Discovering rare variants from populations to families

Author: Amit R. Indap

Boston College
The Graduate School of Arts and Sciences
Department of Biology

Discovering rare variants from populations to families

by
Amit R. Indap

A Dissertation
submitted in partial fulfillment of the requirements
for the degree of
of Doctor of Philosophy

December 2013

Discovering rare variants from populations to families

AMIT R. INDAP

Adviser: Gabor T. Marth

Abstract

Partitioning an individual's phenotype into genetic and environmental components has been a major goal of genetics since the early 20th century. Formally, the proportion of phenotypic variance attributable to genetic variation in the population is known as heritability. Genome wide association studies have explained a modest percentage of variability of complex traits by genotyping common variants. Currently, there is great interest in what role rare variants play in explaining the missing heritability of complex traits. Advances of next generation sequencing and genomic enrichment technologies over the past several years have made it feasible to re-sequence large numbers of individuals, enabling the discovery of the full spectrum of genetic variation segregating in the human population, including rare variants. The four projects that comprise my dissertation all revolve around the discovery of rare variants from next generation sequencing datasets. In my first project, I analyzed data from the exon sequencing pilot of the 1000 Genomes Project, where I discovered variants from exome capture sequencing experiments in a worldwide sample of nearly 700 individuals. My results show that the allele frequency spectrum of the dataset has an excess of rare variants.

My next project demonstrated the applicability of using whole-genome amplified DNA (WGA) in capture sequencing. WGA is a method that amplifies DNA from nanogram starting amounts of template. In two separate capture experiments I compared the concordance of call sets, both at the site and genotype level, of variant calls derived from WGA and genomic DNA. WGA derived calls have excellent con-

cordance metrics, both at the site and genotypic level, suggesting that WGA DNA can be used in lieu of genomic DNA. The results of this study have ramifications for medical sequencing experiments, where DNA stocks are a finite quantity and re-collecting samples maybe too expensive or not possible.

My third project kept its focus on capture sequencing, but in a different context. Here, I analyzed sequencing data from Mendelian exome study of non-sensorineural hearing loss (NSHL). A subset of 6 individuals (5 affected, 1 unaffected) from a family of European descent were whole exome sequenced in an attempt to uncover the causative mutation responsible for the loss of hearing phenotype in the family. Previous linkage analysis uncovered a linkage region on chr12, but no mutations in previous candidate genes were found, suggesting a novel mutation segregates in the family. Using a discrete filtering approach with a minor allele frequency cutoff, I uncovered a putative causative non-synonymous mutation in a gene that encodes a transmembrane protein. The variant perfectly segregates with the phenotype in the family and is enriched in frequency in an unrelated cohort of individuals.

Finally, for my last project I implemented a variant calling method for family sequencing datasets, named Pgmsnp, which incorporates Mendelian relationships of family members using a Bayesian network inference algorithm. My method has similar detection sensitivities compared to other pedigree aware callers, and increases power of detection for non-founder individuals.

Acknowledgements

A knowledge - a precise knowledge - of the laws of heredity will give man a power over his future that no other science has yet endowed him with. I am not going to say that this knowledge is going to create the millennium of the human race; I can only say it will change man's destinies profoundly - wither for good or evil the future alone will show! - William Bateson, 1902

First and foremost I would like to thank my Mom and Dad from supporting me in all my endeavours and being especially supportive in my decision to attend graduate school. It has been a circuitous journey to the Ph.D., but no matter what they have always supported me.

I would like to thank my advisor Gabor T. Marth for giving me an opportunity to work in his group and to my labmates for being great colleagues. Human genetics is a small world and I hope our paths cross again in the future. I would like to thank my former PIs I had the opportunity with to work leading up to my PhD which includes the following people: Michael F. Hammer, Michael Olivier, Andrew G. Clark, and Carlos D. Bustamante. In particular, my undergraduate experience with the Hammer lab and the UBRP program headed by Carol Bender had a great influence on my decision to choose biological research and in particular, human genetics, as a career. Special thanks to the following people for their help, moral support, and company: Melanie Huntley, Colin Maccannell, Wilfried Haerty, Clement Chow, Kirk Lohmueller, Jeremiah Degenhardt, Angela Stevenson, Jannette Bushard Olmsted, Katie Moorhouse, Megan Farrell, Andrew Denninger, Brooke Anderson-White, Heather Gudejko, Michelle Busby, Jonathan Parisi, and Dana Parisi.

And finally, I would like to thank my wife Abha for her love, support, and patience (particularly over this past year). I couldn't have done it without you!

To my parents, Ramakant and Hema.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since the discovery of the laws of inheritance by Gregor Mendel in the 19th century the field of genetics has progressed with new findings made possible by advances in technology. In the early 20th century the fields of population and quantitative genetics were founded with the formulation of new mathematical and statistical methods to study variation and the inheritance of traits. Technological advances led the birth of molecular biology in the 1970s that enabled geneticists to study genes at the molecular level. Today, in the modern post-genomic world, large scale analyses for genetic variation data have given an unparalleled opportunity to uncover the genetic basis of phenotypic traits. Like much of its past history, progress in genetics still depends on methodological and technological advances to study variation at a fine scale.

1.1 Genome-wide Association Studies

Single nucleotide polymorphisms (SNPs) are single base pair differences between individual chromosomes in a population. Genetic association studies genotype SNPs and search for correlations between a SNP genotype and a disease phenotype in a set of affected and unaffected individuals. SNPs that are tested for association either must be the causative allele or be in *linkage disequilibrium* (LD) with the

causative allele. LD is the non-random association of alleles between adjacent loci on a chromosome. SNPs that are in LD with a causative allele serve as a proxy and the association with the disease phenotype is maintained.

The completion of the Human Genome Project (HGP) facilitated the discovery of millions of SNPs and their use in genetic association studies. Shortly after the completion of the HGP, the HapMap project [23] commenced. Its primary aim was to catalog common genetic variation (minor allele frequency or MAF $\geq .05$) in populations throughout the world. The HapMap project truly enabled geneticists to embark on genome-wide association studies (GWAS). The HapMap project and subsequent GWAS study designs were based on the common disease common variant hypothesis (CDCV), which states that genetic risk for complex diseases can be attributed to loci where there are a limited number of common variants segregating in the population.

Prior to GWAS

Prior to the GWAS era there were several early milestones, both technical and methodological, that made association studies feasible. [78]. David Botstein and colleagues in 1980 called for the construction of a genome wide linkage map using restriction length fragment length polymorphisms (RFLPs) as markers [13]. By 1987 the first linkage map of the human genome was constructed [32, 52]. The mapping of Mendelian traits and diseases was a natural application of linkage mapping, but many other traits and diseases followed a multifactorial inheritance pattern. Lander and Botstein [81] first proposed using linkage disequilibrium mapping in 1986, even prior to the first human linkage map being completed. The first demonstrated example of LD between a DNA polymorphism and a disease mu-

tation was between a RFLP allele in the β-globin gene and sickle-cell hemoglobin [68].

Early estimates of how much LD was present in the human population suggested that it would extend to a 100 kbp or less [12] so a high density map would be needed to carry out LD mapping. At the time this was too laborious a task and Lander and Botstein suggested using LD mapping in population isolates who have a greater extent of LD. As the 1980s progressed genetic mapping studies took off as microsatellite markers replaced RFLPs [142]. Family based linkage studies were the primary tool in locating disease genes with LD mapping used to fine map the location of genes first identified by linkage. This approach was first established by Kerem [71] in locating the gene for cystic fibrosis. The first whole-genome LD mapping study resulted in finding the gene responsible for recurrent intrahepatic cholestasis (BRIC) in population isolates in the Netherlands [60]. A similar approach was used to locate the gene for Hirschprung disease in the Mennonite community [116].

In 1996 Neil Risch and Kathleen Merikangas wrote an influential essay promoting the idea that association mapping is a better approach than family based design linkage studies for discovering common variants of small effect [122]. The reasoning behind this is two-fold. First, since linkage studies rely on allele sharing between relatives and if the allele is commonly segregating in the population it can enter the pedigree via multiple founders. Second, if the conferred risk of the allele is small, affected individuals may have the phenotype due to other causes. Association studies avoid the first drawback and are well-suited to uncovering alleles of small to moderate effect. The only drawback at the time Risch and Merikangas published their essay was the lack of the molecular tools to discover and genotype common genetic variation segregating in the human population. To address this roadblock a public-private consortium of companies and academics formed The SNP Consortium

to develop the genotyping technology to identify a genome wide collection of at least 100,000 SNPs [123].

Haplotype Blocks and LD

In the early 2000s after the the completion of the HGP, several publications detailed for the first time LD patterns in the human genome [109, 39]. They all showed that there was a "block" like pattern to LD, where there were regions of strong LD with low haplotype diversity. These regions are called haplotype blocks. There are various methods to computationally define a haplotype block [109, 39], and while there is no gold standard definition, analyzing LD patterns with a variety of methods may be the best approach [64]. No matter how they are identified, the definition of haplotype blocks served to reduce the number of SNPs required in association studies by identifying and typing only the subset of tag SNPs which uniquely identify common haplotypes present in a block [28]. The frequencies of these SNPs can be compared in groups of case and control individuals. This was the underpinning of the HapMap project.

For much of the early to mid-2000s after the HGP was completed the common-disease, common variant hypothesis (CDCV) was the prevailing thought in the human genetics community. Its main tenet is that genetic risk for common diseases is conferred by a single common variant (or small number of them) segregating in the population. If this was true, and taking advantage of the block like structure of LD, association mapping would have strong power to uncover variants associated with disease[114, 121]. Association studies work on the premise that SNP genotypes are correlated with a disease phenotype. Individual SNPs are genotyped and the frequency of alleles are compared between groups of affected and un-affected individ-

uals. SNPs that are tested for association either must be the causative allele or be in linkage disequilibrium (LD) with the causative allele. Individuals are genotyped for SNPs using genotyping arrays which typically 100,000 to 500,000 markers. Data quality of the resulting genotypes is checked, typically by removing SNPs that are not in Hardy-Weinberg equilibrium [6]. Association testing is typically done on a single SNP basis, in which each SNP is testing individually for association with the phenotype. Typically, a 2-by-2 χ^2 is constructed to test individual alleles, under the null hypothesis of no association.

The first successful GWAS utilizing HapMap data was a study by Hoh and colleagues [74] that uncovered a risk allele in the *CFH* gene for age-related macular degeneration (AMD). AMD is a major cause of blindness in the elderly and is characterized by progressive damage to the retina caused by accumulation of extracellular deposits called drusen. Previous linkage studies identified chromosomal regions demonstrating linkage to the phenotype, but failed in discovering a causative allele [1]. Hoh and colleagues designed a genome-wide association study comprised of 96 individuals with and 50 individuals without AMD. The 146 individuals genotyped all had European ancestry. They genotyped 103,000 SNPs spread across the 22 autosomal chromosomes. After careful quality control of SNP genotypes, single marker association was performed by constructing a 2-by-2 contingency table of allele frequencies and computing Pearson's χ^2 test statistic based on the χ^2 distribution under the null hypothesis of no association. Two SNPs, *rs380390* and *rs1329428*, within an intron of the *CFH* gene had significant p-values (after Bonferroni correction). Hoh used genotype data from the Utah-CEPH population from the HapMap project to closely examine LD patterns in this region. Using a haplotype block definition from [39], which is based on D' values, the two associated SNPs were located in a 41 kb LD block contained within the *CFH* gene. Six SNPs genotyped by Hoh were contained

in this 41 kb block and they formed four predominant haplotypes. The highest risk haplotype contained the risk allele for SNP *rs380390* and being heterozygous for this marker increased the odds of having AMD by 4.6. To uncover the functional polymorphism responsible for the association signal in the *CFH* gene, Hoh and colleagues re-sequenced all exons in the gene in each of the 96 affected individuals. They discovered a non-synonymous tyrosine-histidine variant 2 kb upstream of the previously identified haplotype block. The *CFH* gene is part of the innate immune system and regulates against infection. Individuals who carry the risk allele develop AMD due to abnormal *CFH* activity that elicits an inflammatory process [74].

The AMD study described above is exemplary of a successful GWAS result. Other traits examined by GWAS designs have uncovered novel loci, but haven't explained the majority of the heritability. The classic example of this is human height. Twin studies indicate that height is 80% heritable, but a recent meta-analysis of 46 GWAS on height by Peter Visscher and colleagues uncovered 207 significantly associated SNPs which explained approximately 10% of the heritability [82]. The study analyzed 133,653 individuals and imputed genotypes at 2.5M SNPs. The associated variants span 180 genes and are enriched in loci for skeletal growth. The common SNPs genotyped did fail to explain the majority of the heritability of height but the genes found to be associated contributed to understanding biological mechanisms involved in human growth and stature [57].

Missing Heritability and Rare Alleles

There have been notable successes with GWAS based on the CDCV hypothesis, such as studies uncovering variants conferring risk to age-related macular degeneration [49]. But for other traits, such as height, which has a heritability of 80%, GWAS results using common variants has uncovered only 5 to 10% of this heritability

[141, 147, 82]. The inability of GWAS to explain the majority of heritability of traits has been deemed the 'missing heritability problem' [95] and refuted the CDCV hypothesis [41]. Three alternative explanations to the CDCV hypothesis exist. First is that genetic variance can be attributed to a large number of small effect common variants (the infinitesimal model [41]). This was first formalized by Fisher and states that infinitely many unlinked genes have small additive effects such that selection produces negligible changes in allele frequency and variance [55]. Second, there are a large number of rare alleles with large effect (the RALE model [41, 136]). Finally, various forms of genetic, environmental, and epigenetic interactions [41] account for the missing variance. Standard quantitative genetic theory, formulated by R.A. Fisher [94], supports the infinitesimal model, but to uncover such variants would require sample sizes larger than the human population to detect them. Rather than missing then, most heritability is hidden. Evolutionary and population genetic theory support the RALE model. Population genetic theory predicts that the majority of variants segregate at low frequency [52]. If a variant contributes to disease, it is (mildly) deleterious for an individual's fitness and would be purged from the population. Hence, disease-associated variants would be held at low frequency and mutation-selection balance would prevent such deleterious alleles to drift to higher frequency. Many of the large scale sequencing studies show a skew in the allele frequency spectrum and consistently show a large excess of rare alleles when compared to the standard, constant-sized neutral model.

Until very recently, many of the commercial genotyping arrays utilized in GWAS contain only common variants, making it difficult to detect an association signal from rare variants. The only way to discover rare variants is to comprehensively re-sequence large numbers of samples. Using traditional Sanger sequencing, it would be too labor- and cost-intensive to undertake such studies. Only with the advent

of next-generation sequencing (NGS) platforms, which have higher throughput and lower costs, has it been possible to undertake large, population scale sequencing to discover rare variants in the human population. The 1000 Genomes Project picked up where the HapMap Project left off by whole genome sequencing over a thousand individuals from worldwide populations [33, 22] using NGS technology to catalog the full spectrum of human genetic variation down to minor allele frequency of 1 percent. Companies such as Illumina and Affymetrix are already designing new genotyping arrays with rare variants discovered from the 1000 Genomes project and a new wave of GWAS findings are uncovering new associations with the inclusion of rare alleles [25].

1.2 Exome sequencing for Mendelian and complex traits

With the advent of next generation sequencing technologies and methods for genomic enrichment (described fully in Section 1.4), the application of capture sequencing whole exomes has had notable successes in uncovering the causative allele for Mendelian diseases. Since the advent of GWAS some have argued that attention has been diverted from uncovering the genetic basis of Mendelian disease [4]. Traditional linkage mapping of Mendelian disorders in many cases identifies linkage regions several Mbp in size containing potentially hundreds of candidate genes. Traditional Sanger sequencing would be too painstaking and expensive, but whole-exome sequencing provides a more cost-effective and rapid way to locate causative mutations [42]. Successful applications of exome sequencing to Mendelian diseases have been demonstrated in uncovering the causative mutation in Miller Syndrome [106] and

Kabuki Syndrome [105].

Successful Mendelian exome studies have used a discrete filtering approach as opposed to traditional statistical association methods to identify causative mutations [131]. Basic assumptions of this filtering approach are that causative candidate mutations are non-synonymous, extremely rare (often private to the afflicted family), there is complete penetrance of the phenotype, and every affected individual will carry the disease variant. Besides dividing variants as synonymous or non-synonymous, other methods such as PolyPhen [3] or SIFT [79] can computationally predict whether non-synonymous variants will have deleterious effects based on physiochemical or phylogenetic evidence. Based on the previous assumption that causative mutations will be rare, exome variants are filtered against known variant catalogs such as dbSNP, with records matching existing variants in such databases being removed from consideration. Applying such a filtering strategy under a recessive model of inheritance has resulted in at least 11 studies. Applying it to Mendelian diseases under a dominant inheritance model has proved more difficult (only 4 such studies have been published) [131].

Based on published findings, one would assume that applying exome sequencing to Mendelian diseases is as simple as sequencing a small collection of affecteds and applying a discrete filtering approach to uncover causative candidate mutations. The early success stories of exome sequencing may represent the low hanging fruit and more statistically motivated filtering approaches will need to be developed [66]. Also, the assumption of complete penetrance of disease alleles does not necessarily take into account genetic background effects or modifier loci [18]. There is a distinct possibility that under some conditions, some alleles may be non-pathogenic in one background while pathogenic in another. Rather than hard filter against variant

catalogs, a minor allele cutoff should be specified [131].

Studies applying exome sequencing for complex traits have not been as common. In order for exome sequencing for complex disease/traits to be well-powered, sample sizes in excess of 10,000 will need to be collected [72]. The reason why exome sequencing for Mendelian disease have been performed with only a few samples is because the mutations have a large effect size. Effect sizes of variants influencing complex traits are much smaller, hence the need for larger sample sizes. It would be naive to think that variants contributing to the heritability of complex traits only reside in protein coding regions, so why not just sequence the whole genome? High coverage, whole genome sequencing for the sample sizes required for rare variant association studies are not commonplace yet. Also, many of the significant associations resulting from GWAS designs are in non-coding, intergenic regions, making biological interpretation difficult. Protein coding mutations on the other hand can be more straightforward to interpret. Gene-centric, whole exome studies might not explain all the heritability for complex traits, but it can highlight what genes are involved [72].

1.3 Family based sequencing

Historically the study of family pedigrees have played a central role in human genetics research. Linkage analysis studies have had very notable successes [71]. With availability of dense genotyping arrays, the role of family based designs diminished with the rise of GWAS designs. But high throughput sequencing and genomic enrichment techniques, coupled with the increased interest in rare variants and their role in disease have revived the studies of pedigrees. A significant proportion of new variants discovered with high throughput sequencing segregate at low frequencies in the population, but they may be enriched in families with multiple affected

individuals. Studying families then can increase the statistical power of rare variant analysis [89, 88] and uncover casual variants as well as new biological pathways.

A recent example of a family sequencing study from researchers at Johns Hopkins involved targeted sequencing of genes involved in the neuregulin (NRG) signaling pathway in families with multiple affected members with schizophrenia (SZ) [53]. SZ is a genetically heterogenous trait with alleles that are individually rare and potentially specific to individual cases/families [101]. The Hopkins study re-sequenced 120 exons using the Illumina platform in the 10 genes involved in the neuregulin signaling pathway. NRGs are a collection of signaling molecules that bind to receptors and regulate neuronal migration [53]. The investigators sequenced 24 pairs of affected relatives (48 individuals total) and found deleterious variants clustered in certain SZ pedigrees. SZ exhibits allelic heterogeneity [101], so biological pathways might be different between individual patients, but will be similar within families [53]. Supporting this hypothesis, some of the families had multiple NRG pathway variants while others had none [53].

1.4 Overview of DNA Sequencing Methods and Technologies

DNA sequencing is the process of determining the exact order of nucleotides in a DNA molecule. While many genetic discoveries were made with the tools of classical genetics [15], combined with the tools of molecular biology, DNA sequencing has become an essential tool of modern biomedical research. Below, I briefly describe methods of DNA sequencing, starting with Sanger sequencing and ending with current, state-of-the art next generation sequencing platforms.

Sanger sequencing

Sanger sequencing is a method of DNA sequencing developed by Fredrick Sanger and his colleagues in 1977 that is based upon incorporation of dideoxynucleotides by DNA polymerase during the process of *in vitro* DNA replication [124]. The starting materials for Sanger sequencing include single stranded template DNA, DNA primer(s), DNA polymerase, deoxynucleosidetriphosphates (dNTPs) and modified di-deoxynucleosidetriphosphates (ddNTPs). The ddNTPs lack a 3'-OH (hydroxyl) group required for making a phosphodiester bond between two nucleotides, and upon incorporation by DNA polymerase, stops extension of DNA. Traditional Sanger sequencing requires four separate sequencing reactions, each containing the standard dNTPs, but in each reaction only a single ddNTP is added. After successive rounds of denaturation, annealing, and extension the resulting fragments are heat denatured and run on a polyacrylamide gel. Each of the four separate sequencing reactions are loaded into four lanes on the gel. Finally, the individual DNA bands can be visualized by autoradiography and read off from the resulting images.

