution of the genomic abnormality by providing a genomic position, sequence, and estimated length of the insert, which are collectively not available with the PCR electrophoresis assay. The pipeline used to report genomic variants to the COG was not able to detect and report the ITD-AR for these NGS libraries, supporting that the pipeline developed provides novel capabilities.

Table 4.4. Summary FLT3 / ITD detection using Pindel.

ID	Sample	Position	Sequence	Length
Patient 1	Relpase	-	None Detected	-
	Diagnosis	-	None Detected	-
	Remission	-	None Detected	-
Patient 2	Diagnosis	28,608,235	TCTTGGAAACTCCCATTTGAGATCATATTCA	31
	Relapse	-	None Detected	-
	Remission	-	None Detected	-
Patient 3	Diagnosis	28,608,249	ATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGCC	40
	Relapse	28,608,265	ATATTCTCTGAAATCTCCACGGGGG	25
	Remission	-	None Detected	-
Patient 4	Diagnosis	28,608,214	CTTACCAAACTCTAAATTTTCTCTTGGAAACTCCCAT	37
	Relapse	28,608,214	CTTACCAAACTCTAAATTTTCTCTTGGAAACTCCCAT	37
	Remission	-	None Detected	-
Patient 5	Diagnosis	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
	Relapse	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
	Remission	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
Patient 6	Diagnosis	28,608,243	ACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCAC	73
	Relapse	28,608,243	ACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCAC	73
	Remission	-	None Detected	-

A few research groups have tried to establish a method for calculating the ITD-AR using NGS technology. In 2012 Thol et al. used serial dilutions of NGS samples (GS FLX) from AML patients with a known FLT / ITD to help calculate the ITD-AR (Thol et al., 2012). Spencer et al. created two methodologies for calculating ITD-AR using NGS data; one method uses the results directly from Pindel, while the other method uses a *de novo* assembly approach (Spencer et al., 2013). For the direct Pindel method, the ITD-AR was calculated by dividing the number of supporting reads by the coverage of unique reads. For the *de novo* assembly technique, reads that mapped to the juxtamembrane domain were extracted and assembled using Phrap. The mean coverage depth of the inserted sequence in the assembled contig was divided by total coverage of FLT3 exons 14 and 15 (Spencer et al., 2013).

For the dataset analyzed there are several limitations for accurately calculating the ITD-AR. There were no serial dilutions available and the sequencing was WES, versus a more targeted approach as described by Spencer et al (Spencer et al., 2013).

The targeted approach has significantly higher read depth coverage compared to the WES dataset under analysis. It would be beneficial to have a more targeted approach than WES for using NGS in pediatric AML to accurately access the feasibility of using NGS to calculate ITD-AR and other co-occurring mutations.

The key in designing a targeted NGS kit for pediatric AML is to first establish the appropriate sequencing depth required to accurately detect somatic mutations, such as FLT3-ITD, and use library size and read length to determine the total length of the genome that can be captured and sequenced at the desired sequencing depth. Using the calculated total length of the genome to be captured, a limit number of regions can be established by prioritizing known and potential AML cancer genes.

InDels are challenging to detect from NGS data, and their analysis is further complicated as methods for annotating somatic versus germline, and association with disease onset and progression, are still being developed. The goal of this research project was to determine if InDels, beyond the *FLT3* gene, are prevalent in pediatric AML. A necessary step in this analysis is the ability to distinguish between InDels present at the cancer state versus the remission state. Therefore, we decided to leverage a probability calculation using the read depths, and a delta of allelic balances, at the various time-points to further prioritize and understand InDels. The diagnosis, relapse, and remission samples on average had three thousand insertions and 8,000 deletions (Figure 4.16)