Dye terminators and capillary sequencing

In dye-terminator sequencing a fluorescent dye is attached to ddNTPs, thus reducing the number of sequencing reactions from four per sample to one. A similar starting cocktail of DNA template, four species of dNTPs, four fluorescently labeled ddNTPs, and DNA polymerase is used to start the sequencing reaction. The resulting DNA fragments will each have labeled ddNTPs. Rather than using gels to analyze the sequenced fragments, in the late 1990s new automated DNA sequencers used capillary electrophoresis. The products of the sequencing reaction are injected into a capillary tube filled with polymer and high voltage is applied so that the negatively charged DNA travels through the capillaries towards the positive electrode. Prior to reaching the positive electrode the fluorescently labeled DNA fragments pass through a laser

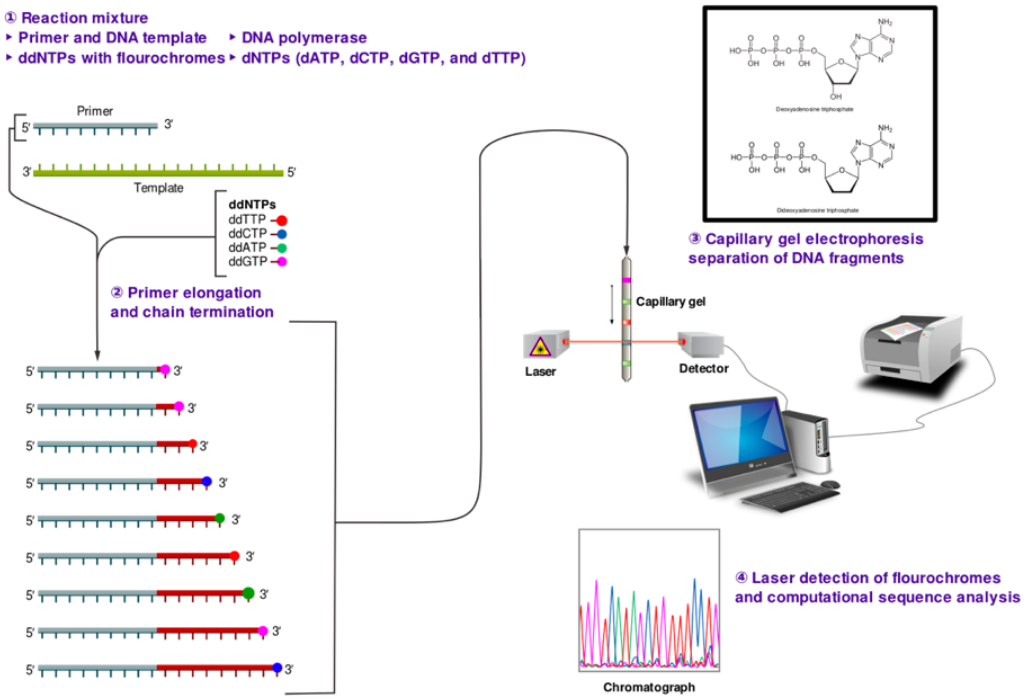


Figure 1.1: **Sanger sequencing** - Sanger sequencing using dye-terminators and capillary electrophoresis, taken from [36]

beam which causes the labeled dyes to fluoresce. Each dye emits a differing wavelength when passing through the laser, and the optical signals are recorded and converted to basecalls by computer software [35]. Figure 1.1 shows an overview of Sanger sequencing using capillary electrophoresis.

Next Generation Sequencing

Capillary-based Sanger sequencing had been the standard method of DNA sequencing, but has limitations to the amount of throughput, scalability, and cost in order to sequence large cohorts of samples. Over the past seven years there has been a shift away from Sanger sequencing to next generation sequencing (NGS) technologies. The key advance with NGS is the ability to sequence DNA in a massively parallel fashion,

enabling the sequencing of gigabase amounts of DNA. Figure 1.2 shows the general steps common to all commercial NGS platforms. A genomic DNA sample is fragmented into a smaller, uniformly represented library of molecules. The sequence of bases are determined by carrying out millions of massively parallel reactions. The resulting sequencing reads are aligned to the reference genome and the sequence of the original sample is determined by the consensus of the aligned reads. There are several commercial vendors offering their own NGS platform, each with unique methods in how parallel sequencing reactions are performed. The following sections describe some of the currently used NGS platforms.

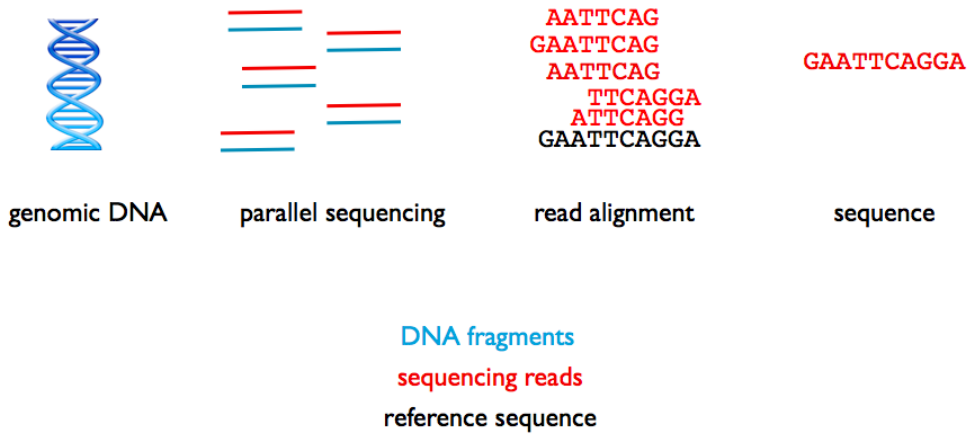


Figure 1.2: **NGS overview** - Common steps involved in all NGS platforms

Illumina

Illumina sequencing reactions occur in an 8-lane flow cell in which DNA fragments are immobilized to adapters that have been ligated to both ends. The flow cell has a dense lawn of primers on its surface, and the addition of unlabeled nucleotides and enzyme initiates *solid-phase bridge amplification*. This step results in priming and extension of the single-stranded DNA templates into double stranded "bridges" on the solid surface of the flow cell. Denaturation results in additional single-stranded DNA templates, and at the end of the bridge amplification step millions of dense clusters of DNA are attached to the solid surface of the flow cell. Bases are determined one at a time by adding a fluorescently labelled dNTP, terminator, primer, and DNA polymerase. The dNTPs have a cleavable fluorophore and have a 3' blocking group. All four dNTPs compete for binding and after laser excitation, the fluorescence is recorded from each cluster. The blocking group is removed and the cycle starts again to determine the base by base composition of DNA of each cluster on the flow cell. The recorded fluorescence images from the clusters are then processed into base calls, with the end result being millions of sequencing reads. These reads are then input for any number of short-read aligners or de-novo assemblers.

454

454 Life Sciences sequencing technology uses a massively parallel pyrosequencing technology to generate reads. Pyrosequencing is based upon detecting the release of pyrophosphate when a new nucleotide is incorporated. The first step in the 454 process is emulsion PCR (em-PCR) where adapter-ligated template DNA fragments are affixed on capture beads in a water-oil emulsion. The DNA on the beads is amplified by PCR and then the beads are placed into micro-titer plate, along with other necessary reagents like DNA polymerase, ATP sulfurylase, and luciferase. The plate is placed

into the sequencing instrument where microfluidics delivers all four nucleotides that flow over the plate. There are millions of DNA fragments attached to the beads in the plate which are sequenced in parallel. The DNA polymerase will add on a complementary nucleotide and upon extension the instrument will record the light emitted. The signals are recorded in flow grams which downstream basecalling software will analyze to generate sequencing reads.

Ion Torrent

Ion Torrent sequencing detects the release of hydrogen ions during the polymerization of DNA. A dNTP is incorporated into a growing DNA strand if its complementary to the leading template strand. Upon incorporation a pyrophosphate and a proton (positively charged hydrogen atom) is released when the new covalent bond is formed. The Ion Torrent platform uses microwells on a semiconductor chip that contain a single-stranded DNA molecule whose sequence needs to be determined and a DNA polymerase. dNTPs are sequentially flooded and if a dNTP is incorporated by the polymerase a proton is released, as previously described. This changes the pH of the microwell, and the changes in pH are measured by an ion sensitive field-effect transistor (ISFET). Hence, each semiconductor chip contains microwells, ISFET detectors, and the proton release will change the current of the resistor. These current changes are transmitted to the computer which then are algorithmically converted into basecalls. Unlike other sequencing technologies, which are based on labelled nucleotides and laser recordings of light release, Ion Torrent does not require any labelled nucleotides or any intermediate signal processing [133]

Pacific Bioscience

Pacific Biosciences (PacBio) uses single molecule real time sequencing (SMRT) to sequence a DNA molecule in a parallelized fashion, without the need to clonally

amplify template DNA. The SMRT technology effectively observes the activity of DNA polymerase in realtime. The DNA template and polymerase is affixed to the bottom of the zero-mode waveguide (ZMW). Phospholinked nucleotides each have a unique fluorescent dye attached to the phosphate group. When the polymerase incorporates a nucleotide a phosphodiester bond is formed and the dye is cleaved off. The ZMW is an optical nanostructure that allows the creation of subdiffraction detection volumes required for single-molecule fluorescence microscopy [77], allowing for the detection of the incorporation of a single nucleotide. There are millions of ZMWs on a SMRT cell allowing for parallel reactions to run.

Capture sequencing

Coinciding with advancements in new sequencing technologies, there have been considerable advancements of methods for genomic partitioning or enrichment. These techniques capture a DNA region of interest and then are sequenced allowing for many individuals to be sequenced, as opposed to whole genome sequencing of smaller number of samples for a similar cost. [128, 137]. While commonly referred to as exome sequencing, because in many cases protein coding regions have been captured, in fact any portion of the genome can be chosen for enrichment [45, 58].

One of the most popular capture techniques was developed by Andreas Gnirke and colleagues working with Agilent Technologies [45]. This aqueous solution phase hybrid selection uses RNA probes to capture DNA regions of interest. First, 200-mer oligonucleotides are constructed on an Agilent microarray and then cleaved off the array. Each oligo consists of 170 bp, target-specific sequence flanked on each end by a 15 bp universal primer sequence. After PCR a T7 promoter is added in a second round of PCR. Finally *in vitro* transcription in the presence of biotin-UTP generates single stranded RNA "baits" used to fish out regions of interest in a "pond" of

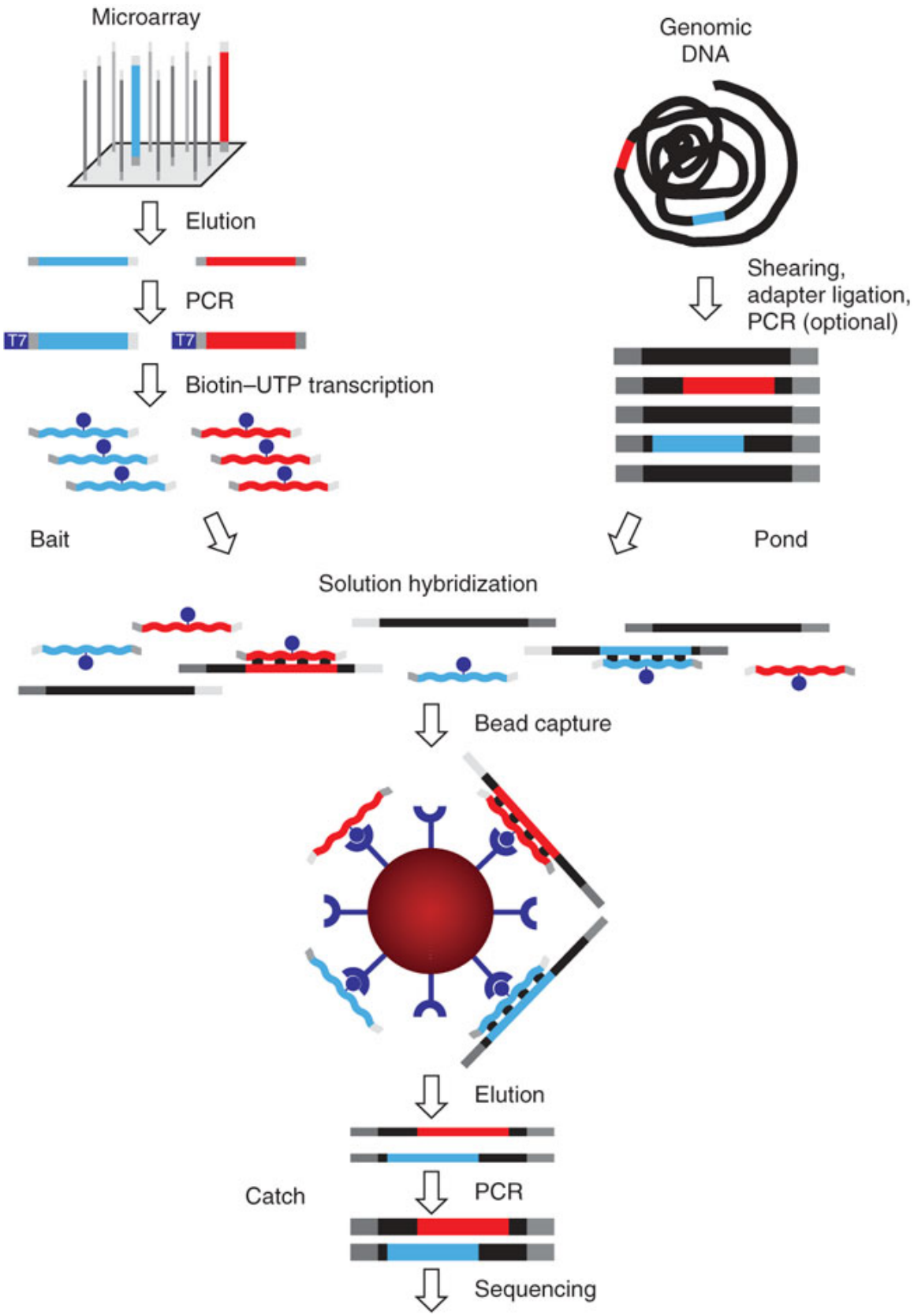


Figure 1.3: **Solution hybrid selection** - Figure taken from [45] showing the steps involved in solution phase hybrid capture.

adapter-ligated, PCR-amplified genomic DNA [45, 137]. There is a vast excess of RNA baits that drives the hybridization process in solution. Streptavidin-coated beads, which have a high affinity for biotin, are used to pull down DNA/RNA hybrids, followed next by PCR amplification, and then finally analyzed on a next-generation sequencing platform. Figure 1.3 shows the steps involved for this capture technology.

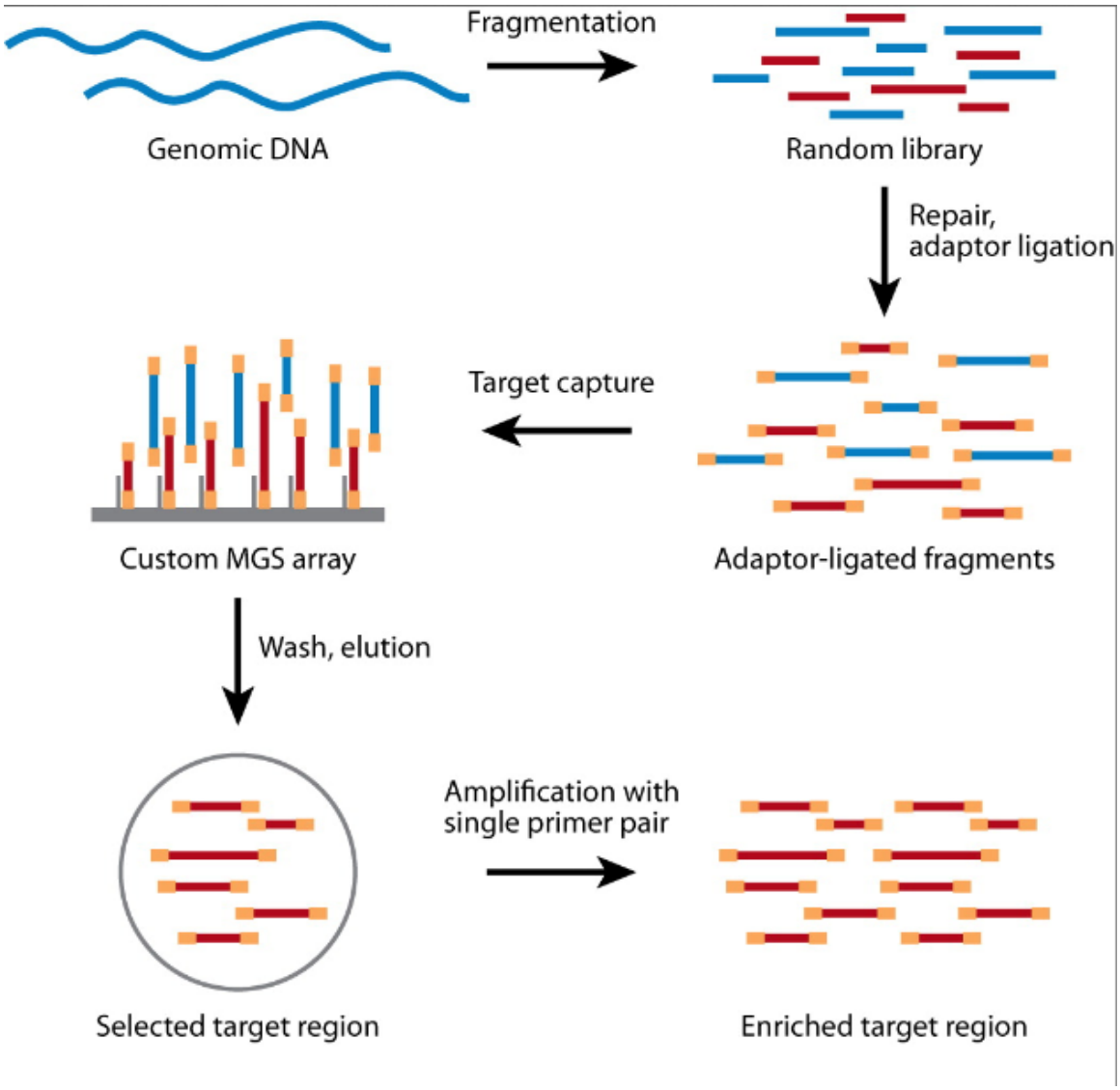


Figure 1.4: **Array hybrid selection** - Figure taken from [137] showing the steps involved in array based hybrid capture.

Another method of capture sequencing is array based hybrid selection using oligonucleotide arrays [137]. The first paper to report the use of the array based hybrid selection was by Hodges et. al. [58] and used the Nimblegen platform. The protocol requires twenty micrograms of genomic DNA that is fragmented into a library of double stranded molecules between 250-1000 bp in length. Prior to array hybridization common PCR adapters are ligated to the genomic DNA. The microarray itself contains 385,000 single stranded 60 base pair oligonucleotides tethered to the surface with the sequence based on the reference human genome assembly. Hybridization of the genomic library to the microarray is carried out for at least two days and heat based elution is performed to recover hybridized material. Universal primers corresponding to the ligated adapters are used for PCR amplification and the target enriched genomic DNA is ready for sequencing [137]. Figure 1.4 shows the steps involved in array based hybrid selection.

1.5 Dissertation Overview

This introductory chapter has given an overview on the increased interest in discovering rare variants and how next-generation sequencing and genomic enrichment technology have enabled their discovery. In Chapter 2 I describe my work in analyzing exon capture data from the 1000 Genomes Pilot project. At the time, this dataset comprising of nearly 700 individuals for over 8000 exons covering 1000 genes was the most comprehensive exome capture study. This work served to introduce me to the informatics issues involved in analyzing NGS data. Chapter 3 continues with an emphasis on capture sequencing, where I demonstrate the applicability of whole genome amplified DNA (WGA) for capture sequencing. This study compared genotype concordance metrics, both at the site and genotype level, showing that WGA derived DNA can be used in place of genomic DNA. This has potential practical implications

for clinical and family-based sequencing studies, where DNA aliquots are a finite resource and recollecting samples is impossible or too expensive. Chapter 4 describes my work in helping uncover a potentially novel mutation involved in non-syndromical hearing loss. In collaboration with investigators at the Medical College of Wisconsin, who had collected from a large family who had a dominantly inherited pattern of hearing loss, I analyzed whole exome sequencing data from a subset of 5 unaffected and 1 affected individuals. Using a discrete filtering approach with a minor allele frequency cutoff I identified a candidate mutation that upon further functional testing is the causative mutation in this family. Finally in Chapter 5, I describe an algorithm, Pgmsnp, which models the relationships of the sequencing data and pedigree relationships in a family sequencing dataset as a Bayesian network to compute posterior genotype probabilities. Pgmsnp has comparable detection sensitivity metrics when compared to similar methods.