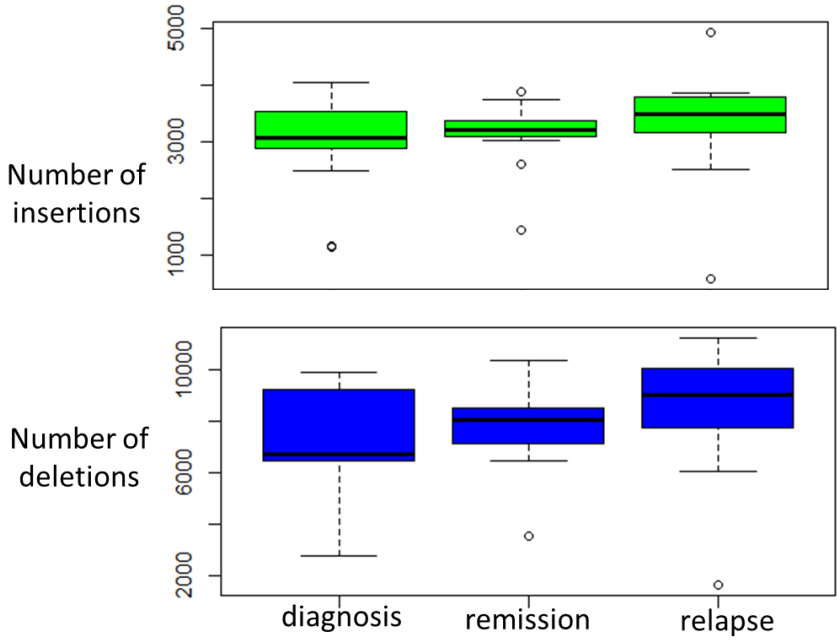

Figure 4.16. Summary insertions (green top panel) and deletions (blue bottom panel) at diagnosis, remission, and relapse state.

The majority of the InDels detected were between 5-20 nucleotides in length per sample. The sequence composition ranged for the InDels, and in general is an important aspect for filtering InDels, as well as SNVs. Recent studies have demonstrated biases in Illumina sequencing, including non-random distribution of the reads sequenced compared to reference genome and non-random distribution of errors, with the majority of the sequencing errors being single point substitution (Minoche, Dohm, & Himmelbauer, 2011). InDels errors occur at a much lower rate compared to substitutions, but increase in regions of homopolymers (Minoche et al., 2011). A filtering strategy was implemented using homopolymer information to help bin InDels that could be related to sequencing artifacts.

There are different ways of utilizing the Pindel algorithm for calling InDels. The samples can be run simultaneously or independently. For this dissertation both options were tested, and InDels not located in homopolymer regions were selected for further downstream analysis. The next step was to calculate the difference between the allelic proportion of the InDels by subtracting the allelic proportion at the cancerous state from the remission state. After applying the filtering strategy, a few InDels were noted in the following genes such as *CDK11B*, *PAPSS1*, *PPA2*, *PDGFRA*, *ORC6*, *POSTN*, and *KRT4*. Upon examining these InDels in the UCSC Genome Browser it was noted that the majority of them were located in regions of clinically significant copy number variations and/or a region were SNVs were reported.

There are limitations to analyzing SV data from NGS. GC-content bias is known to skew average read depth of a genomic region with high or low GC-content by reducing the lower mean read depth compared with bins with medium GC-content (Liu et al., 2013). Also poor capture of an exon probe can influence the ability of an algorithm to accurately detect CNVs. Other issues, like mapability of NGS reads from diverse tumor genomes, contamination with healthy cells, and tumor ploidy, can have a major impact on the accurate detection of SVs.

Chapter 5

DISCUSSION

Next generation sequencing (NGS) technologies provide the potential for developing high-throughput and low-cost platforms for clinical diagnostics. A limiting factor to clinical applications of genomic NGS is downstream bioinformatics analysis for data interpretation. Even targeted NGS approaches like WES, generate massive amounts of data that can be difficult for clinicians and biomedical scientists to analyze and apply appropriately in the medical field. Rigorous bioinformatics methodologies are required to analyze the data with appropriate statistical methods that will ultimately link the genetic data to the disease phenotype.

Collectively, this dissertation highlights the development of bioinformatics pipelines for end-to-end clinical NGS data analysis, including custom variant annotation and functional profiling. Two types of disease cohorts were analyzed, and provided the foundation for developing a strong collaborative approach for analyzing clinical genomic variant data. All datasets analyzed started with WES paired-end library files, and the bioinformatics workflows created delivered focused lists of annotated variants to clinical collaborators.

The projects analyzed in this dissertation required a strong and consistent interaction between clinical researchers and bioinformaticians, because a team science approach was required for appropriate data analysis. Working with clinical collaborators was essential for discussing and maintaining context to the massive data set. Insight gained from this project included the realization that creating a start to end

analysis pipeline requires testing, validation, and the combination of many diverse algorithms, as it is essential that the informatics methodologies can differentiate between normal diversity and disease associated diversity.