Chapter 2

Analysis of the exon sequencing pilot data from the 1000 Genomes Project

2.1 Background

Several publications in the early 2000s gave the first detailed look at the patterns of human genetic diversity and linkage disequilibrium (LD) in the human genome [39, 109]. The results showed a block like pattern of areas of high LD and low haplotype diversity. As a result of these LD blocks the number of SNPs required to be genotyped in an association study could be reduced by only genotyping those loci that uniquely tag common haplotypes in the population. The frequencies of these SNPs could be compared in groups of affected and unaffected individuals, with the assumption that the causative variants were in LD with the SNPs that were genotyped. This was the basic premise of the HapMap Project [23]. The completion of the HapMap ushered in the era of genome wide association studies (GWAS). The first generation GWAS results used common SNPs with minor allele frequency (MAF) above 5 percent. There

are many successful examples of GWAS using common variants discovering genomic regions associated with disease and phenotypic traits [56]. Still, many results of GWAS explain a modest amount of heritability of a trait [96]. Population genetic theory predicts that rare and low frequency variants (defined here respectively as MAF of less than 1 percent and between 1-5 percent) should comprise the bulk of human genetic variation [52] and contribute to the genetic architecture of disease and complex traits [33]. At the time that the HapMap project was being planned, re-sequencing of large numbers of individuals was not feasible and the genotyping arrays used in subsequent GWAS datasets contained only common variants. But over the past 5-7 years new DNA sequencing and genomic enrichment methodologies have removed this obstacle. Hence, in 2008 the 1000 Genomes Consortium was formed with the aim to catalog genetic variation segregating at 1 percent or higher in the human population using high throughput sequencing and genomic enrichment technology. The first phase of the project consisted of three pilot projects. Pilot 1 consisted of low coverage whole genome sequencing of 179 individuals. Pilot 2 consisted of high coverage whole genome sequencing of two trios. Pilot 3 consisted of exon sequencing using genomic enrichment technology of 8140 exons in 697 individuals. The results of all three pilot projects gave the genetics community an unprecedented view of the (spectrum of human genetic variation and drove the development of bioinformatics algorithms and analysis pipelines to analyze data from such experiments. Here I describe my contribution to Pilot 3 of the 1000 Genomes Project.

2.2 Results and Discussion

Data collection

A total of 1.43 Mb of exonic sequence was targeted for capture. Four genome centers collected the data: The Human Genome Sequencing Center Baylor College of

Population code	Population name	N
CEU	CEPH Utah	90
TSI	Toscani Italian	66
CHB	Han Chinese Beijing	109
CHD	Han Chinese Denver	107
JPT	Japanese Tokyo	105
YRI	Yoruban Nigeria	112
LWK	Luhya Kenya	108
		total 697

Table 2.1: Samples sequenced in Pilot 3

Medicine (BCM), the Broad Institute (BI), the Sanger Centre (SC), and the Genome Institute at Washington University (WU). The sequencing platforms used to generate the data included 454 Titanium/FLX and Illumina GAII. The genomic enrichment methods used were Nimblegen liquid phase capture and Agilent solid phase capture [45]. There was considerable heterogeneity in original capture intervals used by each center. The original exon target intervals were derived from Consensus Coding Sequencing Project (CCDS) gene models [115], and each of the center specific coordinate files were intersected with the CCDS gene model. The final interval files consisted of 1.43 Mbp of exonic sequence, representing 8279 exons, spanning 942 genes. A total of 697 samples were sequenced from 7 world populations, as shown in Table 2.1.

SNP variant calling results

There were two pipelines employed to discover SNPs in Pilot 3. The first was executed at Boston College (BC) and the second was executed at the Broad Institute (BI). Figure 2.4 in the Section 5.4 shows the steps employed in each pipeline. The main difference in the complementary pipelines were the alignment and variant calling software used. BC aligned the sequencing data using MOSAIK [85] while BI used MAQ [91] and SSAHA2 [107]. The variant caller at BC, Gigabayes, made SNP calls on all 697 samples simultaneously. BI used their software called UnifiedGenotyper [30] which called variants in each of the 7 populations individually, and then merged into a single callset. Several iterations of comparing call sets and fine tuning parameters

Iteration	Intersection	Union
1	10847	20695
2	10100	17613
3	12358	18277
4	12758	19890

Table 2.2: Comparison and tuning of BC and Broad pipeline results lead to a convergence of call sets.

Population	YRI	LWK	CHB	CHD	JPT	CEU	TSI	All
Technology	ILL,454	454	ILL,454	ILL,454	ILL,454	ILL,454	ILL	ILL,454
SNPs	5175	5459	3415	3431	2900	3489	3281	12758
dbSNP %	53.8	50.1	52.6	50.3	57.9	65.9	65.6	30.36
TsTv	3.56	3.67	3.74	3.64	3.67	3.47	3.53	3.82
Coverage (1st quartile)	18x	19x	18x	30x	20x	20x	20x	19x
Coverage (median)	27x	25x	22x	36x	26x	43x	57x	29x
Coverage (mean)	52x	25x	40x	49x	43x	69x	71x	48x
Coverage (3rd quartile)	42x	32x	37x	44x	54x	98x	118x	49x

Table 2.3: Callset and capture metrics of Pilot 3. The callset metrics were derived from the intersection of the BC and Broad pipelines.

lead to a convergence of callsets, as shown in Table 2.2. The final callset release of SNPs was the intersection of calls made by the BC and BI pipelines. This resulted in a high quality callset of 12758 SNPs. Per population SNP counts, dbSNP fractions, transition transversion ratios are shown in Table 2.3, as well as summaries of median target coverage. Overall, 70% of the SNPs discovered in the exon sequencing pilot had not been previously cataloged in dbSNP. SNP calls were validated by Sanger sequencing and validation summaries are shown in Table 2.4. The validation rates for singletons (meaning the alternate allele is segregating only in one chromosome) is 93.8% and for low frequency variants (meaning the alternate allele is segregating 2-5 chromosomes) is 98.8%, demonstrating that the intersection of calls between the BC and BI pipelines has very high sensitivity.

Allele frequency spectrum

At the time of release the Pilot 3 callset was the largest catalog of coding variation. One way to summarize and quantify the levels of genetic diversity in the dataset is

Alternate allele count (AC)	AC=any	AC=1	AC=2-5
Samples	697	697	697
sites	95	177	166
segregating	92	166	164
validation rate	96.8%	93.8%	98.8%

Table 2.4: Validation results of the Pilot 3SNP callset taken from Marth et. al. [100]

	site 1	site 2	site 3	site 4	site 5
chr 1	0	0	1	0	0
chr 2	0	0	1	0	1
chr 3	0	0	0	1	0
chr 4	0	1	1	0	0
chr 5	1	1	1	0	0
chr 6	0	0	0	0	0
counts	1	2	4	1	1

Table 2.5: Example AFS for a sample of six chromosomes and 5 segregating sites

to calculate the allele frequency spectrum (AFS), also known as the site frequency spectrum [139]. The AFS summarizes the distribution of allele frequencies in a sample of chromosomes. An illustrative example of an AFS is given in Table 2.5. Here, there are 5 segregating sites in a sample of 6 chromosomes. Sites 1, 4, and 5 are present in a single chromosome, site 2 is present in two chromosomes, and site 3 is present in 4 chromosomes. Assuming that the presence of the ancestral nucleotide is denoted as 0 and the mutant nucleotide is denoted by 1, the frequency of a segregating site can range between 1 to n-1 chromosomes, otherwise the site is not polymorphic [139]. Hence, the AFS describes the numbers of segregating sites present in 1 to n-1 chromosomes.

Not all variant sites in the callset had the same number of genotypes in each of the 7 populations. It was necessary to project down the AFS to a common sample size of 100 chromosomes to compare the spectra of each population. Essentially, this involves averaging possible re-samplings of the larger sample size to the smaller one using the hypergeometric distribution [93]. The AFS projection was applied to the

Pilot 3 data using the software package $\delta a \delta i$ [48]. Figure 2.1 shows the AFS of all 7 exon sequencing pilot populations. The neutral, expected AFS is denoted in the figure as theta/x. This spectrum is based on the standard coalescent model of a constant sized, Wright-Fisher population. The expected counts in the spectrum are computed from Watterson's formula [52]. Compared to the neutral AFS, there is a vast excess of singleton and doubleton (derived allele count of 1 and 2, respectively) SNPs in the dataset. The AFS are quite similar for each continental population, with African (YRI and LWK) populations exhibiting the largest number of segregating sites. The Japanese (JPT) AFS has a lower number of low and rare frequency sites compared to other populations. The YRI and CEU populations in Pilot 3 were also part of Pilot 1. Restricting the Pilot1 callset to those SNPs in regions sequenced by the Pilot3, we can compare the AFS of the two study designs. Figure 2.3 displays the AFS of the Pilot 1 and Pilot 3 datasets for the CEU and YRI populations. It shows clearly that the deeper coverage Pilot 3 data is more effective at discovering singleton and low frequency variants [47]. Figure 2.2 displays the AFS of the complete 697 individual (1394 chromosomes) SNP callset, focusing on those sites with alternate allele count of between 1-20. It partitions the data between those sites already in dbSNP v129 and those that are novel. Clearly, the majority of variants segregating in 1-5 chromosomes are not present in dbSNP, demonstrating that deep exon re-sequencing is an effective tool in discovering novel coding variation in the human genome.

Per-base heterozygosity

Per base heterozygosity for Pilot 3 were calculated at non-synonymous, 2-fold, 3-fold, and 4-fold degenerate sites in autosomal target regions. Targeted base pairs were included in the analysis if they had at least 10x or greater coverage in the MOSAIK alignments and had a genotype call in at least 100 chromosomes. Site degeneracy was calculated based on the Gencode [51] annotation model. The results are shown

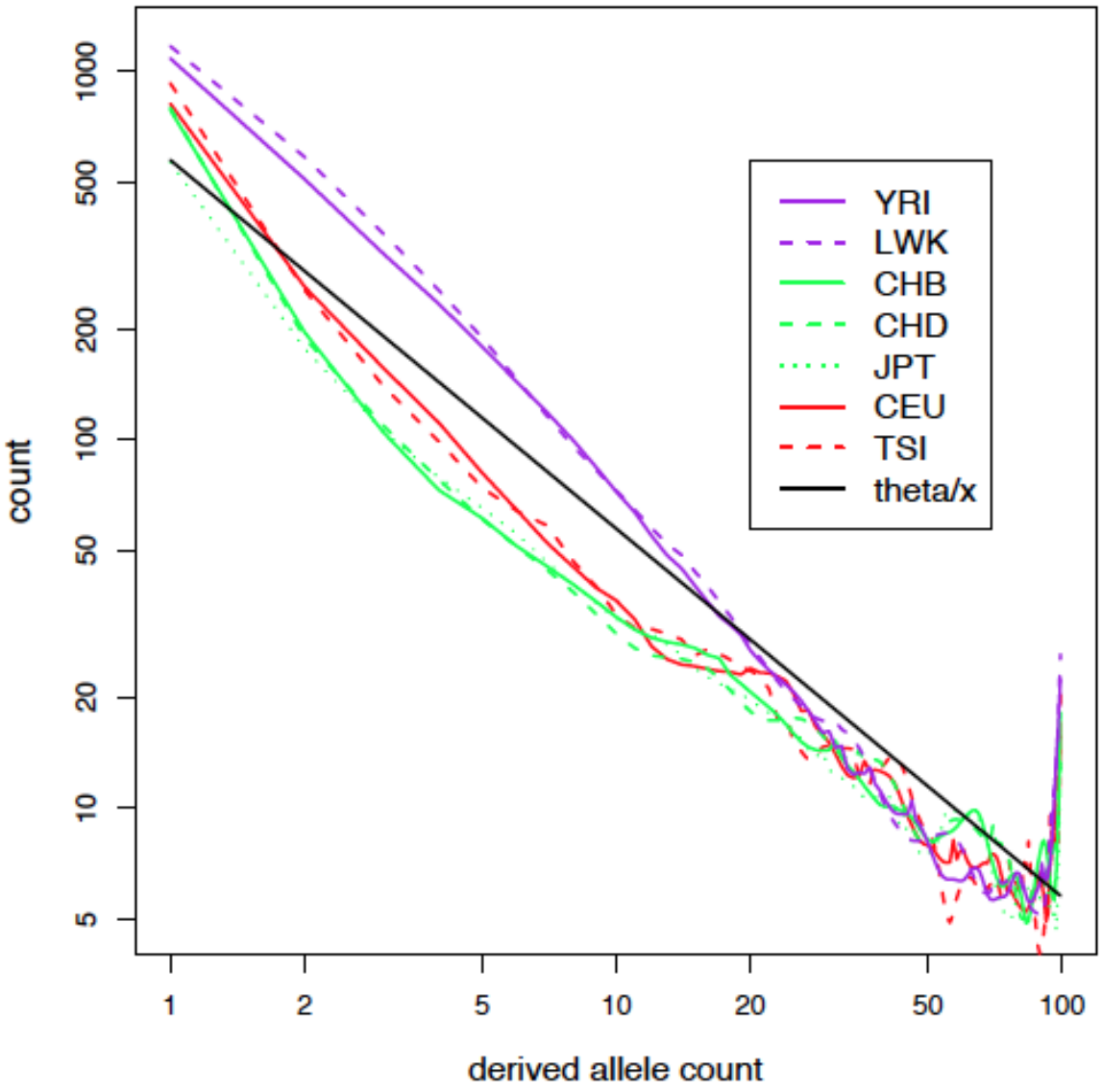


Figure 2.1: **Allele frequency spectrum of the Pilot 3 populations.** - Spectra have been downsampled to a common sample size of 100 chromosomes.

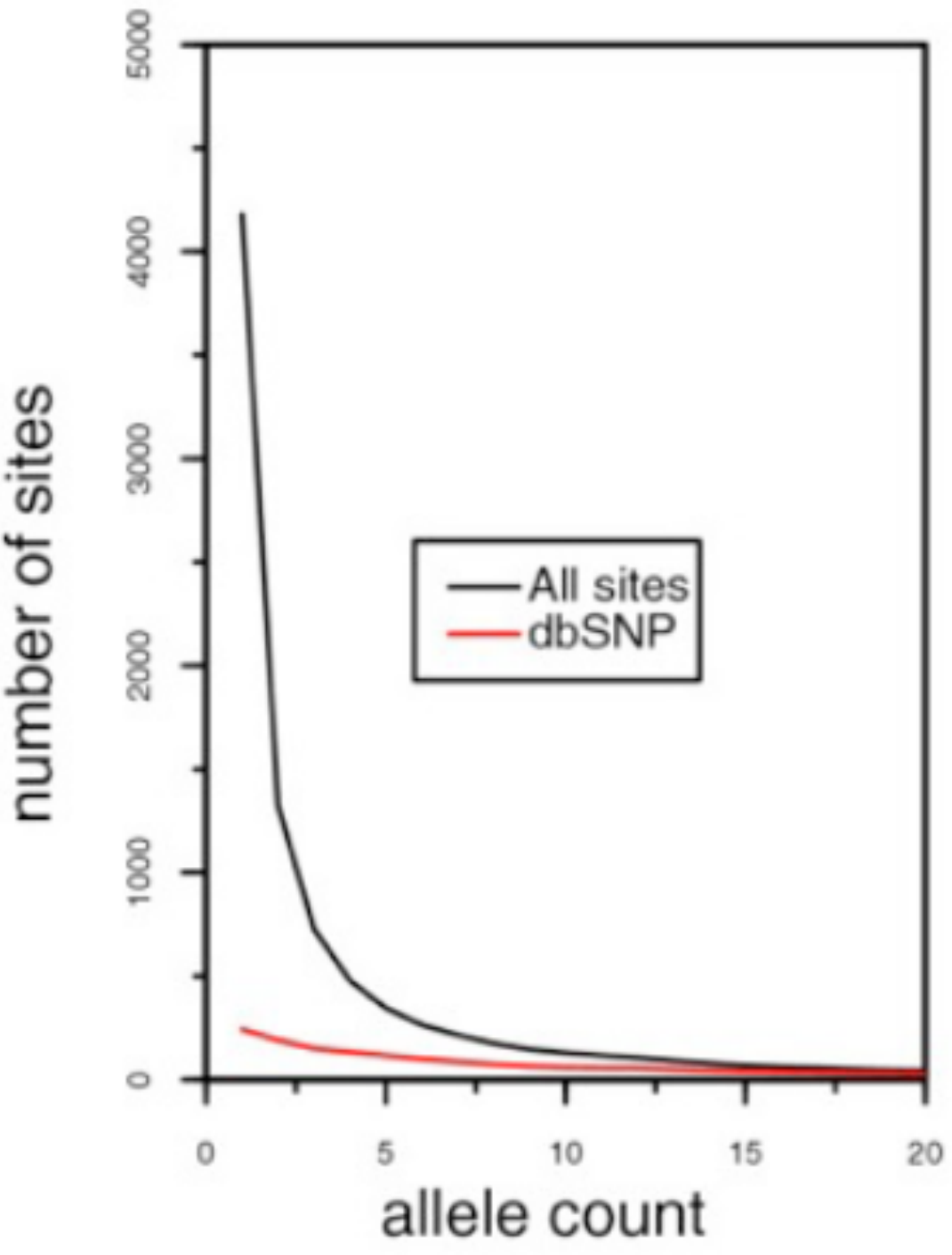


Figure 2.2: **Allele frequency spectrum of all 697 individuals in Pilot 3, focusing on only those sites with derived allele counts of 20 or less.** - Data has been partitioned to distinguish dbSNP and novel variants.

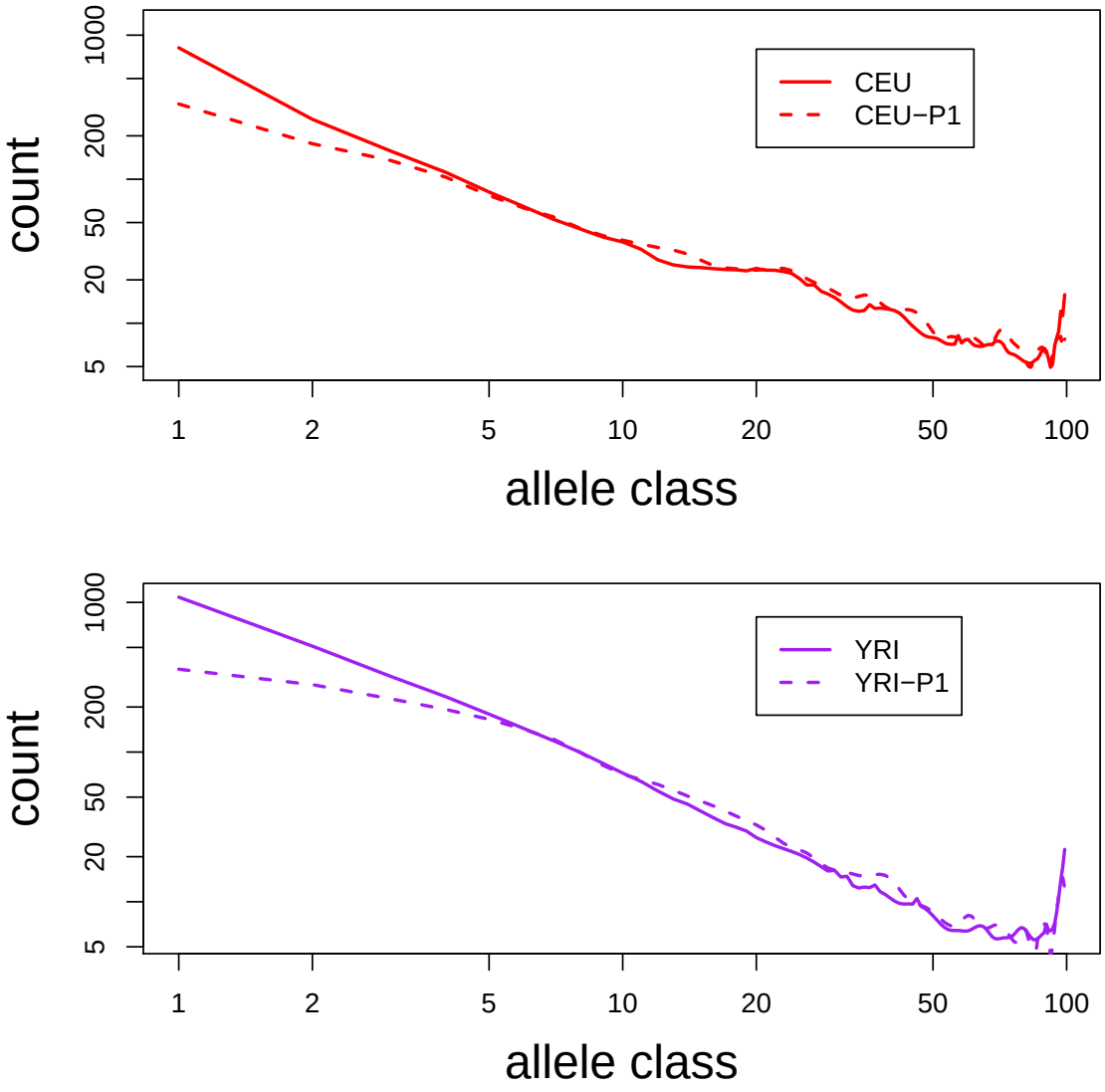


Figure 2.3: **Allele frequency spectrum the Pilot 3 and Pilot 1 datasets** - Spectra have been downsampled to a common sample size of 20 chromosomes.

Site category	total sites	YRI	LWK	CHB	CHD	JPT	CEU	TSI
All	1315794	4.42	4.52	3.34	3.35	3.26	3.54	3.50
4-fold	210575	9.24	9.16	6.60	6.63	6.43	7.12	7.04
3-fold	20990	5.01	5.41	4.24	4.39	4.60	3.59	3.59
2-fold	257486	6.04	6.16	4.44	4.42	4.37	4.74	4.68
non-synonymous	854682	2.74	2.86	2.19	2.21	2.12	2.31	2.29

Table 2.6: Per-base heterozygosity measurements at non-synonymous and n-fold degenerate sites. All values are in units of 10e-4

in Table 2.6. Overall patterns of the data indicate that heterozygosity is highest in the the African populations (YRI, LWK) and 4-fold degenerate sites exhibit the highest amounts of variation. Non-synonymous sites clearly show reduced amounts of heterozygosity, which suggests the force of negative selection constrains the amount of variation that is observed at non-synonymous sites.

2.3 Conclusions

Pilot 3 resulted in a high-quality dataset that fully characterized the spectrum of genetic variation in protein coding regions of the human genome. In terms of bioinformatics advances, the project drove the development of tools to effectively analyze capture sequencing as well as whole genome sequencing datasets [40, 85]. The main biological insights from the AFS results presented here show that there is a vast excess of singleton and low frequency variants segregating in the human genome when compared to the expected AFS from the standard neutral model of a constant sized Wright-Fisher population. A plausible explanation for this pattern is that recent, explosive population growth over the past 10,000 years has resulted in an excess of rare genetic variation. Indeed, this has been confirmed by a recent study from Clark and Keinan [70] who made demographic inferences from the Pilot 3 AFS [47, 100], as well as other re-sequencing studies [26]. More than likely, many of these newly arisen variants are mildly deleterious, suggesting a reason why heterozygosity levels

at non-synonymous sites are low. Similar patterns of excess of rare variants were also seen in the Exome Sequencing Project (ESP) [134], a much larger, whole exome re-sequencing project of 3528 individuals from European and African American ancestry. Hence, if many of these newly arisen mutations have a (mildly) deleterious affect on phenotype, natural selection has not had enough time to remove them from the population. Cataloging rare coding variants in the human genome is essential to understanding the role these variants play in complex as well as Mendelian disease [38, 24]. Pilot 3 was a pioneering study, laying down the bioinformatic groundwork for future exome re-seqeuncing studies in the genomics community.

2.4 Methods

Bioinformatics data processing

Boston College (BC) was one of the two contributors to the Pilot 3 callset. The official SNP callset release was the intersection of calls between the BC and Broad pipelines. The steps involved in BC data processing pipeline are shown in Figure 2.4. Parameter values given to the MOSAIK aligner were -act 35, -bw 37, -mhp 200, -mm 14. Base quality scores were re-calibrated with the programs CountCovariates and TableRecalibration, which are part of GATK [30]. PCR duplicate removal was performed with the program MarkDuplicates from the software package Picard [132]. The program Gigabayes is an updated version of the program PolyBayes [99], adapted for analyzing high-throughput sequencing data. It calculates genotype likelihoods and uses a genotype prior to calculate posterior genotype probabilities of samples, as well as a posterior probability of a SNP. Post-filtering of Gigabayes calls involved removing SNP variants that did not have a Phred scaled quality score of at least 40 and at least one individual with a polymorphic genotype with a Phred scaled genotype quality score of at least 10. The details of the Broad pipeline are described in [100].

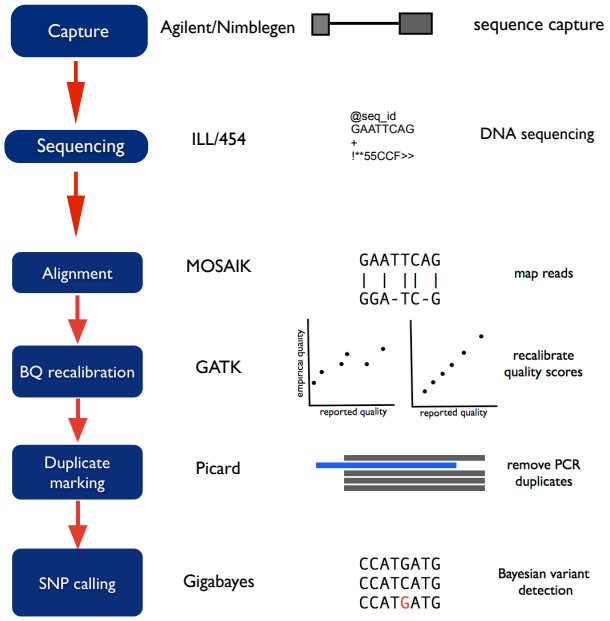


Figure 2.4: **Bioinformatics pipeline for Boston College** - The BC variant calls were produced by a four step process including alignment, base quality recalibration, duplicate marking, and variant detection.

Intersecting the Boston College and Broad callsets

The intersection of SNP callsets from the Boston College and Broad pipelines formed the official Pilot 3 release. If genotypes did not agree between pipelines for variants intersecting at the site level, those individual genotypes were filtered out. There were 4 iterations of comparison and fine tuning of pipelines during the course of the pilot. This lead to a convergence call set that formed the official release, as shown in Table 2.2.

Allele Frequency Spectrum analysis

To construct the un-folded allele frequency spectrum of the Pilot 3 callset the first step was to determine the orthologous (and ancestral) base in the panTro2 (chimpanzee) genome assembly. Next, since not all variant sites will have the same number of chromosomes sampled (either due to differing sample sizes of the populations studied, or missing data) the AFS was projected down to a common sample size of 100 chromosomes, and then plotted. This was done by using the software δaδi [48]. Projecting down the AFS involves averaging over possible re-samplings of the larger sample size to the smaller one using the hypergeometric distribution [93].

Per-base heterozygosity

A basic measure of genetic variation heterozygosity. For the Pilot 3 data heterozygosity was measured at non-synonymous, 2-fold, 3-fold, and 4-fold degenerate sites. Degeneracy was determined by the exon reading frame of target bases based on the Gencode gene model annotation [51]. To calculate heterozygosity equation 2.1 was used.

$$\sum_{i}^{m} 2p_i(1-p_i) \tag{2.1}$$

The p_i refers to the frequency of the reference allele at the ith site and m refers to the number of sites. As the number of sites approaches ∞, the result is an estimate of π, a measure of nucleotide diversity [43]. In order to be included in the analysis, a site had to have had at least 10x coverage in the MOSAIK alignments in at least 50 samples (100 chromosomes).

Chapter 3

Variant discovery in targeted re-sequencing using whole genome amplified DNA

3.1 Background

There has been considerable focus in human genetics on characterizing rare variation in the human population, and the role these variants play in human diseases to account for the "missing heritability" in genome-wide association studies using common variants [96, 95]. Until recently, the discovery of genetic variants was the rate-limiting step due to the prohibitive cost of sequencing large numbers of samples using traditional Sanger sequencing. Over the past five years, next generation sequencing (NGS) technologies have replaced traditional Sanger sequencing as the predominant method of DNA sequencing [8, 98]. The main advantage of NGS over traditional Sanger sequencing is its cheaper cost and higher throughput. NGS has had a profound impact on the field of human genetics because it is now possible to sequence large numbers of individuals to fully describe the spectrum of human

genetic variation, from common to rare variation [33]. In parallel to the developments of new sequencing technologies, improved methods have been developed to enrich specific subsets of the genome for next generation sequencing. While commonly referred to as exome sequencing, because in many cases protein coding regions have been enriched, in fact any portion of the genome can be chosen for target enrichment [58, 45]. Capture sequencing allows many individuals to be sequenced for particular regions of interest, as opposed to whole genome sequencing a smaller number samples at the same cost [128]. This also provides greater sensitivity for SNP detection compared to whole genome sequencing [21]. Exome capture sequencing has yielded many successful examples for uncovering causative mutations in Mendelian disease [106, 7], and describing the full extent of rare variation in protein-coding portions of the genome that whole genome sequencing may have missed because high-coverage, whole genome sequencing is still not common practice [100].

While the discovery of genetic variation is no longer a rate-limiting step for human genetic analysis, the application of NGS and sequence capture technologies can be limited by the amount of DNA available [83]. In particular, probands that have been collected for a clinical study maybe difficult to sample again. Previously collected DNA samples gradually decay in quality over time, and non-invasive collection techniques, such as buccal swabs, may result in insufficient amounts of DNA [83]. Several rounds of NGS or capture array sequencing may deplete original stock aliquots of samples. Whole genome amplification (WGA) is a method to overcome such challenges, and can yield micrograms of WGA DNA from nanogram starting amounts of template.

Previous studies have shown that WGA DNA performs well on high-density SNP genotyping arrays [145, 9, 54]. Three recent studies have investigated the use of WGA

DNA in NGS. Murphy et. al. [103] investigated the use of a WGA protocol performed *in situ* on laser capture micro-dissection cancer cells for the discovery of structural variants in a tumor genome using Illumina mate-pair sequencing. Tao et. al. [67] showed that WGA DNA has favorable sequence capture metrics when comparing to genomic DNA when adapting the NimbleGen capture array for use on the Illumina GA sequencing platform. El Sharawy et. al. [34] investigated the use of WGA DNA in a NGS microdroplet-based PCR sample enrichment pipeline experiment of 384 exons with 3 HapMap samples and showed there was strong genotype concordance with both genomic and WGA DNA SNP calls to HapMap III genotypes. In this paper we describe the results of variant calls using WGA DNA for a single sample for two separate capture sequencing experiments on the Agilent SureSelect platform, and compare them to variant calls made with genomic DNA for the same samples. While the results in this study are based on a limited number of samples, our results suggest that WGA samples have a high sensitivity in detecting variant alleles identified with genomic DNA, and can be used effectively in re-sequencing studies.

3.2 Results and Discussion

Capture metrics of WGA and genomic DNA

We analyzed capture sequencing metrics of genomic and WGA sample pairs for two capture experiments, a chr12 custom array and a whole exome capture array. Tables 3.1 and 3.2 contain capture metrics from the program CalculateHsMetrics from the software package Picard [132]. The average target coverage for the whole exome capture experiments were 92x (WGA) and 80x (genomic). The average target coverage for the chr12 capture experiments were 432x (WGA) and 224x (genomic). WGA samples in both capture experiments had a higher number of PF (passed filter) reads thus higher average target because they were sequenced in a separate flow-cell lane,

dataset	chr12 WGA	chr12 Genomic	chr12 WGA subset
Read Length	101	101	101
Target territory	3871678	3871678	3871678
PF reads	57462846	33441588	33441588
PF unique reads	2688176	13002406	19870208
PF unique reads aligned	23607248	11566780	17673567
% Selected bases	87	83	87
% Usable bases on target	32	26	40
Mean target coverage	432	224	342
% Target bases 2x	98	98	98
% Target bases 10x	97	97	97
% Target bases 20x	97	96	96
% Target bases 30x	96	95	96

Table 3.1: Sequencing capture metrics of chr12 genomic, WGA, and WGA subset experiments

dataset	whole exome WGA	whole exome genomic	whole-exome WGA subset
Read Length	101	101	101
Target Territory	49649722	49649722	49649722
PF reads	258222898	105316652	105316652
PF unique reads	93557436	79549416	62080571
PF unique reads aligned	70593742	62036624	47366440
% Usable bases on target	17	35	28
Mean target coverage	92	80	63
% Target bases 2x	92	92	91
% Target bases 10x	85	86	82
% Target bases 20x	79	81	74
% Target bases 30x	74	75	65
% Selected bases	83	83	83

Table 3.2: Sequencing capture metrics of whole exome genomic, WGA, and WGA subset experiments

while the genomic DNA samples were multiplexed. For both sequencing experiments a large percentage of reads were marked as duplicates, as the percentage of usable bases on target for each of the capture experiments does not exceed 40%. Despite the high duplicate read fraction both samples in the whole exome capture experiment had 80% of targeted bases with at least 20x coverage. For the smaller chr12 capture experiment, over 90% of targeted bases had at least 20x coverage.

Since the WGA capture experiments had a larger sequencing library compared to the genomic, a random subset of reads were selected from the starting fastq files to match the number of PF reads of the genomic sequencing library (see Tables 3.2, 3.1, and

Dataset	chr12 WGA	chr12 Genomic	whole-exome WGA	whole-exome Genomic
SNPs	4642	4592	29600	30316
dbSNP %	98.4	98.6	98.6	98.6
TsTv overall	2.42	2.41	2.83	2.82
TsTv novel	1.47	1.48	1.81	1.89
TsTv known	2.44	2.43	2.85	2.84
INDELs	491	482	2197	2215
dbSNP %	34.8	34.0	34.2	34.8

Table 3.3: Variant callset summary of whole exome and chr12 experiments

Section 3.4). The average target coverage for the chr12 WGA subsetted BAM (342x) is higher than the chr12 genomic experiment, even though the starting number of PF reads is the same. This can be attributed to higher percentage of usable bases on target, as calculated with HsMetrics. Similarly, the whole-exome WGA subsetted BAM average target coverage (63x) is less than the genomic sample, despite starting with the same number of PF reads. The percent usable bases on target are lower in the whole-exome WGA subset than the whole exome genomic sequencing experiment.

Next, we explored the relationship between GC% and median target coverage for both capture experiments. Previous studies have shown that lower sequencing coverage occurs in regions with high GC% [31]. GC% of targets for each capture experiment was calculated. Next, the targets were placed in four bins according to the first, median, and third quartiles of capture target GC%, based on the boxplots shown in Figure 3.1. In addition to boxplots of GC% of capture targets of the two experiments, Figure 3.1 shows the GC% of the whole genome and chr12 for comparison. Targets were placed in the appropriate bin and within each bin, a box plot of median target coverage was made for genomic and WGA DNA, as shown in Figure 3.2. The results show that for genomic DNA, chr12 capture targets in the fourth bin (with GC% greater 51%) have lower coverage than targets in the other three bins. For the corresponding WGA DNA, targets in the first (GC% less than 38%) and fourth bins have a similar distribution of median target coverage. Whole exome capture targets in the fourth bin (GC% greater 59%) had lower amounts of coverage than targets with lower GC%

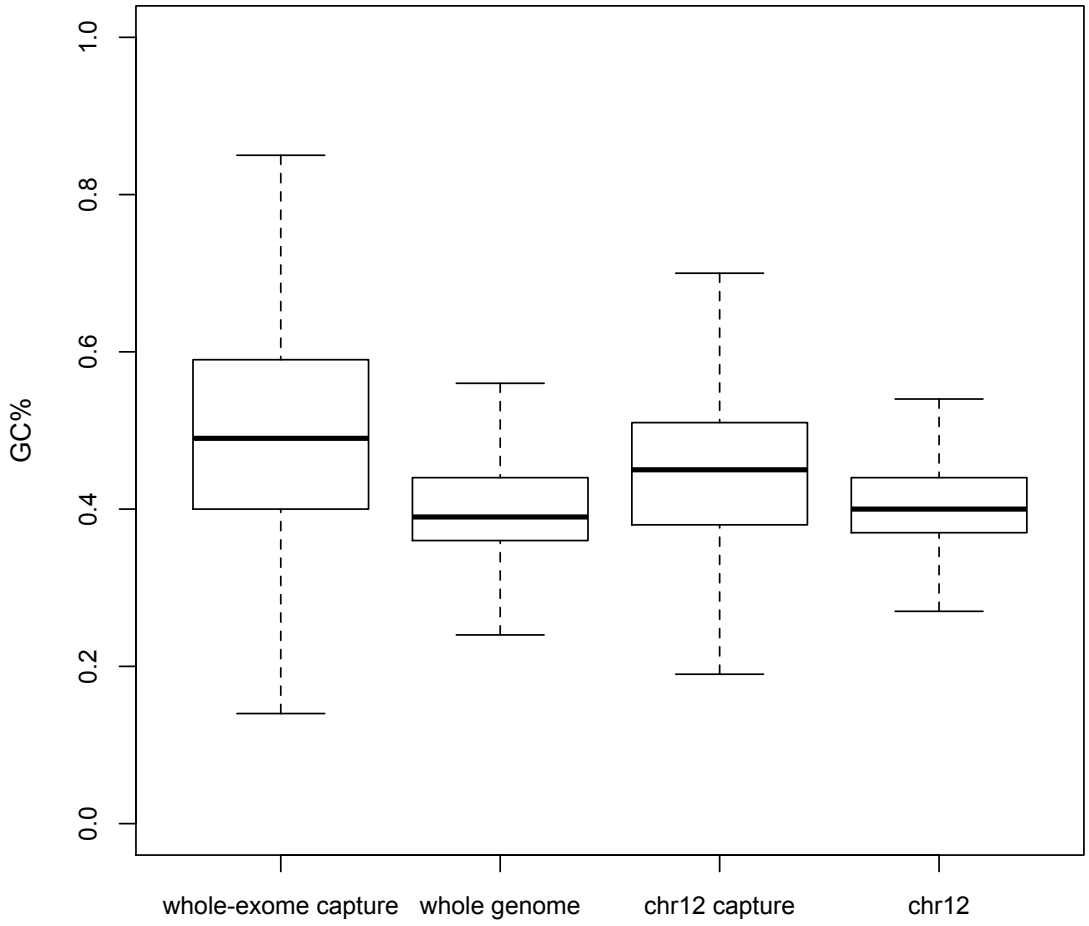


Figure 3.1: **Boxplot of target GC percentage** - Boxplot of GC percentage of whole exome and chr12 capture targets as well as genome wide and chromosome 12 wide GC percentage

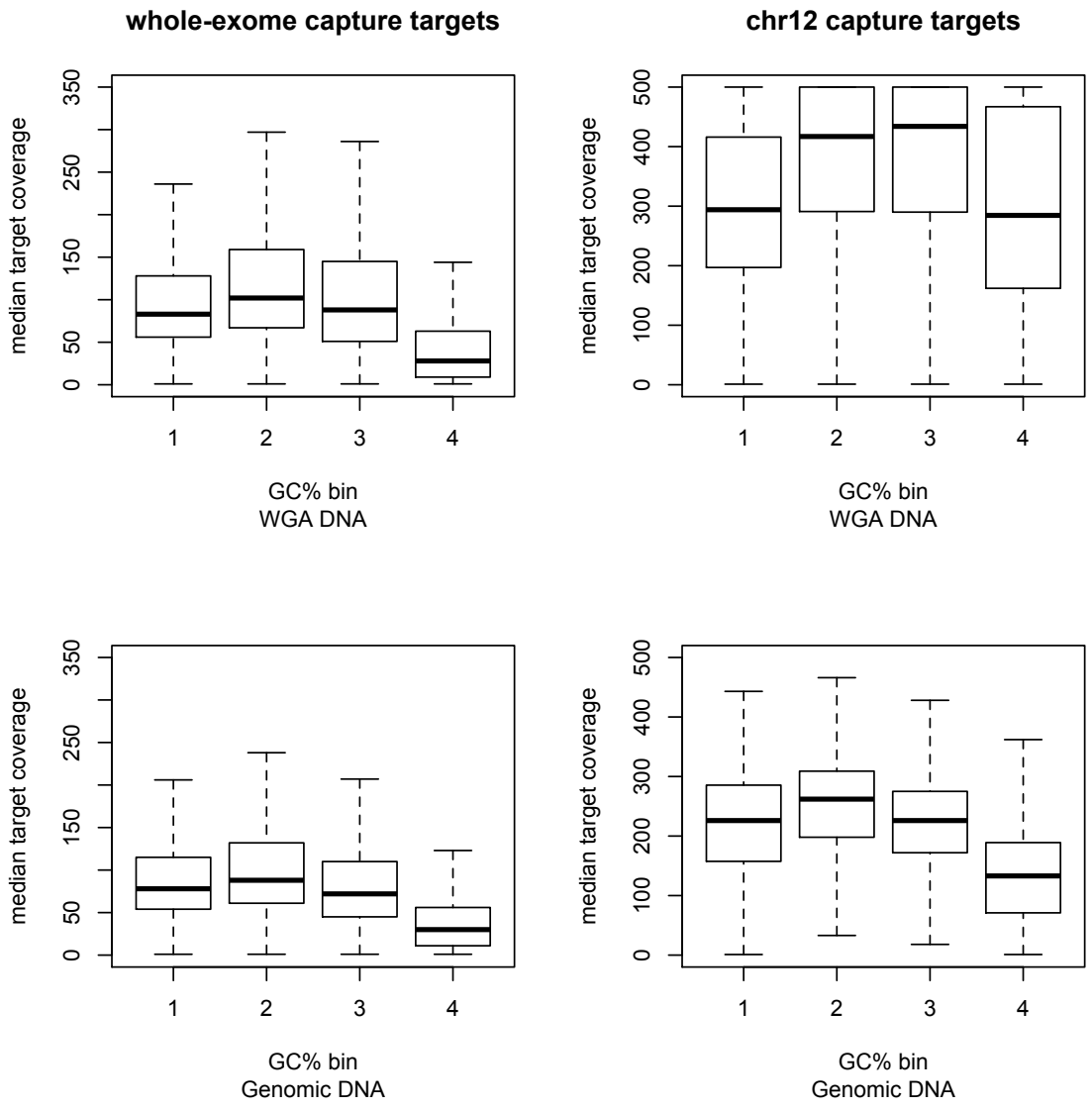


Figure 3.2: **Boxplots of median target coverage** - Boxplots of median coverage of targets binned according to quartiles of GC% of capture targets for chr12 and whole-exome capture experiments.

for both genomic and WGA samples. Since the chr12 capture targets were over a much smaller interval (3.87 Mbp), its harder to make any definitive statement regarding GC% and lower sequencing coverage, but the patterns of coverage seen in both capture experiments examined here are in line with previous studies [31, 21].

Overall variant counts and Venn analysis

Table 3.3 shows the counts and callset metrics of the individual SNP and INDEL callsets, after post-call filtering (described in Section 3.4). For all SNP callsets the dbSNP fraction is 98%. The overall transition-transversion (TsTv) ratio for the WGA and genomic chr12 callsets are 2.42 and 2.41 respectively. The overall TsTv ratio for the WGA and genomic whole-exome callsets are 2.83 and 2.82, respectively. The TsTv values of novel SNPs found in each of the capture experiments is considerably reduced, suggesting these may be false positive calls.