The analysis of a Mendelian disease typically starts with a family pedigree and genotype (allelic) information from the affected child (proband), mother, father, and /or siblings. The end goal of the analysis is to provide a list of inherited alleles that are biologically relevant to the disease. To aid in the analysis of Mendelian diseases, an informatics pipeline was established to help provide a mechanism for filtering / selecting candidate alleles within a large dataset (Figure 5.1). As with any informatics approach, validation is required by other methods. Validating the biology of a genomic variant requires extensive time and resources.

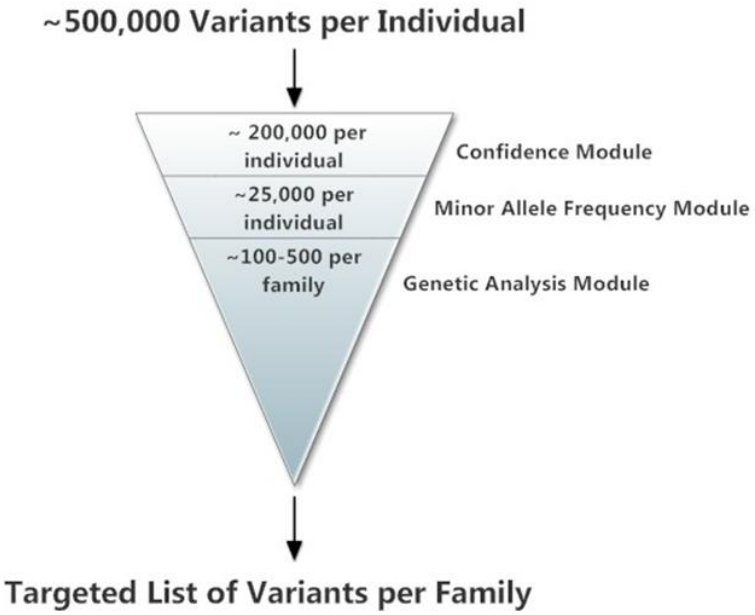

Figure 5.1. Workflow for filtering a large genomic dataset generated from a Mendelian disease cohort

The code and accompanying documentation supporting the Mendelian disease work is being deposited on github (https://github.com/ecrowgey), enabling public access and dissemination. Furthermore, a methods paper is in preparation for releasing the code in a peer-reviewed scientific journal. The manuscript will leverage data from a sickle cell disease cohort, and will focus on enhancing the code to work with larger pedigrees.

Cancer genomic projects, unlike Mendelian diseases, often focus on non-inherited genomic mutations. Often times these studies start with a distinct clinical diagnosis of cancer determined by cytogenetics, white blood cell counts, and diseased phenotype. Genomic sequencing and genotyping is often carried out with either a targeted approach using a panel of known oncogenes, or through an exploratory sequencing strategy such as WES or WGS. Although in theory a clinical diagnostic kit should be a targeted approach to help reduce costs and detection of off-target variants.

Pediatric AML is a challenging disease to generate a targeted list of genomic alterations as the disease is currently considered to be extremely heterogenic between sub-populations of patients in regards to type and location of genomic mutations. Therefore, it is a difficult transition from a general approach such as WES and /or WGS to a targeted diagnostic kit, as different genomic alterations have different biases. For example, the detection of cryptic translocations is more suited for WGS versus a targeted amplicon approach as the breakpoint locations are diverse. Illumina has a few targeted approaches available, including an adult myeloid leukemia NGS kit called TruSight Myeloid Sequencing Panel (http://www.illumina.com/products/trusight-myeloid.html).

The panel of genes comprising the TruSight Myeloid kit was designed by a consortium of experts and focuses on genes frequently mutated in: AML, myelodysplastic syndrome, myeloproliferative neoplasms, chronic myelogenous leukemia, and juvenile myelomonocytic leukemia. Leveraging a collaboration with key members of the Children's Oncology Group, we are in the process of working with Illumina and adding ~50 additional genes to the current TruSight Myeloid sequencing panel. The goal is to broaden the application of the current kit, by adding in key pediatric AML gene candidates. This will enable a very targeted and specific kit for pediatric patients.