We performed Venn analysis of the WGA and genomic callsets to see how variants overlapped based on coordinate intersection. Figure 3.3 shows four Venn diagrams for SNP and INDEL sites in each of the capture experiments. Visual inspection indicates there is a high fraction of site-level concordance of SNP calls, with 97% and 99% of the union of SNP sites lying in the intersection for the whole exome and chr12 capture callsets. Slightly lower numbers of 87% and 90% were found for INDEL sites. Overall TsTv ratios for SNPs in the intersection were similar to those calculated for each individual callset. TsTv ratios of novel sites were slightly higher in the intersection, when compared to the original callsets. The TsTv values of the genomic and WGA unique fractions for the whole-exome capture experiment are considerably lower, suggesting these are lower quality calls. The unique fractions of the chr12 capture experiment are much smaller, making it difficult to interpret the differences in value of their TsTv ratios.

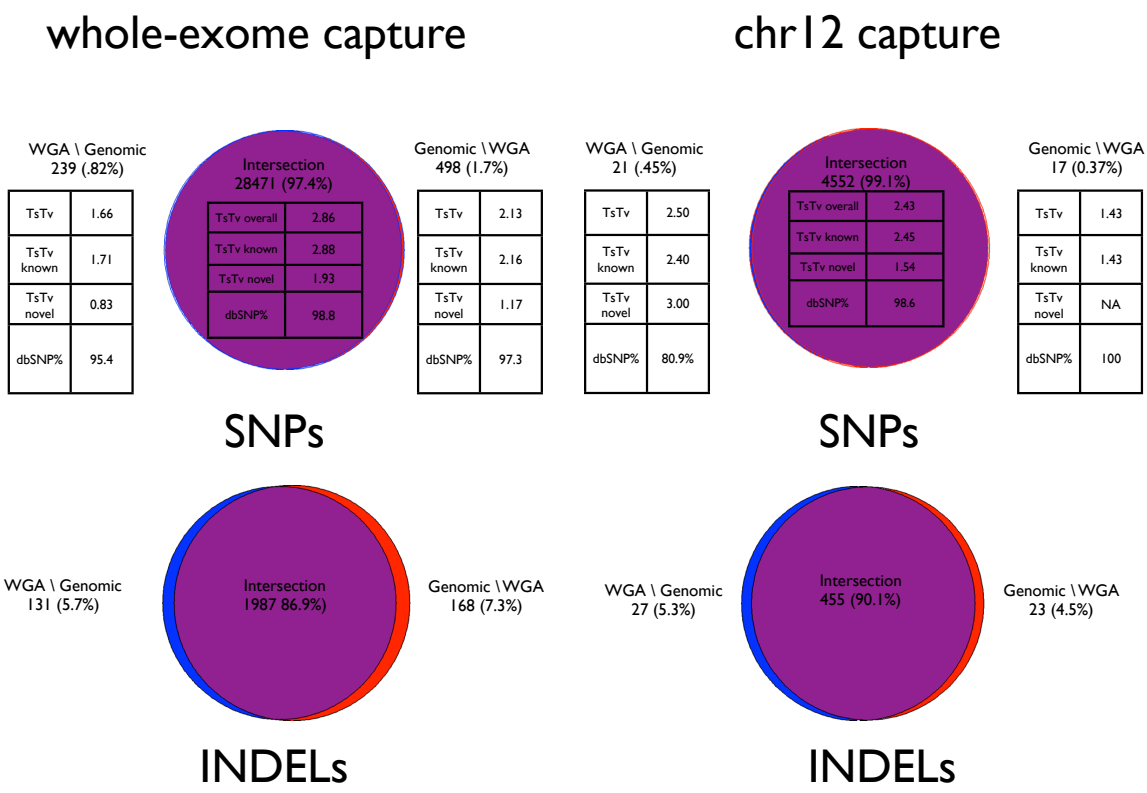


Figure 3.3: **Venn diagrams of SNP and INDEL variant calls** - Venn diagrams of SNP and INDEL variant calls. The top row also shows TsTv ratios and dbSNP fractions of SNPs in each portion of Venn diagram.

Downsampling alignments and subsetting reads

Since the WGA samples were run as a single lane but the genomic samples were multiplexed, we downsampled reads from each BAM to examine the effect of coverage on the numbers of discovered variants. A total of 100 bootstrap sub-samples of reads were performed (see Section 3.4). In addition to downsampling the reads from the aligned BAM file, a subset of fastq reads were chosen at random to match the starting number PF reads in the WGA library for both experiments (see Tables 3.1 and 3.2).

Figure 3.4 shows the median number of variants discovered as a function of average target coverage for SNPs and INDELs, for each capture experiment. The randomly

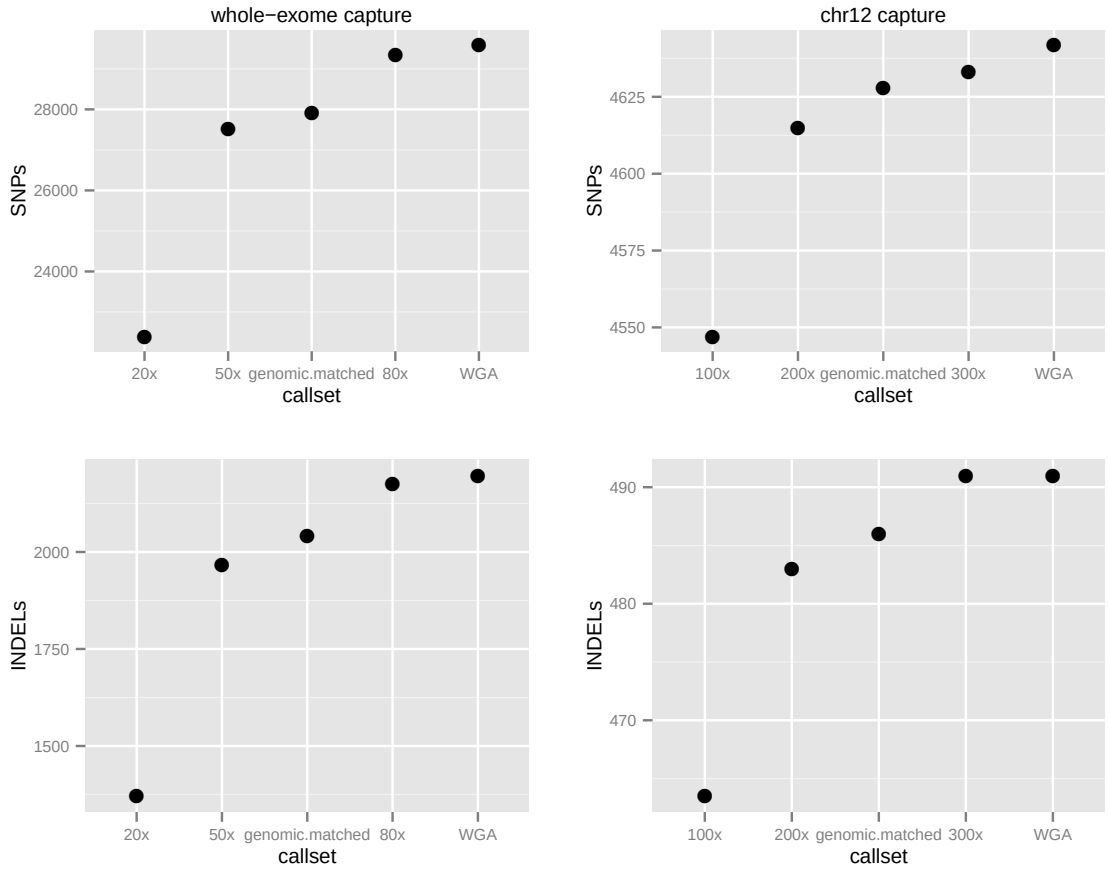


Figure 3.4: **Numbers of variants discovered in downsampled and subsetted WGA BAMs** - Median number of SNPs and INDELs called from 100 bootstrap subsampled BAM files from whole-exome and chr12 capture experiments for WGA DNA samples. Plot also includes number of variants discovered in the WGA subsetted BAM that matched the starting read count of the genomic sample.

Dataset	NRS	NRD
whole-exome capture SNPs	98.28	0.63
whole-exome capture INDELs	91.17	13.46
chr12 capture SNPs	99.63	0.29
chr12 capture INDELs	94.07	10.7

Table 3.4: NRS and NRD values for WGA derived whole-exome and chr12 capture callsets when comparing to genomic derived callsets.

chosen subset of WGA reads to match the number of PF reads in the genomic sequencing experiment is shown figure as genomic.matched on the x-axis, and sorted in ascending order of target coverage. As expected, downsampling BAMs reduces the number of called variants, with the original WGA BAM having the largest number of called variants. The datapoint that most closely matches the target coverage of the non-WGA sample is 80x for the whole-exome plot. The median number of SNPs and INDELs found (29350 and 2174) closely match the numbers of variants found in genomic derived variant calls listed in Table1. The datapoint that most closely matches the target coverage non-WGA sample is 200x for the chr12 plot. The median number of SNPs and INDELs found (4615,483), again closely match what was found in the genomic derived calls listed in Table 3.3.

Genotype concordance

We used two measures of genotype concordance, non-reference sensitivity (NRS) and non-reference discrepancy (NRD) [30, 72], shown in Figure 3.5, to compare genotypes made with WGA and genomic DNA. NRS measures the proportion of sites called variant in the comparison callset (genomic) that are also called variant in the evaluation callset (WGA). NRD measures the proportion of differing genotypes between the WGA and genomic callsets, at sites called in both data sets, excluding concordant homozygous reference calls.

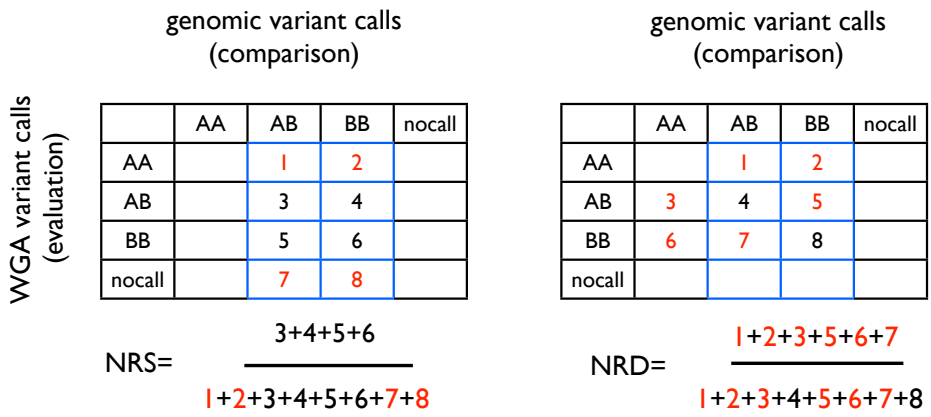


Figure 3.5: **Calculating NRS and NRD genotype concordance metrics** - Illustrates how the metrics of non-reference discrepancy (NRD) and non-reference sensitivity (NRS) are calculated

The NRS and NRD values for SNPs and INDELs for each capture experiment are shown in Table 3.4 and the concordance matrices from which they were calculated are shown in Figure 3.6. For the chr12 capture experiment, of the 17 sites that contribute to the decrease in SNP NRS of the WGA call set, six are heterozygous sites in the genomic DNA that were not called in WGA DNA. Of the 28 sites contributing to the decrease in INDEL NRS, 18 were heterozygous genotypes in genomic DNA, that were evenly split as homozygous reference or no calls in WGA DNA. For the 13 sites contributing SNP NRD, eight were WGA heterozygous sites, called homozygous non-reference in genomic DNA. The greatest contribution to INDEL NRD came from sites that were called heterozygous in WGA DNA, but homozygous reference

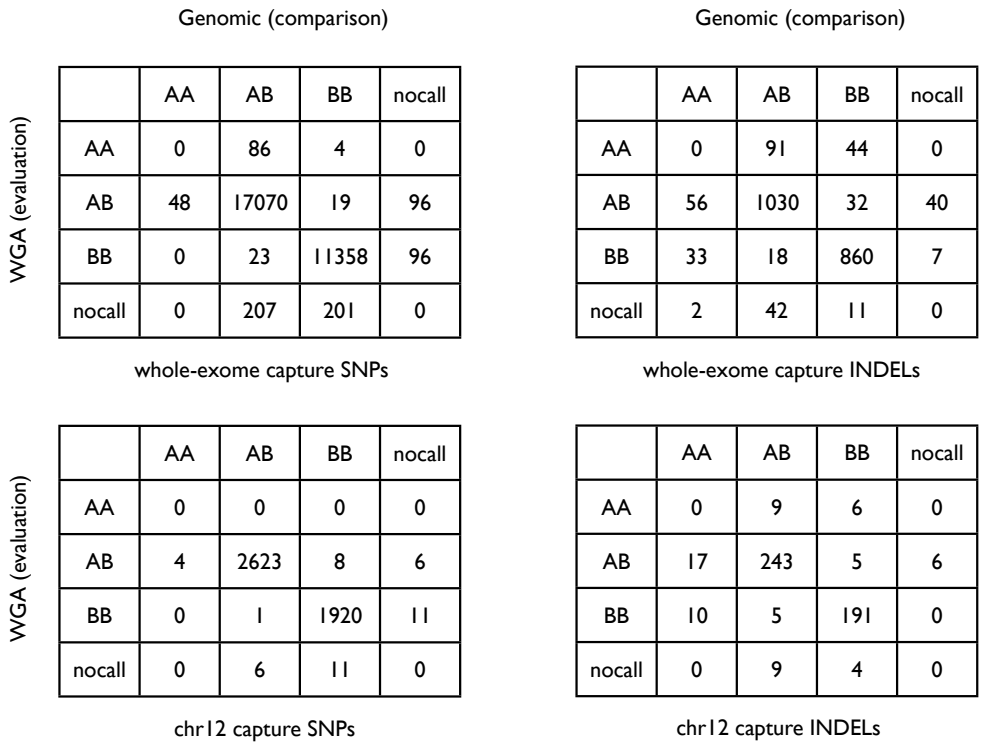


Figure 3.6: **Genotype concordance matrices** - Genotype concordance matrices for chr12 and and whole-exome SNP and INDEL callsets from which concordance metrics of NRS and NRD were calculated from.

in genomic DNA.

Next, genotype concordance for each bootstrap downsampled chr12 capture BAM was calculated by comparing its calls to the ones made from the original genomic BAM file. NRS and NRD values were summarized by calculating their median value across all 100 downsampled BAMs. In addition, NRS and NRD of the subsetted WGA BAM was calculated by comparing its genotypes to the original genomic BAM. Figure 3.7 shows the effect of downsampling and subsetting on genotype concordance metrics. Unexpectedly, two of the three downsampled datasets have slightly higher SNP and INDEL NRS values than the original WGA callset. This

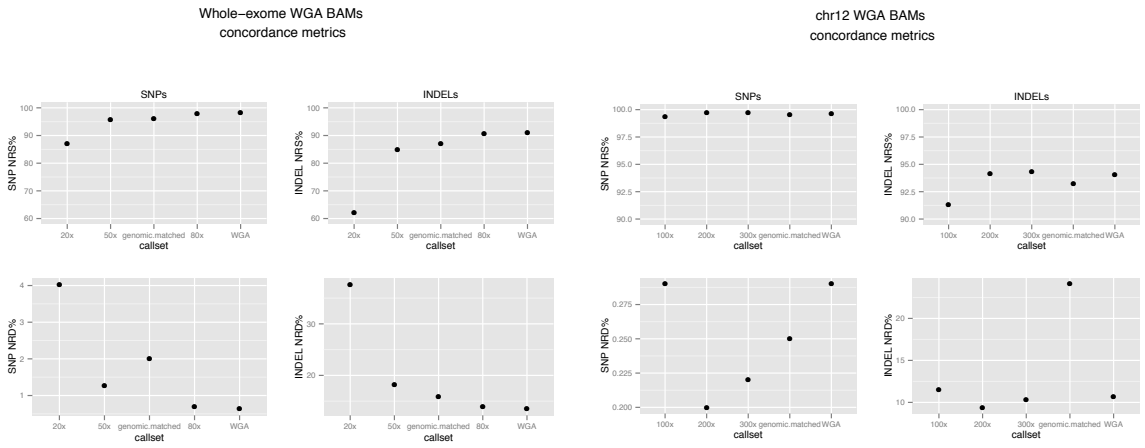


Figure 3.7: **Genotype concordance metrics of downsampled subsetted WGA BAMs** - Median values of NRS and NRD metrics for SNP and INDEL variants calculated from 100 bootstrap sub- sampled BAM files from whole-exome and chr12 capture experiments for WGA DNA samples. Plot also includes NRS and NRD metrics of the WGA subsetted BAM that matched the starting read count of the genomic sample.

includes the NRS of the 200x downsampled BAM, which most closely matches the coverage of the genomic sample. Similarly, the original WGA callset has a higher NRD values than some of the lower coverage, downsampled BAMs (including the 200x downsample BAM). The INDEL NRD for the genomic matched WGA BAM is clearly an outlier on the graph. This might be attributed to sampling error, but since the WGA fastq files were subsetted only once, it's difficult to say. This unexpected pattern can potentially be attributed to the smaller capture interval in the chr12 experiment and the fewer numbers of variants called, as the relationship between concordance metrics and lower coverage, downsampled BAMs is clearer in

the whole-exome capture experiment (see below). Also, since a technical replicate of genomic sequencing was not performed, it is difficult to ascertain what the expected genotype discrepancies should be between genomic and WGA derived variant calls.

The NRS and NRD values and the genotype concordance matrix from which they were calculated for the whole-exome capture experiment are also shown in Table 3.4 and 3.6, respectively. Of the 498 sites that contribute to the decrease of SNP NRS of the WGA call set, the majority come from sites either called heterozygous or homozygous non-reference in genomic DNA but were no calls in WGA DNA. The majority of sites contributing to the decrease of INDEL NRS come from sites called heterozygous in genomic DNA, but called homozygous reference in WGA DNA. Sites contributing most to SNP NRD are heterozygous calls in genomic DNA, called homozygous reference in WGA DNA, for both SNP and INDEL variants. The concordance metrics of the WGA whole exome downsampled BAMs to original genomic DNA calls, also shown in 3.7, reinforce the intuitive expectation that the lower coverage WGA callsets result in higher NRD and lower NRS values. The one exception is the SNP NRS of the genomic matched subsetted WGA BAM, which had a NRD value of 2%. This could be attributed to sampling error, since the subsetting was only performed once, and not multiple times like the downsampling. The SNP and INDEL NRS of the downsampled 80x BAMs, which match the average coverage of the genomic BAM, are only slightly lower than the original WGA BAM. Also, the SNP and INDEL NRD values are slightly higher than the original WGA BAM. Still, in each comparison, the original WGA call set had the lowest NRD and highest NRS values relative to lower coverage downsampled and subsetted callsets. As with the chr12 experiment, the genomic sequencing was not repeated, so it difficult to quantify the expected genotype discrepancies and sensitivity of the WGA derived variant calls.

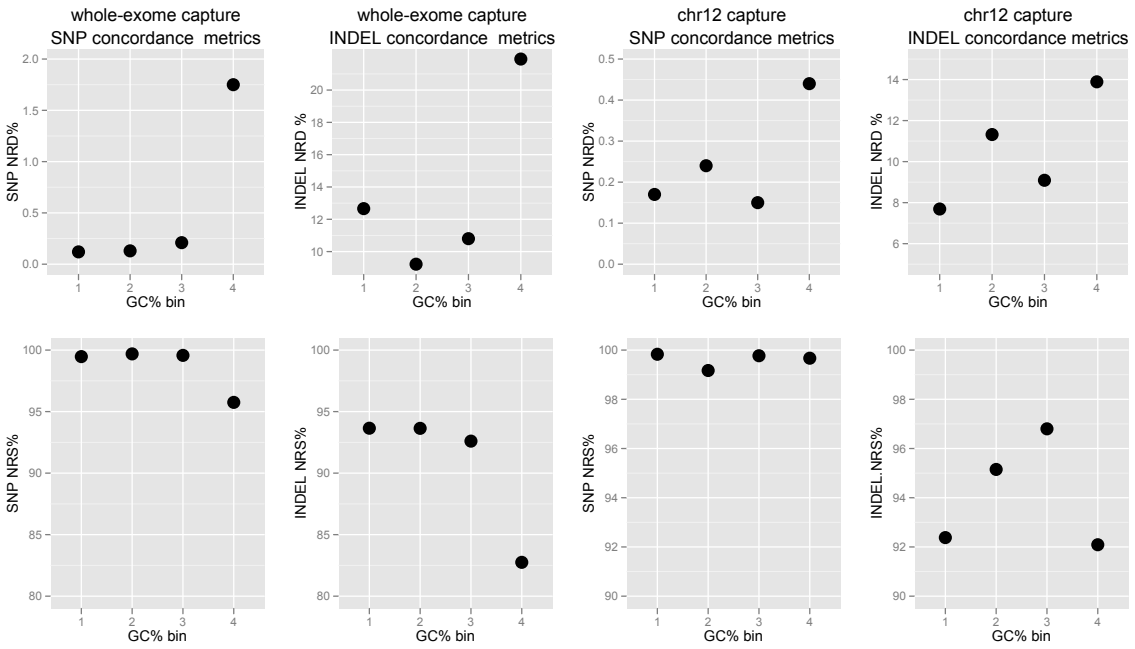


Figure 3.8: **Genotype concordance metrics as a function of GC%** - NRS and NRD of variants binned according to GC% based on quartiles of GC% in capture targets. First row shows is NRD values for SNP and INDELs and second row shows NRS values of SNPs and INDELs for each of the chr12 and whole-exome capture experiments.

Targets with higher amounts of GC% have lower amounts of median target coverage for both capture experiments and both types of DNA, as described above. Figure 3.8 shows NRS and NRD metrics for each bin, based on GC% of targets. For the original WGA whole-exome callset, the greatest number of genotype discrepancies and lowest detection sensitivities, for both SNP and INDEL variants, occur in targets with the highest GC%. The patterns are less clear for the original WGA chr12 callset, again most likely attributable to the smaller size of capture region. For both SNP and INDEL variants, the greatest numbers of genotype discrepancies are in targets with the highest GC%. The pattern is less clear for variant detection sensitivity, INDELs in target regions with the highest amount of GC% have the lowest sensitivity, but this is not true for SNPs.

Allele bias in SNP variant calls

To investigate whether there is any evidence of allele bias in SNP variant calls, all calls from the original chr12 and whole-exome WGA datasets were divided into four groups: concordant genotypes, unique genomic calls (these are sites that contribute to NRS), discordant genotypes (these are sites that contribute to NRD), and WGA unique. In each group, the percentage of each six possible reference/alternate allele combinations was calculated. The results are shown in Figure 3.9. We tested to see if there were statistically different proportions of each reference/alternate allele combinations (see Section 3.4) between the four groups. The resulting p-values are in the appendix. For the whole-exome capture SNPs, there was a significant difference in proportion between concordant CG SNPs and each of the three other categories. Also, there was a significant difference in proportion of GT SNPs between concordant and WGA unique categories. For the chr12 capture set there was no significant difference in proportion of SNPs between any of the four categories for each of the 6 different allele combinations. The interpretation of the statistical analysis of allele bias must be tempered by the fact that the analysis is based on a small sample size of matched genomic / WGA samples, lack of technical replicates,and the reduced target region for the chr12 capture experiment. But even with this in mind, results suggest that allele bias does not play a significant role in SNP variant discovery with WGA DNA.

Validation of SNP variant calls

Sequencing derived SNP variant calls were validated by comparing genotypes to Affymetrix 6.0 Human SNP array genotypes for the same sample. The 6.0 array has over 900,000 variants covering the whole genome, hence only those array genotypes that overlapped a capture target interval were examined. For the whole-exome capture array there were a total of 11831 overlapping SNPs and for the custom chr12 capture array there were a total of 1435 overlapping SNPs. See Section 3.4 section for

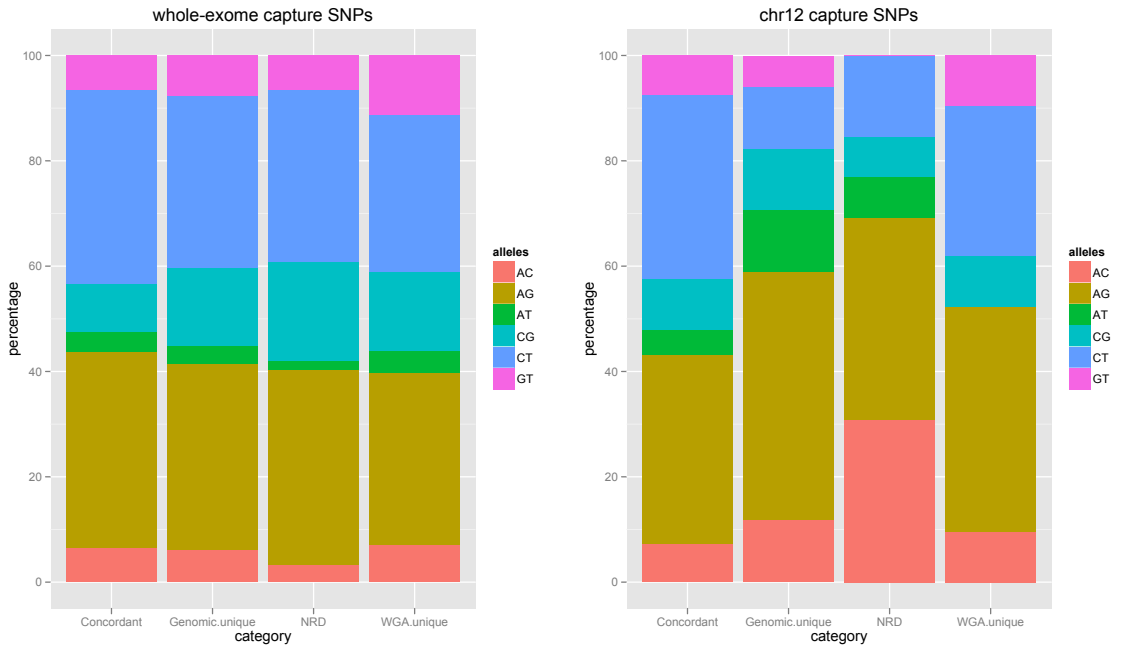


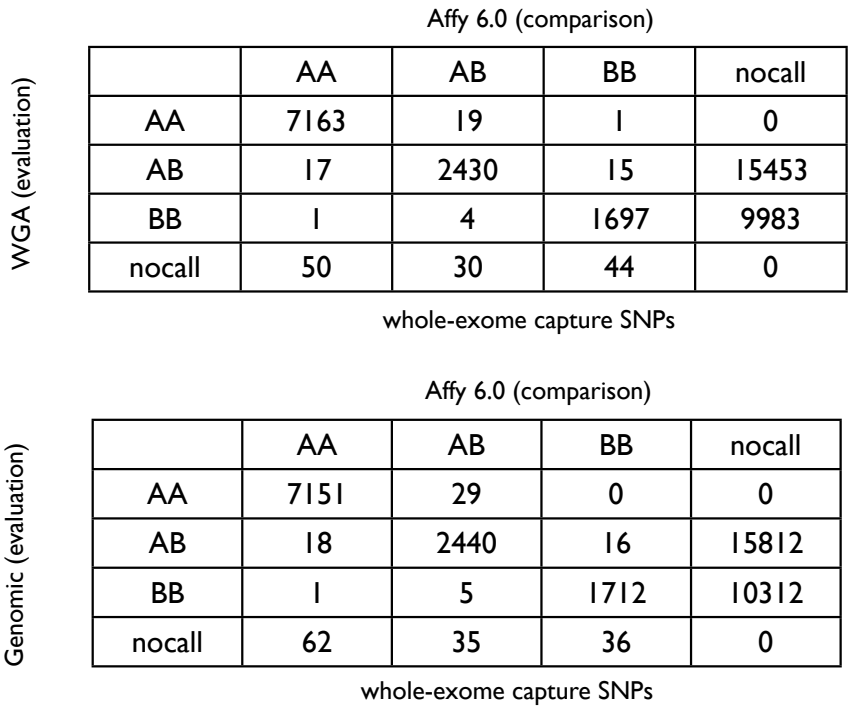
Figure 3.9: **Allelic proportions of SNPs in whole-exome and chr12 capture experiments** - Allelic proportions of whole-exome and chr12 capture SNPs in each of four categories: concordant genotypes, unique sites in original genomic and WGA call sets, and discordant genotypes contributing to NRD for chr12 and whole-exome capture experiments when comparing WGA derived SNP genotypes to genomic DNA callset.

more details. The NRS and NRD metrics of the WGA and genomic sequencing based SNP genotypes when compared to Affymetrix 6.0 SNP array genotypes for both capture experiments is shown in Table 3.5. The NRD for the WGA whole-exome capture sequencing derived genotypes when compared to the SNP array genotypes is 1.3% and the NRS value is 97.78%. The NRD for the genomic whole-exome capture sequencing derived genotypes when compared to the SNP array genotypes is 1.6% and the NRS value is 97.66%.

The concordance matrix for the WGA whole-exome comparison to capture array genotypes is shown in the top panel and the genomic concordance matrix is shown in the bottom panel in Figure 3.10. For sites that contribute to a decrease in whole-exome capture NRS, the read coverage and pileup of bases was investigated. For the

Dataset	NRS	NRD
WGA whole-exome capture SNPs	97.78	1.30
Genomic whole-exome capture INDELs	97.66	1.30
WGA chr12 capture SNPs	82.60	22.20
Genomic chr12 capture SNPs	83.00	22.60

Table 3.5: NRS and NRD values for WGA derived whole-exome and chr12 capture callsets when comparing to Affymetrix 6.0 derived genotypes.



	AA	AB	BB	nocall
AA	7163	19	1	0
AB	17	2430	15	15453
BB	1	4	1697	9983
nocall	50	30	44	0

	AA	AB	BB	nocall
AA	7151	29	0	0
AB	18	2440	16	15812
BB	1	5	1712	10312
nocall	62	35	36	0

Figure 3.10: **Affymetrix genotype concordance matrices whole exome** - Genotype concordance matrices of WGA and genomic DNA SNP calls to Affymetrix genotypes for the whole exome capture experiment.

94 sites in the WGA whole-exome capture call set that contribute to a decrease in NRS, 20 had minimal coverage and were called homozygous reference. The remaining sites have an overwhelming majority reads with mapping quality 0 spanning the SNP position and were not called. Similarly, for the 100 sites that contribute to the decrease in NRS in the genomic DNA whole-exome capture derived genotypes, 29 had minimal coverage and were called homozygous reference. The remaining sites had reads spanning the SNP position with mapping quality values of zero and not called. There are a total of 68 SNP positions common to both WGA and genomic callsets that contribute to a loss of NRS when comparing the Affymetrix SNP array genotypes.

Affy 6.0 (comparison)

WGA (evaluation)

	AA	AB	BB	nocall
AA	565	137	0	0
AB	0	417	5	2269
BB	0	47	229	1675
nocall	19	7	3	0

chr12 capture SNPs

Affy 6.0 (comparison)

Genomic (evaluation)

	AA	AB	BB	nocall
AA	565	138	0	0
AB	0	416	4	2227
BB	0	47	231	1667
nocall	19	5	1	0

chr12 capture SNPs

Figure 3.11: **Affymetrix genotype concordance matrices chr12** - Genotype concordance matrices of WGA and genomic DNA SNP calls to Affymetrix genotypes for the chr12 capture experiment.

The NRS and NRD values when comparing the WGA, chr12 capture sequencing SNP genotypes to the SNP array genotypes are 82.6% and 22.3%. The NRS for the genomic DNA, chr12 capture SNP genotypes when compared to SNP array genotypes is 83% and the NRD is 22.6%. The concordance metrics for the chr12 custom array SNP genotypes when compared to the SNP array derived genotypes is shown in 3.11. The top panel shows the concordance matrix for the WGA chr12 capture array and in the bottom panel is the genomic chr12 concordance matrix. For both comparisons, the majority of sites that contribute to the loss of sensitivity in the sequencing derived SNP calls are sites that were called heterozygote on the genotyping array. Careful visual inspection and examination of read pileups in the WGA and genomic BAM files revealed no evidence of an alternate allele and hence were called homozygous reference. There are total of 144 SNP position common to both WGA and genomic call sets that contribute to a loss NRS when comparing to the Affymetrix SNP array genotypes.

Allele bias in SNP variant validation calls

To investigate if there were any biases in the comparisons of the sequencing derived genotypes to the Affymetrix array based genotypes the percentage of each six possible reference/alternate allele combinations was calculated in sites that contributed to concordant, NRS, and NRD categories. The results are shown in Figure 3.12. To test if there were statistically different proportions of each reference/alternate allele combinations between groups we applied the same pairwise.fisher.test when comparing the WGA derived SNP calls to the genomic derived SNP calls (see Section 3.4). The resulting p-values of the analysis are in the Appendix. The only significant differences in proportion detected were AT SNPs when comparing the chr12 genomic and whole exome capture calls to the corresponding Affymetrix array derived genotypes.

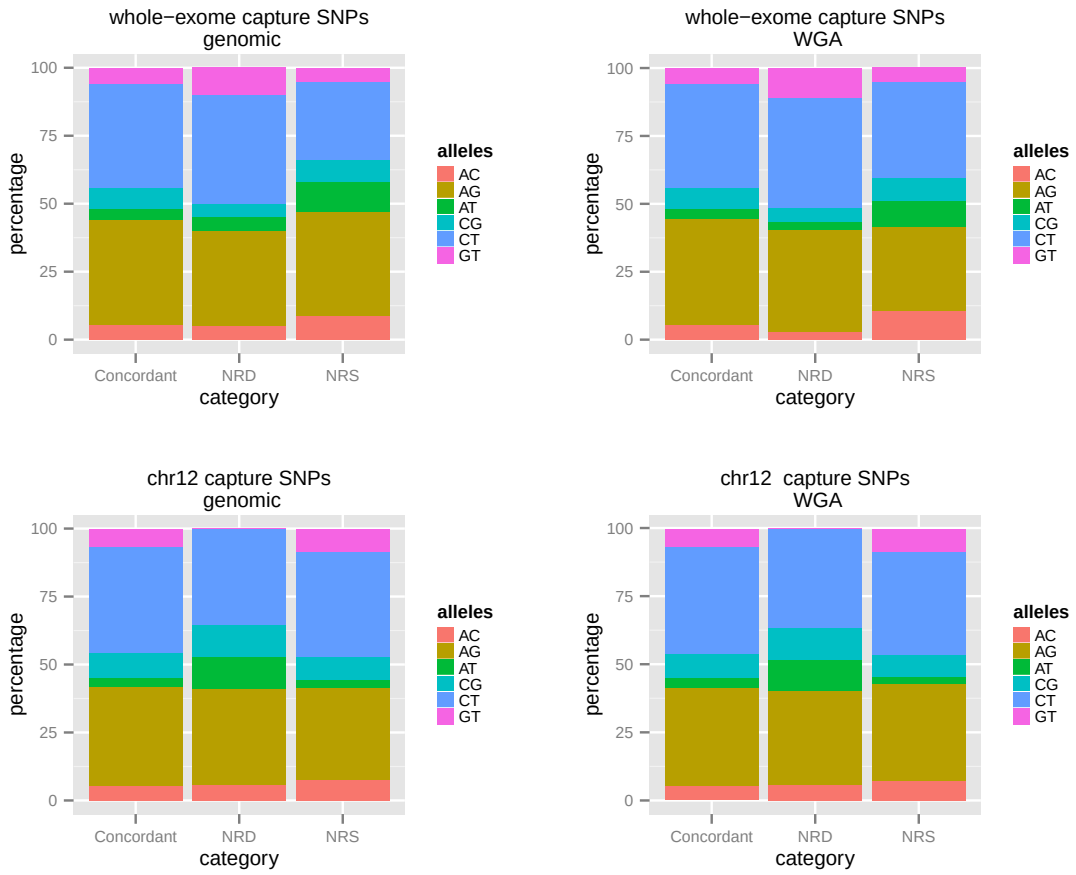


Figure 3.12: **Allelic proportions of Affymetrix SNPs** - Allelic proportions of whole-exome and chr12 capture SNPs in each of 3 categories: concordant genotypes, NRD contributing, and NRS contributing for chr12 and whole-exome capture experiments when comparing WGA and genomic SNP call sets to Affymetrix SNP array genotypes.

3.3 Conclusions

The study described here provides an in-depth assessment of the suitability of WGA DNA for targeted resequencing and variant discovery using next generation sequencing. We evaluated whole exome as well as targeted genomic enrichment using Agilent SureSelect technology, and compared findings from WGA samples to results obtained with genomic DNA from the same individual, as well as validated a subset of SNP variant calls with Affymetrix SNP array genotypes. Overall, Venn analysis showed that the numbers of SNPs and indels called in the whole exome and chr12 capture callsets using WGA or genomic DNA is very similar, with the vast majority of variant sites shared between datasets. The concordance metric NRS demonstrates that using WGA DNA has high sensitivity for SNP sites with values of 98.28% and 99.63% for the whole exome and chr12 sequence capture callsets, respectively. The NRS for INDELs is lower at 91.17% and 94.07%. SNP NRD values for the whole-exome and chr12 callset were both less than 1%, but were an order of magnitude higher for INDEL calls. The lower values of these metrics may be due to slight differences in alignment of reads between genomic and WGA DNA in regions that contain INDEL variants. The majority of discrepant genotypes between WGA and genomic DNA involve heterozygous genotypes and statistical analysis suggests that these are enriched for GC alleles, at least in the whole-exome capture data. Validating a subset of the SNP made with genomic and WGA DNA that overlap sites on the Affymetrix 6.0 SNP array showed high sensitivity and high genotype accuracy for the whole exome capture callset. The sensitivity and genotype concordance numbers for the chr12 capture array were not as high, but the loss of sensitivity can be explained by lack of evidence of the alternate allele in the read pileup or poor zero mapping quality values spanning the SNP position. Downsampling and subsetting of reads to achieve lower coverage in WGA callsets (or match the starting number reads in the genomic sequencing experiment) consistently resulted in lower genotype concordance

and sensitivity metrics for the whole exome capture experiment, in contrast to the chr12 capture experiment. This difference may be due to statistical fluctuations of read sampling in the downsampling process, combined with the much smaller size of the chr12 capture region. Coverage and concordance metrics correlated with GC% of target intervals, with target intervals above the 3rd quartile of each respective capture array having less coverage and poorer concordance metrics. Our work complements the study of ElSharawy [34] who used a greater number of matched genomic / WGA samples in showing both genomic and WGA samples had high concordance and sensitivity metrics to HapMap III sites, but whose study examined only 384 exons. A limitation of our study is that we only have 1 genomic/WGA sample pair for each of the capture experiments, and the chr12 experiment captured a much smaller region of genomic DNA. Since the genomic sequencing was not repeated, we cannot know the expected discrepancy for a technical replicate, but we were able to validate a subset of our SNP calls that overlapped sites on the Affymetrix SNP array. Thus, our conclusions about allele bias, and the relationship between GC% content and genotype concordance must be taken with caution, but overall suggest that WGA samples can be used effectively in re-sequencing studies and thus offer a promising alternative for variant discovery studies using archived DNA.

3.4 Methods

WGA and genomic DNA sample preparation

Two sample sets were analyzed in this study. One sample was from a family cohort [73] that was sequenced for a 3.87 Mbp region on chr12 using a custom designed SureSelect capture array from Agilent. The second sample was from a single family that was whole exome sequenced using the Agilent SureSelect All Exon kit. In both

cases, the genomic DNA was originally isolated from blood samples. A REPLI-g Mini Kit (Qiagen) was used to prepare WGA DNA from 15 ng of starting genomic DNA.

Sequence Capture

We used two different Agilent SureSelect kits to perform sequence capture on the samples used in this study. The first was a custom array designed to capture a 3.87 Mbp region on chromosome 12 . The second was an Agilent SureSelect All Exon kit designed to capture at total of 49.4 Mbp of exonic sequence spanning the whole genome. The standard Agilent SureSelect protocol for Illumina paired-end sequencing was used which requires 3 of micrograms of starting genomic DNA.

DNA sequencing

Samples were paired-end sequenced on an Illumina GAII machine with read lengths of 101 bp. with insert size for the genomic and WGA whole exome capture samples being each 370 bp, respectively. Insert sizes for the genomic and WGA chr12 capture samples were both 320 bp, respectively. Both sets of genomic DNA samples were multiplexed with other samples not part of this study, while each of the corresponding WGA DNA samples were sequenced in an individual flow cell lane. Fastq files were generated via the Illumina CASAVA pipeline v1.8. The starting number of passed filter reads is shown in additional Tables 3.2 and 3.3, as well as additional metrics of capture experiments.

Bioinformatics Pipeline

We applied the same bioinformatics pipeline to WGA and genomic DNA samples as shown in figure S4. All programs from the Genome Analysis Toolkit (GATK) were from version v1.6-5-g557da77 [30]. All programs from Picard were from v1.50 [132].

Fastq files were aligned to the human reference sequence GRCh37 with the program MOSAIK v2.0.113q [85]. Parameter values to MosaikAligner were as follows: -act 35, -bw 37, -mhp 200 -mm 14. Capture metrics for the whole exome and chr12 capture experiments were calculated using the program CalculateHsMetrics in Picard. Base quality scores were recalibrated with the GATK programs CountCovariates and TableRecalibration. PCR duplicates were marked using the program MarkDuplicates, which is part of Picard. SNP and INDEL variants were discovered using the GATK program UnifiedGenotyper. Parameters used for running UnifiedGenotyper were as follows: -stand_call_conf: 10, -stand_emit_conf: 30, -glm: BOTH, -out_mode: BOTH, -hets: .001. Each member of the WGA/genomic sample pair was called independently as a single sample. SNP variant calls were filtered using the GATK program VariantFiltration with the following filtering parameters:

((MQ0 / (1.0 * DP)) $>$ 0.05) || DP $<$ 5 || QUAL $<$ 30.0 || QD $<$ 5.0 || HRun $>$ 5.0 || SB $\geq$ -0.10

INDEL variant calls were filtered with the following:

((MQ0 / (1.0 * DP)) $>$ 0.05) || SB $\geq$ -1.0 || QUAL $<$10

Where MQ0 = Number of reads with mapping quality zero, DP = depth of coverage, QUAL= Phred scaled quality score, HRun = Largest contiguous homopolymer run of variant allele in either direction, QD = Variant Confidence/Quality by Depth, and SB = Strand Bias.