Additionally the kit will be implemented using an error corrected sequencing (ECS) approach to allow for better detection of low-level variants (Schmitt et al., 2012). Detecting residual leukemic cells may provide key information for understanding the transition from remission to relapse in pediatric AML. ECS uses single molecule indexing to provide a strategy for overcoming the error rate of NGS (Kirsch & Klein, 2012). This approach uses read families, or multiple reads generated from a unique index, to calculate the probability of a variant being real versus background sequencing error.

Recently ECS has been reported to detect rare hematopoietic sub-clones with TP53 mutations in healthy elderly individuals (Young, A, Wong, TN, Ley TJ, Link, DC., Druley, 2014). Interestingly, 9 of the 20 healthy patients analyzed had a variant in TP53 with an allele frequency between one in ten thousand. The rare variants were verified using digital droplet PCR, and the results supported the ECS allele frequencies. A major advantage to ECS is the ability to distinguish low level variants

from sequencing artifacts, which is important for cancer genomics as gaining insight of residual markers of cancer may lead to better treatment strategies.

The ECS study for the pediatric AML project will begin in early 2016 and will start with a small pilot dataset consisting of ~100 patients. InDel detection and calling has not yet been applied with ECS data. The initial goal is to detect FLT3 / ITD, and determine if there are any residual clones in the remission samples. Furthermore, samples with known SNVs in oncogenes, such as KRAS, will also be tested for residual markers of disease at the remission state.

Collectively, this dissertation project touches upon relevant and current topics in precision medicine including the use of a high performance computer infrastructure to help translate big data into knowledge. The expansion of data types for clinical molecular biology data is driving the need for bioinformatics and access to community based projects with large sample sizes (Figure 5.2). Clinicians, researchers, and bioinformaticians need to work in solidarity to help drive the use of genomic data in the clinical setting.

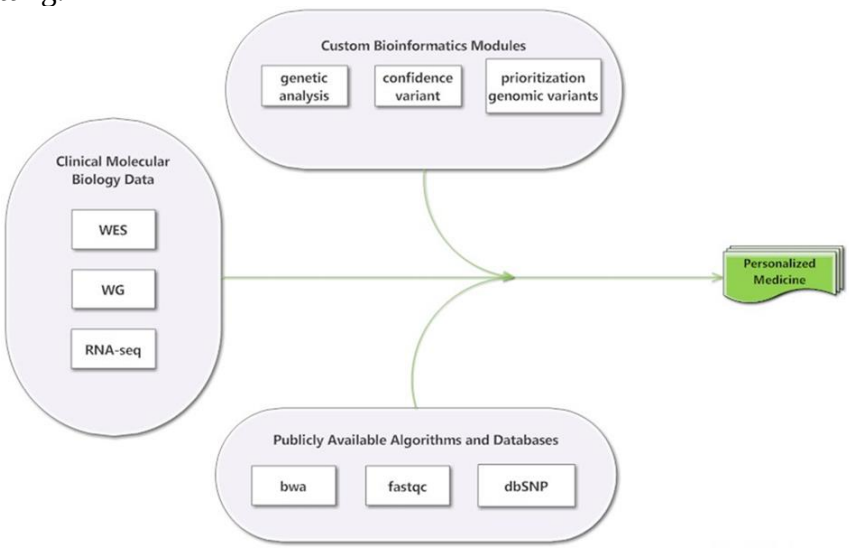

Figure 5.2. Overview of components for precision medicine.

REFERENCES

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73. http://doi.org/10.1038/nature09534

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–76. http://doi.org/10.1038/nrg2958

ALVING, A. S., CARSON, P. E., FLANAGAN, C. L., & ICKES, C. E. (1956). Enzymatic deficiency in primaquine-sensitive erythrocytes. *Science (New York, N.Y.)*, *124*(3220), 484–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13360274

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–73. http://doi.org/10.1038/nprot.2010.116

Andersson, A. K., Ma, J., Wang, J., Chen, X., Gedman, A. L., Dang, J., … Downing, J. R. (2015). The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nature Genetics*, *47*(4), 330–337. http://doi.org/10.1038/ng.3230

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–9. http://doi.org/10.1038/nature07517

Bodini, M., Ronchini, C., Giaco, L., Russo, A., Melloni, G. E. M., Luzi, L., … Riva, L. (2015). The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations. *Blood*, *125*(4), 600–605. http://doi.org/10.1182/blood-2014-05-576157

Cancer Genomics. (2014). Elsevier. http://doi.org/http://dx.doi.org/10.1016/B978-0-12-396967-5.00017-7

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., & Reese, J. (2014).

BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics*, *15*, 104. http://doi.org/10.1186/1471-2105-15-104

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–9. http://doi.org/10.1038/nbt.2514

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. http://doi.org/10.4161/fly.19695

Collins, F. S., & Varmus, H. (2015). A New Initiative on Precision Medicine. *The New England Journal of Medicine*. http://doi.org/10.1056/NEJMp1500523

Connolly, R. M., Nguyen, N. K., & Sukumar, S. (2013). Molecular pathways: current role and future directions of the retinoic acid pathway in cancer prevention and treatment. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *19*(7), 1651–9. http://doi.org/10.1158/1078-0432.CCR-12-3175

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, *38*(1), 75–81. http://doi.org/10.1038/ng1697

Cutting, G. R. (2014). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics*, *16*(1), 45–56. http://doi.org/10.1038/nrg3849

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–8. http://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. http://doi.org/10.1038/ng.806

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3), e1003348. http://doi.org/10.1371/journal.pgen.1003348

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics*, *24*(7), 344–352. http://doi.org/10.1016/j.tig.2008.04.005

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. http://doi.org/10.1093/nsr/nwt032

Firth, H. V, Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, *84*(4), 524–33. http://doi.org/10.1016/j.ajhg.2009.03.010

Gamis, A. S., Alonzo, T. A., Perentesis, J. P., & Meshinchi, S. (2013). Children's Oncology Group's 2013 blueprint for research: acute myeloid leukemia. *Pediatric Blood & Cancer*, *60*(6), 964–71. http://doi.org/10.1002/pbc.24432

Gargis, A. S., Kalman, L., Berry, M. W., Bick, D. P., Dimmock, D. P., Hambuch, T., … Lubin, I. M. (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology*, *30*(11), 1033–6. http://doi.org/10.1038/nbt.2403

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biology*, *12*(9), 228. http://doi.org/10.1186/gb-2011-12-9-228

Gilliland, D. G., & Griffin, J. D. (2002). The roles of FLT3 in hematopoiesis and leukemia. *Blood*, *100*(5), 1532–42. http://doi.org/10.1182/blood-2002-02-0492

Graham, J. M., & Hennekam, R. C. (2014). Genetics of common malformations. *European Journal of Medical Genetics*. http://doi.org/10.1016/j.ejmg.2014.05.007

Hayakawa, F., Towatari, M., Kiyoi, H., Tanimoto, M., Kitamura, T., Saito, H., & Naoe, T. (2000). Tandem-duplicated Flt3 constitutively activates STAT5 and MAP kinase and introduces autonomous cell growth in IL-3-dependent cell lines. *Oncogene*, *19*(5), 624–31. http://doi.org/10.1038/sj.onc.1203354

Huang, H., Hu, Z.-Z., Arighi, C. N., & Wu, C. H. (2007). Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Frontiers in Bioscience : A Journal and Virtual Library*, *12*, 5071–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17569631

Ingrid Lobo. (2008). Environmental Influences on Gene Expression. *Nature Education*, *1*(1), 39.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., … von Mering, C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, *37*(Database issue), D412–6. http://doi.org/10.1093/nar/gkn760

Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and "race". *Nature Genetics*, *36*(11 Suppl), S28–33. http://doi.org/10.1038/ng1435

Kannry, J. L., & Williams, M. S. (2013). Integration of genomics into the electronic health record: mapping terra incognita. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *15*(10), 757–60. http://doi.org/10.1038/gim.2013.102

Kirsch, S., & Klein, C. A. (2012). Sequence error storms and the landscape of mutations in cancer. *Proceedings of the National Academy of Sciences*, *109*(36), 14289–14290. http://doi.org/10.1073/pnas.1212246109

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. http://doi.org/10.1101/gr.129684.111

Kosan, C., & Godmann, M. (2016). Genetic and Epigenetic Mechanisms That Maintain Hematopoietic Stem Cell Function. *Stem Cells International*, *2016*, 1–14. http://doi.org/10.1155/2016/5178965