Downsampling and Subsetting of Reads in WGA and Genomic BAM files

To investigate the relationship between sequence coverage and number of variants discovered, aligned reads from both WGA BAM files were downsampled to different levels average target coverage using the Picard v1.50 program DownsampleSam. Since UnifedGenotyper restricted its variant calling to target capture interval regions, only aligned reads that had a minimum 1-bp overlap with a target interval were considered in the downsampling process by removing off target alignments by using the pairToBed program in BEDTools package [119] For the chr12 WGA BAM, 100 downsampled BAM files were generated with average target coverages of 100x, 200x and 300x, respectively. For the whole-exome WGA BAM, 100 downsampled BAM files were generated at coverage levels of 20x, 50x, and 80x. Since the WGA prepared samples had higher sequence coverages, the coverage range of the downsampled BAMs were chosen so they would closely overlap the coverage of the original genomic DNA sample. Due to the stochastic nature of the downsampling process, as well as variation in capture efficiency between targets, it was difficult to get exact match in the number of reads between WGA and genomic BAMs. The number of reads needed to achieve a desired coverage was determined by solving this equation: C=(N $\times$ L)/G, where C is the coverage, N is the number of reads, G is the size of the genome (in this case the total length in base pairs of capture array targets), and L is the read length value (101 bp).

In addition to downsampling the reads from the WGA BAMs for both capture experiments, an exact number of read pairs were randomly sampled from the initial WGA fastq files to match the starting number of genomic DNA fastq read pairs. This was accomplished by writing a Python script that randomly selects a specified number of

read pairs from a fastq file. Once the subset of fastq read pairs were selected they were put through the same bioinformatics pipeline applied to the original data.

Callset comparison metrics

We compared the variant calls from genomic and WGA using three types of metrics. The first was site level intersection to see if the same genomic position was called variant in both callsets. The other two types of metrics were non-reference sensitivity (NRS), and non-reference discrepancy (NRD), shown in Figure 3.5. NRS measures the fraction of sites called variant in the comparison callset that are also called variant in the evaluation callset. For this study the evaluation callset are the WGA variant calls and the comparison callset are the genomic variant calls. Sites called homozygous reference or no-call in the evaluation calls, but were variant in the comparison callset reduce NRS. NRD measures the accuracy assigned genotypes called by both datasets. It excludes concordant homozygous reference calls. To calculate these values, the VCF files of the WGA and genomic callsets were merged using the GATK program CombineVariants and then calculated in Python.

SNP validation with Affymetrix 6.0 Human SNP array

The SNP variant calls for WGA and genomic DNA for both capture sequencing experiments were compared to Affymetrix 6.0 Human SNP array derived genotypes for the same samples. SNP array genotypes were called with Birdseed v2. The 6.0 Human SNP array contains a genomewide collection of more than 900,000 sites. For a SNP array variant to be included in the validation analysis it must overlap a target region on the capture array and have a confidence score of at least 0.05. Only those variants that met these two conditions were considered. Based on these criteria there were a total of 11831 SNPs on the 6.0 array that overlapped the whole exome capture targets and 1435 SNPs that overlapped the custom chr12 capture targets. Similar

to the comparison of WGA calls to genomic DNA calls, the VCFs of sequencing and array derived genotypes were merged using the GATK program CombineVariants. The sequencing derived genotypes were evaluated by comparing them to the array based genotypes and the NRS and NRD concordance metrics were calculated. Only sites that have PASS in the filter column of the individual VCFs were included when calculating NRS and NRD from the CombineVariants derived VCF.

Statistical analysis of allele bias in SNP calls

For both the whole-exome and chr12 capture experiments, genomic and WGA SNP call sets were merged, and then placed into 4 categories: concordant, uniquely called genomic, differing genotypes (NRD contributing), and WGA uniquely called SNPs. The counts of each of the 6 possible allele combinations in each category were tallied. To test the null hypothesis that the proportion of SNPs are equal across all 4 categories, the pairwise.fisher.test using the Bonferroni correction method was applied in succession to each of the 6 possible allele combinations in R [120]. The pairwise.fisher.test is part of the CRAN R package fmsb [104] . The significance level $\alpha = .05$ was chosen. The appendix contains of p-values for the whole-exome and chr12 capture experiments.

A similar analysis was performed when comparing the sequencing derived SNP calls to Affymetrix array derived genotypes for genomic and WGA capture experiments. The sequencing and Affy callsets were merged (only SNPs on the Affymetrix array that overlapped a target capture region were included) and placed into concordant, NRS, or NRD contributing categories. The appendix contains p-values for the whole exome and chr12 comparisons to the array based genotypes.

Sample Ascertainment

All samples and protocols for this study have been reviewed and approved by the IRB of the Medical College of Wisconsin. In accordance with the approved protocols, all participants provided written informed consent to participate in the study. Only adult individuals were included in the study.

Chapter 4

Discrete filtering approach to prioritize variants in a Mendelian exome study of non-sensorineural hearing loss

4.1 Background

Hearing is an important biological function and species capable of sensitive sound detection have a potential selective advantage. Sound transduction is an intricate process, and unsurprisingly up to 1 percent of the approximately 20,000 human genes in the human genome are involved in hearing [37]. The mammalian auditory system is comprised of the external, middle, and inner ear (cochlea). Components of the external ear include the pinna, which is the part of the ear that lies external on the head, and ear canal. Sound waves travel through this canal and vibrate the ear drum. This movement is transferred to the middle ear by movement of three small bones, the malleus, incus, and stapes. The movement of the stapes transmits sound

waves to the fluid filled inner ear. The cochlea is where sound waves are converted to electrochemical signals which is transmitted by the auditory nerve to higher levels of the auditory system [37].

A recent World Health Organization report indicated that over 360 million people worldwide suffer from disabling hearing loss (HL) [61] HL is fairly common in human populations with congenital deafness occurring 1 in every 1000 births [146]. Late onset, progressive instances of HL are genetic in origin, with genes playing a critical role with aging associated HL [146]. Non-syndromic hearing loss is not associated with any other clinical symptoms, while syndromic hearing loss is associated with other abnormalities in the body [130]. Non-syndromic sensorineural hearing loss (NSHL) is deafness associated with alterations to the structures of the inner ear [146]. Over 1000 mutations in 60 genes have been cataloged [108]. There is considerable genetic heterogeneity with NSHL, making the search for the genetic basis of deafness a challenging task [146]. The first gene identified to cause HL was the X-linked *POU3F4*, identified by linkage mapping in 1995 [29]. Genomic enrichment and next generation technology have vastly accelerated the discovery of new causal loci and have identified a dozen new loci [146]. Previous studies have applied a two step approach of first performing linkage analysis and then following up with exome capture sequencing [129]. This is more efficient than Sanger sequencing of candidate loci. Recent examples of using this approach include the study by Walsh et. al. [140] who combined homozygosity mapping with exome sequencing to identify a novel, non-synonymous variant in *GPSM2*, a G-protein signaling modulator essential for cell polarity. Another study by Yariz and colleagues [148] identified a frameshift deletion and a compound heterozygote in *OTOGL*, which is a protein associated with the cellular membrane of the inner ear.

In this study we describe a whole exome capture experiment where six members from a larger family pedigree of individuals of European descent living in Northern Wisconsin were diagnosed with NSHL. Previous linkage analysis identified a 18 Mbp size linkage interval with a LOD score of 3 on chromosome 12. No mutations were found in previously implicated candidate genes involved in HL, suggesting a novel, causal variant segregates in the family. Using a discrete filtering approach with a minor allele frequency cutoff identified a non-synonymous mutation in *TMTC2*, a transmembrane protein, that segregates perfectly with the phenotype in the family and is enriched in a set of 200 unrelated individuals with the same form of hearing loss. Functional studies suggest the electrophysiology is altered in cells lines that contain the variant. The work here also suggests that hard filters against variant catalogs to narrow down the list of candidate variants may need to be adjusted to use a minimum minor allele frequency cutoff when studying genetic heterogeneous traits with potentially incompletely penetrant alleles.

4.2 Results and Discussion

.

Capture sequencing metrics

The pedigree of the family that participated in this study is shown in Figure 4.1. A subset of six individuals from the pedigree were whole-exome sequenced using the Agilent SureSelect All Exon Kit (see section 5.4). Table 4.1, shows the capture sequencing metrics of the samples (see section 5.4 for more details).

Figure 4.3 shows variability in coverage for the samples that underwent whole-exome sequencing, as summarized by box plots for each sample. The median values range

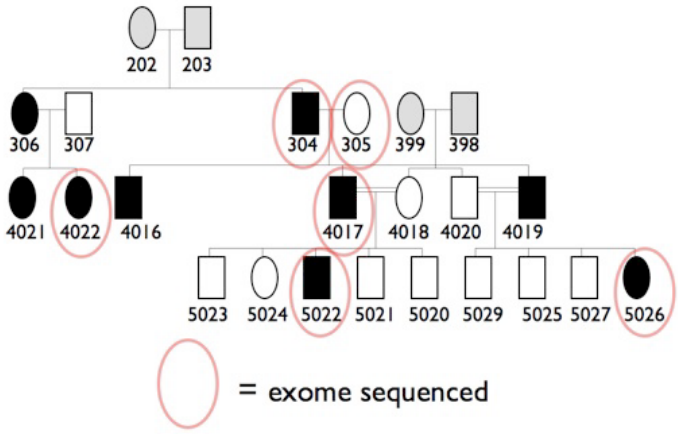


Figure 4.1: **Pedigree of hearing loss family** - Individuals of pedigree selected for exome analysis. Individuals with filled shapes are affected. Individuals in gray are deceased.

from 50x to 60x for each sample. Figure 4.4 shows the per-base cumulative coverage for each sample. Despite the variability in per target coverage, for each sample studied, at least 80 percent of targeted bases had at least 20x coverage. Focusing in targeted bases that lie in the chr12 linkage region, figure 4.5 shows the box plots of median target coverage for each sample. Median values range between 65x-75x. Figure 4.6 shows the per-base cumulative coverage in the chr12 linkage region. All samples have at least 85-90% of linkage region bases with at least 30x coverage. Table 4.2 shows the numbers of genes, exonic sequence, and Aglilent capture targets in the chr12 linkage region. Of the 120 kbp of exonic sequence in the linkage region, nearly 92% of them are covered by an Agilent target.

dataset	sample_3_4	sample_3_5	sample_4_16	sample_4_22	sample_5_22	sample_5_26
Read Length	101	101	101	101	101	101
Target Territory	49649722	49649722	49649722	49649722	49649722	49649722
total reads	105316652	110371248	86940416	104367328	95981328	114601152
total unique reads	79549416	77491632	73449923	85295071	81154142	89518704
total unique reads aligned	62036624	60005207	57696024	66775648	63191839	69840387
% Usable bases on target	35	33	38	38	31	37
% Selected bases	83	79	81	81	62	82

Table 4.1: Sequencing capture metrics

Linkage region	chr12:78475869-96475869
Number of genes	104
Number of transcripts	214
Number of exons	779
Exonic sequence (bp)	120584
Number of Agilent targets	723
Agilent targets (bp)	160376
Exonic sequence covered by Agilent (bp)	110741
Exonic sequence not covered by Agilent (bp)	9843

Table 4.2: Summary of genes, targets, and exonic sequence in linkage region

Discrete Filtering

Discovering the causal variant from the background of non-pathogenic polymorphisms is a key challenge in analyzing exome sequencing data for Mendelian traits. Previous studies analyzing exome sequence data to identify causative alleles of Mendelian disease have utilized a discrete filtering approach [7, 106, 105]. This approach searches for variants shared by all affected individuals sequenced. Next, assuming the causative mutation is novel, candidate variants are filtered against known catalogs of genetic variants, such as dbSNP or the 1000 Genomes Project. Next, variants can be stratified by their function (i.e. synonymous, non-synonymous, loss of function) and functional impact (benign, damaging, conserved). Methods like SIFT [79], and PolyPhen [3] use multiple alignments from related sequences and/or physiochemical properties of mutations to predict any potential deleterious affect. Methods like phyloP [112] measure evolutionary conservation by measuring rates of mammalian evolution at an individual nucleotide level inferred from whole genome alignments of

multiple species.

We applied a discrete filtering strategy shown in Figure 4.8 and described in detail in Section 4.4. The first step of filtering out variants not conforming to a dominance inheritance pattern was applied exome wide to variants found in all target regions. Subsequent steps were focused on the linkage interval on chromosome 12. Table 4.3 shows the numbers of variants found after applying each filter step. All of the 24 SNPs and 2 INDELs that remained were present in the 1000 Genomes PhaseI European callset from February 2012. Annotating these remaining variants showed that 5 were non-synonymous SNPs. Table 4.4 shows the minor allele frequency of the 5 non-synonymous and 1 UTR SNPs in the European Phase1 1000 Genomes callset and the European NHBLI Exome Sequencing Project (ESP) [134]. All but one variant are segregating at high frequency, while the single rare variant in the gene *TMTC2* is segregating at 1 percent in 1000 Genomes and .76% in ESP. Next, PolyPhen2 [3] and SIFT [79] classifications and phyloP [112] score was obtained for the TMTC2 variant. Table 4.5 shows the results of these tools. Both PolyPhen2 and SIFT classify the TMTC2 variant as tolerated and benign. The phyloP score is 1.40. phyloP measures evolutionary conservation at the individual nucleotide level, and positive scores suggest evolutionary conservation. Looking at the descriptive statistics of phyloP scores of all exonic nucleotides in the linkage region, shown in Table 4.6, reveals that a score of 1.40 near the 50th percentile of all phyloP scores. While the phyloP score suggests evolutionary conservation, half the other exonic sites in the linkage region have larger phyloP scores, making the interpretation of its score unclear.

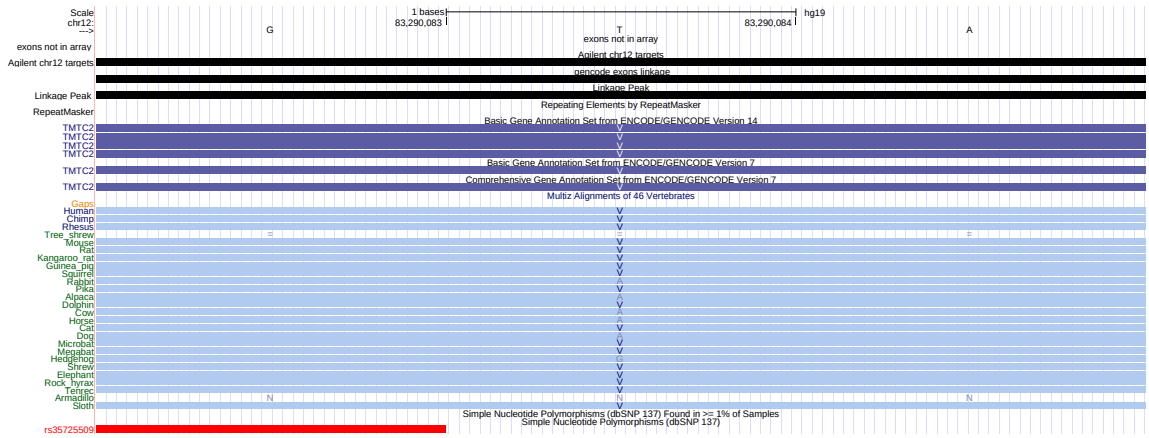


Figure 4.2: **Putative causal variant in *TMTC2* gene** - A non-synonymous *G*TA to *A*TA mutation in TMTC2 resulting in a Valine to Isoleucine replacement.

	Freebayes callset	Dominant inheritance	reside in linkage peak	1000G membership	variant annotation
SNPs	84357	1574	24	24	5 nsyn, 4 syn, 14 intronic, 1 UTR
INDELs	12170	138	2	2	1 fs, 1 intronic

Table 4.3: Numbers of variants after each discrete filter step

TMTC2 annotation

Based on the the evidence showing that the *TMTC2* variant is segregating at low frequencies in the 1000 Genomes and ESP datasets, it was flagged as a possible causal mutation. Figure 4.2 shows the position of the mutation resulting in a Valine to Isoleucine replacement. The specific function of TMTC2 is unknown but it is a transmembrane protein containing a tetratricopeptide repeat motif [65]. The tetratricopeptide repeat (TPR) is a structural motif consisting of 34 degenerate amino acids. It is found in a number of proteins that mediate protein-protein interaction. [11]. Performing a PFAM search [117] with the amino acid sequence of TMTC2 shows that it contains three TPR domains, but the amino acid residue (381) which the non-synonymous mutation changes, does not seem to be in a TPR domain. Utilizing the web server TMHMM [138], which predicts transmembrane spanning regions in protein sequences, indicates that mutated amino acid residue is cytoplasmic.

Gene	rsid	Annotation	PhaseI 1KG MAF %	ESP MAF %
TMTC2	rs35725509	nsyn	1	.76
LRRIQ1	rs3765044	nsyn	26	27
LLRIQ1	rs17012533	nysn	26	27
POC1	rs2230283	nsyn	36	34
USP44	rs3812813	nsyn	55	45
TSPAN19	rs7962577	5-UTR	50	46

Table 4.4: Minor allele frequencies of non-synonymous candidate mutations

SIFT	PolyPhen2	phyloP
Tolerated	Benign	1.40

Table 4.5: Results from functional impact methods for *TMTC2* mutation

Exons missed by capture

While Figures 4.5 and 4.6 show very good coverage metrics for Agilent targets in the linkage region, another important point to address are the genes in the linkage region whose exons were not covered by a target in the capture array. Table 4.2 shows that 9.8kb of exonic base pairs were not covered by an Agilent target. Table 4.7 lists the genes and exonic base pairs missed. Note, the numbers in the second column of the table do not add up to that in table 4.2 due to the fact that some exonic intervals may have been counted more than once due to alternative transcripts of the same gene. Two genes on the list, *PTPRQ* and *OTOGL* have been previously associated with hearing loss [148, 127]. The three genes *PTPRQ, OTOGL, TMTC2* all span a 2.4 Mbp region on chr12. While there is certainly a distinct possibility there could be other pathogenic mutations segregating in these genes, additional genotyping of the *TMTC2* mutation and functional experiments suggest it has an affect on phenotype.

4.3 Conclusions

Here I described a discrete filtering approach to identify a putative causative mutation for non-syndromic sensorineural hearing loss. The results show the utility of using the discrete filtering approach to narrowing down a list of candidate variants, as well

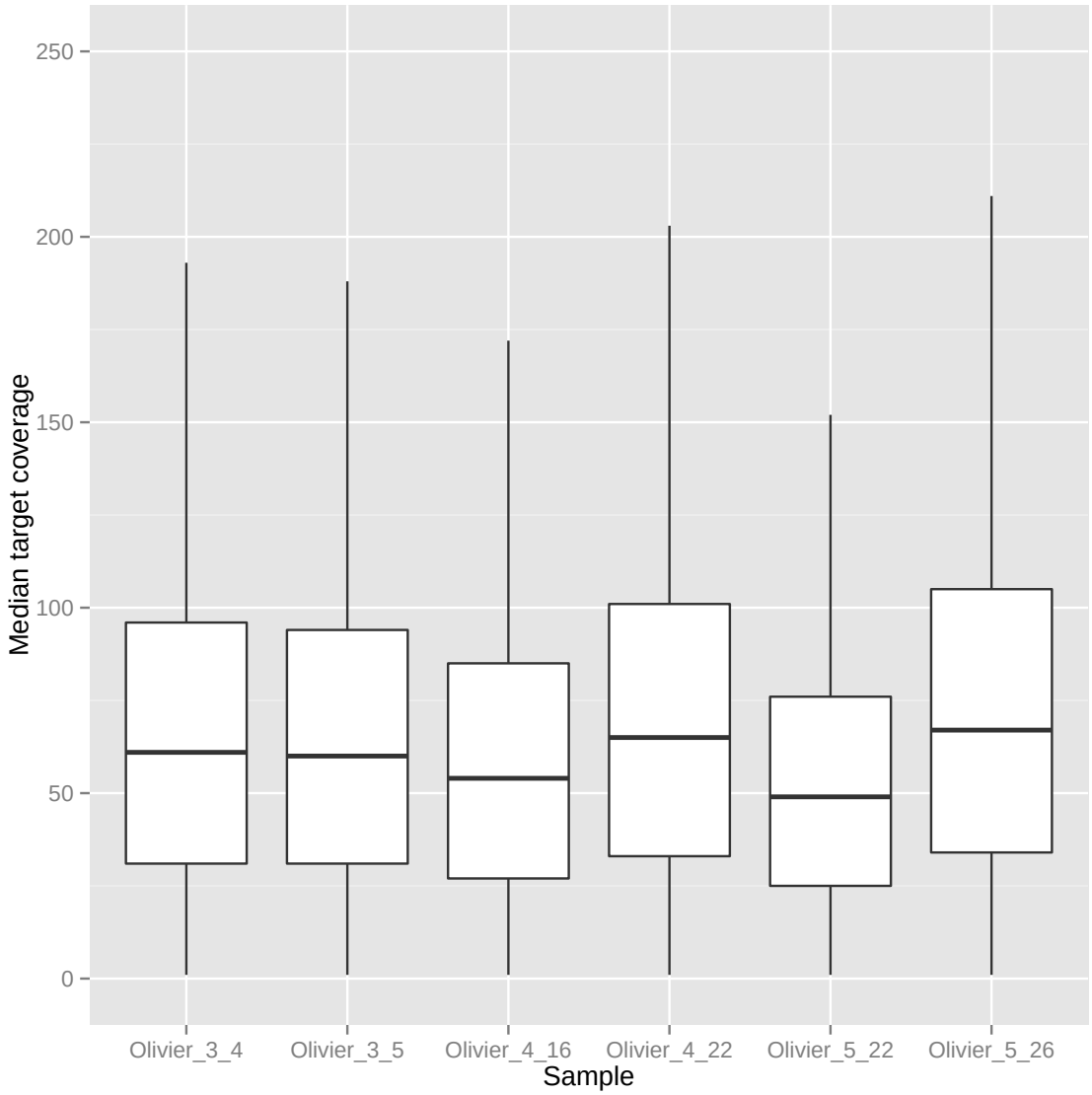


Figure 4.3: **Median target coverage boxplot** - Whole-exome median target box plot of the six samples investigated in this study