Kozlowski, P., de Mezer, M., & Krzyzosiak, W. J. (2010). Trinucleotide repeats in human genome and exome. *Nucleic Acids Research*, *38*(12), 4027–39. http://doi.org/10.1093/nar/gkq127

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., … Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, *22*(8), 1525–32. http://doi.org/10.1101/gr.138115.112

Lennartsson, J., & Ronnstrand, L. (2012). Stem Cell Factor Receptor/c-Kit: From Basic Science to Clinical Implications. *Physiological Reviews*, *92*(4), 1619–1649. http://doi.org/10.1152/physrev.00046.2011

Lennerz, J. K., & Stenzinger, A. (2015). Allelic Ratio of KRAS Mutations in Pancreatic Cancer. *The Oncologist*, *20*(4), e8–e9. http://doi.org/10.1634/theoncologist.2014-0408

Lepri, F. R., Scavelli, R., Digilio, M. C., Gnazzo, M., Grotta, S., Dentici, M. L., …

Dallapiccola, B. (2014). Diagnosis of Noonan syndrome and related disorders using target next generation sequencing. *BMC Medical Genetics*, *15*, 14. http://doi.org/10.1186/1471-2350-15-14

Li, H. (2013). Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM, *00*(00), 1–3.

Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., … Wang, J. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature Biotechnology*, *29*(8), 723–30. http://doi.org/10.1038/nbt.1904

Liu, B., Morrison, C. D., Johnson, C. S., Trump, D. L., Qin, M., Conroy, J. C., … Liu, S. (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, *4*(11), 1868–81. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3875755&tool=pmcentrez&rendertype=abstract

Longo, D. L., Döhner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute Myeloid Leukemia. *New England Journal of Medicine*, *373*(12), 1136–1152. http://doi.org/10.1056/NEJMra1406184

Louie, E., Ott, J., & Majewski, J. (2003). Nucleotide frequency variation across human genes. *Genome Research*, *13*(12), 2594–601. http://doi.org/10.1101/gr.1317703

Ludwig, M. (2002). Functional evolution of noncoding DNA. *Current Opinion in Genetics & Development*, *12*(6), 634–639. http://doi.org/10.1016/S0959-437X(02)00355-6

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, *21*(16), 3448–9. http://doi.org/10.1093/bioinformatics/bti551

Mager, J., Glaser, G., Razin, A., Izak, G., Bien, S., & Noam, M. (1965). Metabolic effects of pyrimidines derived from fava bean glycosides on human erythrocytes deficient in glucose-6-phosphate dehydrogenase. *Biochemical and Biophysical Research Communications*, *20*(2), 235–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5850686

Manwar Hussain, M. R., Khan, A., & Ali Mohamoud, H. S. (2014). From genes to health - challenges and opportunities. *Frontiers in Pediatrics*, *2*, 12. http://doi.org/10.3389/fped.2014.00012

McCandless, S. E., Brunger, J. W., & Cassidy, S. B. (2004). The Burden of Genetic Disease on Inpatient Care in a Children's Hospital. *The American Journal of Human Genetics*, *74*(1), 121–127. http://doi.org/10.1086/381053

McGarvey, P. B., Zhang, J., Natale, D. A., Wu, C. H., & Huang, H. (2011). Protein-centric data integration for functional analysis of comparative proteomics data. *Methods in Molecular Biology (Clifton, N.J.)*, *694*, 323–39. http://doi.org/10.1007/978-1-60761-977-2_20

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … Depristo, M. A. (2010). The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data, 1297–1303. http://doi.org/10.1101/gr.107524.110.20

McLornan, D., Percy, M., & McMullin, M. F. (2006). JAK2 V617F: a single mutation in the myeloproliferative group of disorders. *The Ulster Medical Journal*, *75*(2), 112–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16755940

Meshinchi, S. (2003). Activating mutations of RTK/ras signal transduction pathway in pediatric acute myeloid leukemia. *Blood*, *102*(4), 1474–1479. http://doi.org/10.1182/blood-2003-01-0137

Meshinchi, S., Alonzo, T. A., Stirewalt, D. L., Zwaan, M., Zimmerman, M., Reinhardt, D., … Radich, J. P. (2006). Clinical implications of FLT3 mutations in pediatric AML. *Blood*, *108*(12), 3654–61. http://doi.org/10.1182/blood-2006-03-009233

Meshinchi, S., & Appelbaum, F. R. (2009). Structural and functional alterations of FLT3 in acute myeloid leukemia. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *15*(13), 4263–9. http://doi.org/10.1158/1078-0432.CCR-08-1123

Meshinchi, S., Woods, W. G., Stirewalt, D. L., Sweetser, D. A., Buckley, J. D., Tjoa, T. K., … Radich, J. P. (2001). Prevalence and prognostic significance of Flt3 internal tandem duplication in pediatric acute myeloid leukemia. *Blood*, *97*(1), 89–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11133746

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, *12*(11), R112. http://doi.org/10.1186/gb-2011-12-11-r112

Narzisi, G., O'Rawe, J. A., Iossifov, I., Fang, H., Lee, Y.-H., Wang, Z., … Schatz, M. C. (2014). Accurate de novo and transmitted indel detection in exome-capture

data using microassembly. *Nature Methods*. http://doi.org/10.1038/nmeth.3069

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, *9*(5), 471–472. http://doi.org/10.1038/nmeth.1938

Novelli, E. M., Ramirez, M., & Civin, C. I. (1998). Biology of CD34+CD38- cells in lymphohematopoiesis. *Leukemia & Lymphoma*, *31*(3-4), 285–93. http://doi.org/10.3109/10428199809059221

Ostronoff, F., Othus, M., Gerbing, R. B., Loken, M. R., Raimondi, S. C., Hirsch, B. A., … Meshinchi, S. (2014). NUP98/NSD1 and FLT3/ITD coexpression is more prevalent in younger AML patients and leads to induction failure: a COG and SWOG report. *Blood*, *124*(15), 2400–7. http://doi.org/10.1182/blood-2014-04-570929

Pui, C.-H., Carroll, W. L., Meshinchi, S., & Arceci, R. J. (2011). Biology, risk stratification, and therapy of pediatric acute leukemias: an update. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *29*(5), 551–65. http://doi.org/10.1200/JCO.2010.30.7405

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–54. http://doi.org/10.1038/nature05329

Rosen, D. B., Minden, M. D., Kornblau, S. M., Cohen, A., Gayko, U., Putta, S., … Cesano, A. (2010). Functional characterization of FLT3 receptor signaling deregulation in acute myeloid leukemia by single cell network profiling (SCNP). *PloS One*, *5*(10), e13543. http://doi.org/10.1371/journal.pone.0013543

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., … Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928–33. http://doi.org/10.1038/35057149

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. http://doi.org/10.1073/pnas.74.12.5463

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, *109*(36), 14508–14513. http://doi.org/10.1073/pnas.1208715109

Schneider, J. A., Pungliya, M. S., Choi, J. Y., Jiang, R., Sun, X. J., Salisbury, B. A., & Stephens, J. C. (2003). DNA variability of human genes. *Mechanisms of Ageing and Development*, *124*(1), 17–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12618002

Schuback, H. L., Arceci, R. J., & Meshinchi, S. (2013). Somatic characterization of pediatric acute myeloid leukemia using next-generation sequencing. *Seminars in Hematology*, *50*(4), 325–32. http://doi.org/10.1053/j.seminhematol.2013.09.003

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., … Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, *305*(5683), 525–8. http://doi.org/10.1126/science.1098918

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–11. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29783&tool=pmcentrez&rendertype=abstract

Shu, L.-L., & Yang, M. (2012). [Imatinib in treatment of thrombocythemia and other myeloproliferative diseases]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi / Zhongguo Bing Li Sheng Li Xue Hui = Journal of Experimental Hematology / Chinese Association of Pathophysiology*, *20*(6), 1507–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23257463

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–32. http://doi.org/10.1038/nrg3642

Singleton, M. V, Guthery, S. L., Voelkerding, K. V, Chen, K., Kennedy, B., Margraf, R. L., … Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, *94*(4), 599–610. http://doi.org/10.1016/j.ajhg.2014.03.010

Spencer, D. H., Abel, H. J., Lockwood, C. M., Payton, J. E., Szankasi, P., Kelley, T. W., … Duncavage, E. J. (2013). Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. *The Journal of Molecular Diagnostics : JMD*, *15*(1), 81–93. http://doi.org/10.1016/j.jmoldx.2012.08.001

Stirewalt, D. L., Pogosova-Agadjanyan, E. L., Tsuchiya, K., Joaquin, J., & Meshinchi, S. (2014). Copy-neutral loss of heterozygosity is prevalent and a late event in the pathogenesis of FLT3/ITD AML. *Blood Cancer Journal*, *4*, e208.

http://doi.org/10.1038/bcj.2014.27

Testa, U., & Pelosi, E. (2013). The Impact of FLT3 Mutations on the Development of Acute Myeloid Leukemias. *Leukemia Research and Treatment*, *2013*, 1–14. http://doi.org/10.1155/2013/275760

Thiede, C. (2002). Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood*, *99*(12), 4326–4335. http://doi.org/10.1182/blood.V99.12.4326

Thol, F., Kölking, B., Damm, F., Reinhardt, K., Klusmann, J.-H., Reinhardt, D., … Heuser, M. (2012). Next-generation sequencing for minimal residual disease monitoring in acute myeloid leukemia patients with FLT3-ITD or NPM1 mutations. *Genes, Chromosomes & Cancer*, *51*(7), 689–95. http://doi.org/10.1002/gcc.21955

Tropeano, M., Ahn, J. W., Dobson, R. J. B., Breen, G., Rucker, J., Dixit, A., … Collier, D. A. (2013). Male-biased autosomal effect of 16p13.11 copy number variation in neurodevelopmental disorders. *PloS One*, *8*(4), e61365. http://doi.org/10.1371/journal.pone.0061365

Wang, Z., Liu, X., Yang, B.-Z., & Gelernter, J. (2013). The role and challenges of exome sequencing in studies of human diseases. *Frontiers in Genetics*, *4*, 160. http://doi.org/10.3389/fgene.2013.00160

Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, *11*(5), R53. http://doi.org/10.1186/gb-2010-11-5-r53

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–71. http://doi.org/10.1093/bioinformatics/btp394

Young, A. L., Wong, T. N., Hughes, A. E. O., Heath, S. E., Ley, T. J., Link, D. C., & Druley, T. E. (2015). Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing. *Leukemia*, *29*(7), 1608–1611. http://doi.org/10.1038/leu.2015.17

Young, A, Wong, TN, Ley TJ, Link, DC., Druley, T. (2014). Rare Hematopoietic Subclones Harboring Leukemogenic TP53 Mutations Are Detectable Via Error-Corrected Sequencing in Healthy Elderly Individuals. *Blood*, *124*(21), 2907–2907.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, *14 Suppl 1*, S1. http://doi.org/10.1186/1471-2105-14-S11-S1

Zwaan, C. M., Meshinchi, S., Radich, J. P., Veerman, A. J. P., Huismans, D. R., Munske, L., … Griesinger, F. (2003). FLT3 internal tandem duplication in 234 children with acute myeloid leukemia: prognostic significance and relation to cellular drug resistance. *Blood*, *102*(7), 2387–94. http://doi.org/10.1182/blood-2002-12-3627

If you want to number your bibliographic entries, change the style of the items to *Bib Entry - numbered*.

Appendix A

COPYRIGHT PERMISSION FOR CHAPTER 3

Chapter 3 is modified from a journal article, “An Integrated Approach for Analyzing Clinical Genomic Variant Data from Next Generation Sequencing”, published in Journal of Biomolecular Techniques, an open access journal. As the author of this article, I hold the copyright and agree to it being modified for use in this dissertation.

Appendix B

COPYRIGHT PERMISSION CHAPTER 4

Chapter 4 is modified from a journal article, "Development of Bioinformatics Pipeline for Analyzing Clinical Pediatric NGS Data" published in the American Medical Informatics Association Joint Summits Translational Proceedings, and open access journal. As author of this article, I hold the copyright and agree to it being modified for use in this dissertation.

Appendix C

IRB PROTOCAL APPROVAL

The genomic sequencing analyzed for this dissertation was de-identified. Following the University of Delaware's institutional review board (IRB) procedures, a project description was submitted (565619-1: Development of Bioinformatics Methodologies for Clinical NGS Genomic Data) and was determined to be exempt from IRB review according to federal regulations (exemption category 46.101(b)(4)).