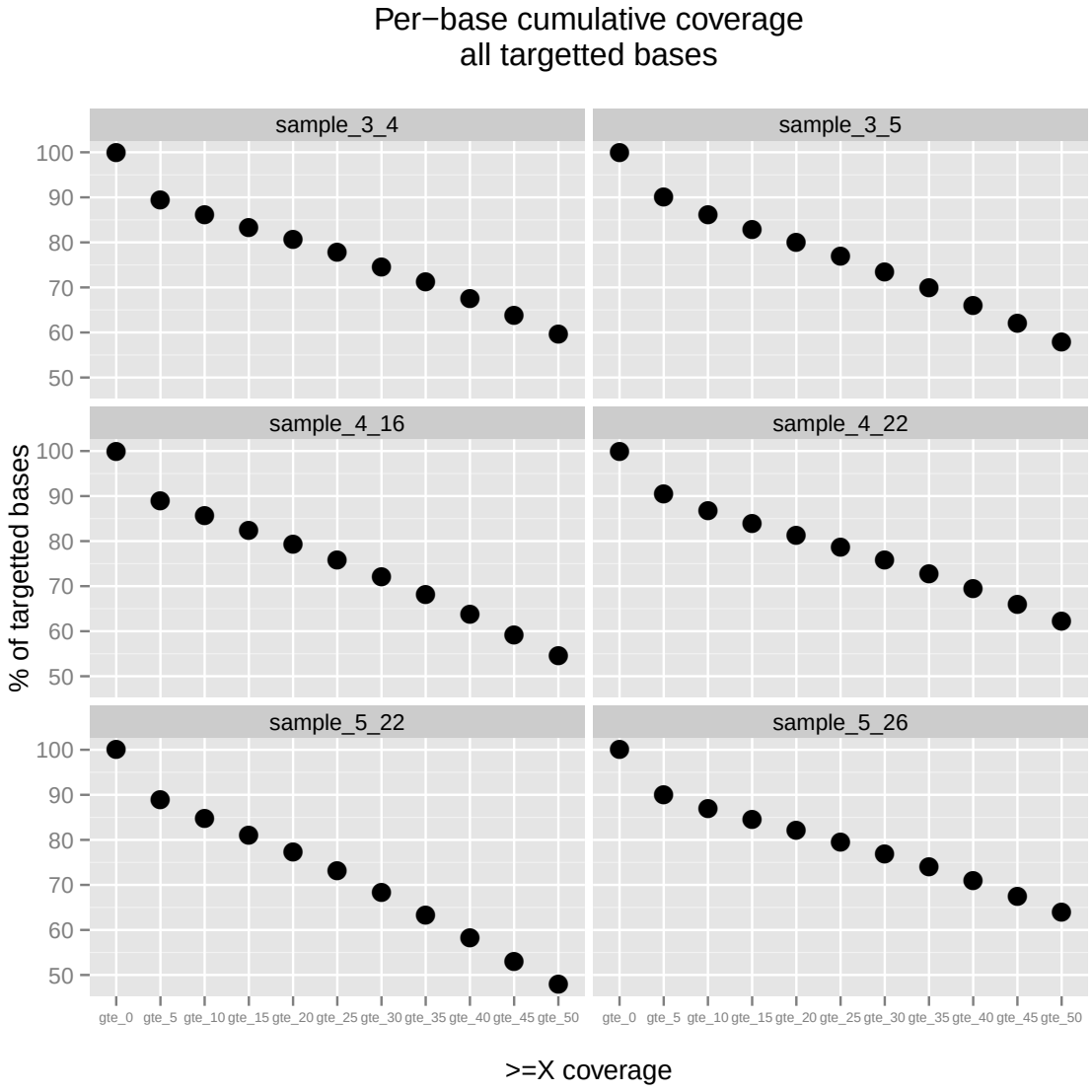


Figure 4.4: **Cumulative per-base coverage** - Whole-exome per-base cumulative coverage for the six samples investigated in this study

Min	1st Quartile	Median	Mean	3rd Quartile	Max
-9.16	0.34	1.48	1.30	2.26	2.94

Table 4.6: Summary of placental mammal phyloP scores of exonic basepairs in chr12 linkage region

Gene	exonic sequence missed (bp)
AC024909.1	2254
ALX1	129
ATP2B1	1779
BTG1	510
C12orf12	2961
C12orf37	1082
CCDC41	99
CEP290	330
CLLU1	1976
CLLU1OS	180
CRADD	68
DUSP6	1760
EEA1	60
LIN7A	219
LRRIQ1	6929
LTA4H	232
METAP2	132
MGAT4C	3467
MRPL42	20
NAV3	185
OTOGL	805
PAWR	663
PLXNC1	1452
POC1B	773
PPFIA2	2336
PPP1R12A	2261
PTPRQ	13150
TSPAN19	242
VEZT	6735

Table 4.7: Genes missed by Agilient capture

as some of its challenges. The assumption that pathogenic, causal mutations should not be present in variant catalogs did not hold in this study. Of the non-synonymous SNPs that remained, after filtering against the 1000 Genomes and Exon Sequencing Project variant catalogs, only TMTC2 was segregating at low frequency. Despite functional impact and evolutionary conservation of the Valine to Isoleucine substitution shown to be benign and difficult to interpret, it was still selected for follow up

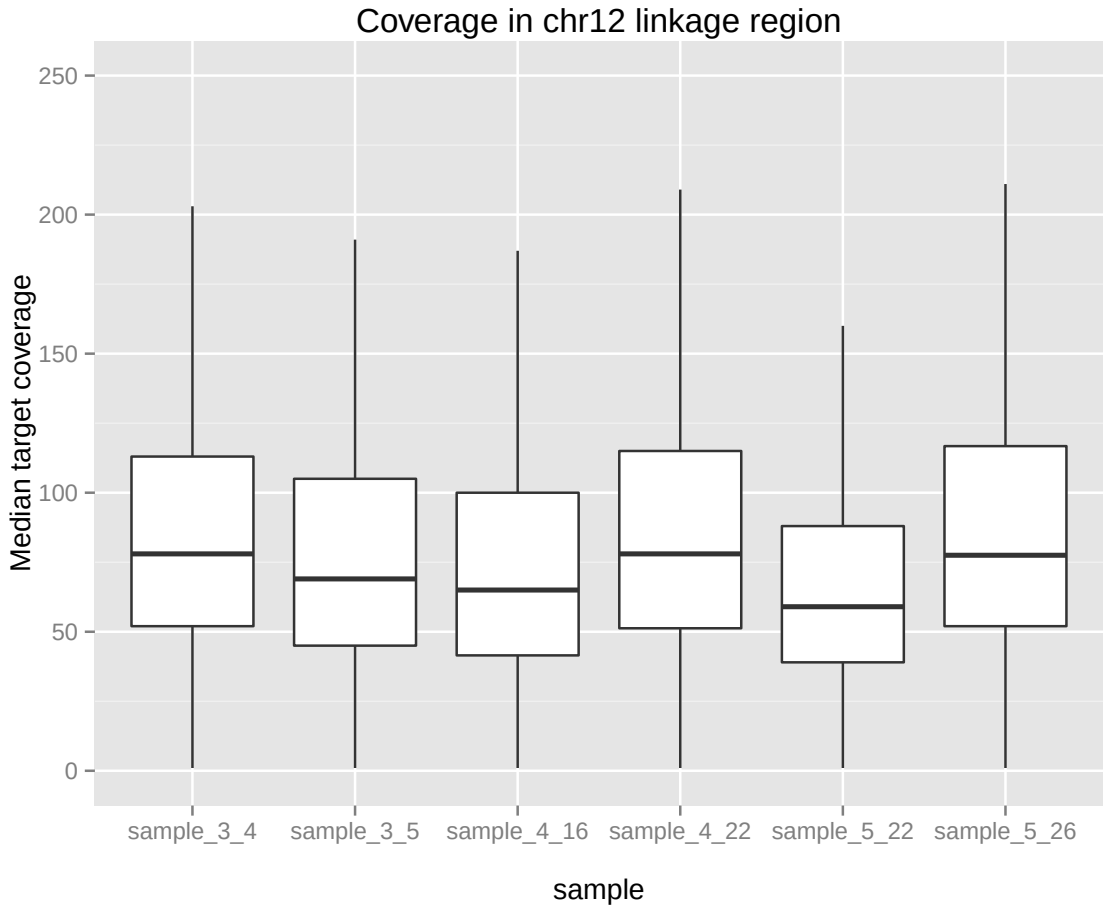


Figure 4.5: **Median target coverage box chr12 linkage region** - Median target box plot in chr12 linkage region for the six sample investigated in this study

genotyping and functional analysis due to its rare frequency. The protein product of *TMTC2* is a transmembrane protein with a tetratricopeptide repeat (TPR) motif. Other proteins with this motif have been shown to been involved in mediating protein-protein interactions. Follow up genotyping in a cohort of 200 unrelated individuals suffering from NSHL showed that the TMTC2 variant was segregating at 3 percent, nearly 4 times as high as in the European ESP population. The mutation was genotyped and present in every affected member of the proband studied that was not selected for exome sequencing (M. Olivier, personal communication). The technical shortcomings of exons in the linkage region not being covered by the capture array in genes previously implicated in hearing loss does not diminish our findings.

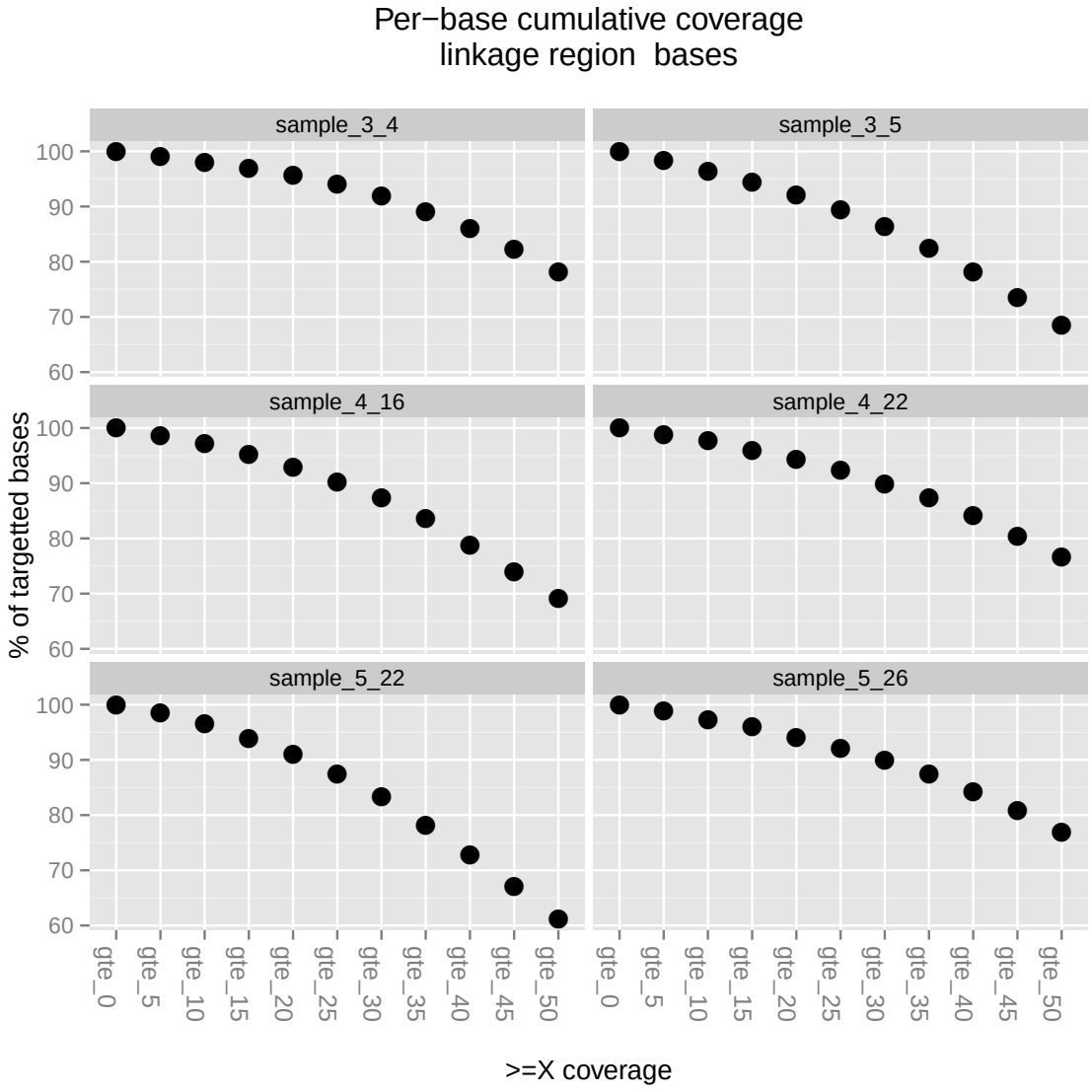


Figure 4.6: **Cumulative per-base coverage chr12 linkage region** - chr12 linkage region per-base cumulative coverage for the six samples investigated in this study

If there were other pathogenic mutations in the genes *PTPRQ* and *OTOGL*, they likely are on the same haplotype, and it would be difficult to disentangle the affects of any other mutations. It is not necessarily clear if the TMTC2 variant has complete penetrance, and if so, rather than requiring complete absence of candidate mutations from variant catalogs, it might be necessary to employ a minor allele frequency cutoff when filtering Mendelian exome callsets [131]. Other limitations of exome sequencing that could effect our conclusions are the existence of potential functional variants in non-coding regions not covered by the capture array. Variants in enhancer or silencer elements could modulate the amount of wild-type or mutant transcript, leading to phenotypic variance [17].

4.4 Methods

DNA sequencing

All samples were pair-end sequenced on an Illumina GAII machine with read lengths of 101 base pairs with insert sizes ranging from 350 to 370 base pairs, respectively. Fastq files were generated from Illumina's CASAVA v1.8 pipeline.

Sequence Capture

All samples underwent genomic enrichment using the Agilent SureSelect All Exon kit, which is designed to capture a total of 49.4 Mbp of exonic sequence spanning the whole genome. The standard Agilent SureSelect protocol for Illumina paired-end sequencing was used, requiring 3 micrograms of starting genomic DNA.

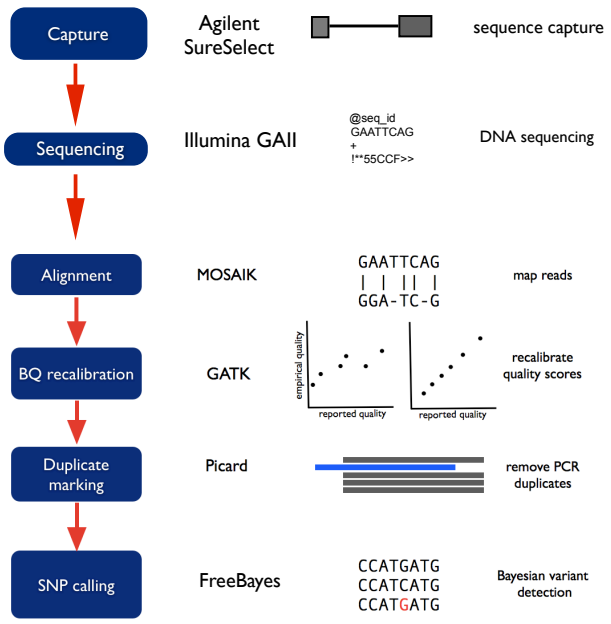


Figure 4.7: **Bioinformatics Pipeline** - Bioinformatics pipeline for hearing loss exome study

Bioinformatics Pipeline

The bioinformatics pipeline depicted in Figure 4.7 was applied to all six samples in the study. Fastq files were aligned the human reference sequence GRCh37 with then program MOSAIK v2.0.113 [85]. Parameters given to MosaikAligner were: -act 35, -bw 37, -mhp 200, and -mm 14. All programs from the Genome Analysis Toolkit (GATK) [30] were from version GenomeAnalysisTK-1.0.5974. Base quality scores were re-calibrated with GATK programs CountCovariates and TableRecalibration. Duplicate marking was performed with Picard v1.45 program MarkDuplicates [132]. SNP and INDEL variants were called with FreeBayes v0.8.9 [40] with all samples called jointly using the parameters –min-alternate-count 5, –min-alternate-qsum 40, –binomial-obs-priors, –allele-balance-priors. Variants were called in slightly modified

target capture intervals for the Agilent All Exon kit that included plus 50 bp upstream and downstream of original starting and ending coordinates. Post-filtering of the FreeBayes calls required a minimal QUAL value of .5 and a maximal read depth of 1000.

Discrete filtering of variants

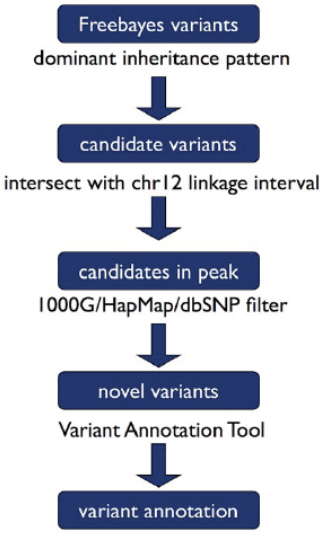


Figure 4.8: **Exome filtering steps** - A series of discrete filtering steps was applied to narrow the list of candidate mutations.

Figure 4.8 shows the discrete filtering steps to narrow down the list of candidate causative variants. The list of SNPs and INDELs called by FreeBayes were filtered by removing any variant site that did not conform to a dominant inheritance pattern. This means that the single unaffected individual was required to be homozygous reference, while the remaining affected individuals were required to be heterozygote

or homozygote non-reference. Since there was a strong linkage signal on chr12 based on linkage analysis from previous genotyping, we further restricted variants to those found only under the linkage interval. Finally, assuming that only non-synonymous mutations are functional, variants were annotated as synonymous or non-synonymous using the Variant Annotation Tool (VAT) [80].

Functional impact using PolyPhen2 and phyloP

Functional impact predictions from PolyPhen2 were obtained from the Polyphen-2 and SIFT webservers `http://genetics.bwh.harvard.edu/pph2/` and `http://sift.jcvi.org`. phyloP scores for exonic regions in the linkage interval were retrieved using the UCSC Table Browser [69].

Using kinship coefficient to select individuals to sequence

Table 4.8 shows the condensed identity coefficients for each possible pair of individuals that were sequenced from the pedigree shown in Figure 4.1. They were calculated by the program idcoefs [2]. There are 9 possible condensed identity coefficients and they give a complete probability distribution for identity by descent (IBD) between single loci of two individuals [94]. The first 6 coefficients quantify the probability of being inbred and since there is no consanguinity in the family studied, these probabilities are 0. Coefficients 7-9, denoted in Table 4.8 as Δ_* give the probability of sharing 2, 1, or 0 genes IBD between a pair of individuals. When deciding which individuals to exome sequence in a family it is most advantageous to select individuals that are the most distantly related. These relationships are quantified by the identity coefficients in Table 4.8.

Indvidual 1	Individual 2	Δ_7	Δ_8	Δ_9
sample_3_4	sample_3_5	0	0	1
sample_3_4	sample_4_22	0	0.5	0.5
sample_3_4	sample_4_16	0	1	0
sample_3_4	sample_5_22	0	0.5	0.5
sample_3_4	sample_5_26	0	0.5	0.5
sample_3_5	sample_4_22	0	0	1
sample_3_5	sample_4_16	0	1	0
sample_3_5	sample_5_22	0	0.5	0.5
sample_3_5	sample_5_26	0	0.5	0.5
sample_4_22	sample_4_16	0	0.25	0.75
sample_4_22	sample_5_22	0	0.125	0.875
sample_4_22	sample_5_26	0	0.125	0.875
sample_4_16	sample_5_22	0	0.5	0.5
sample_4_16	sample_5_26	0	0.5	0.5
sample_5_22	sample_5_26	0.0625	0.375	0.5625

Table 4.8: Pairwise kinship coefficients for individuals exome sequenced

Chapter 5

SNP Variant discovery in pedigrees using Bayesian networks

5.1 Background

Next generation sequencing technologies have reduced the cost and increased the throughput of DNA sequencing experiments by sequencing DNA molecules in a massively parallel fashion [102]. This has enabled geneticists to sequence large numbers of individuals to properly characterize the numbers of rare variants segregating in the human population. Projects like the 1000 Genomes have provided the genetics community with a comprehensive catalog of genetic variants that include rare and low frequency loci [33]. There has been increased attention to the role that rare variants might play in explaining the missing heritability in genome wide association studies that previously SNP genotyped only common variants.

While association studies using unrelated individuals have had success [56], family sequencing studies offer a different avenue to uncovering new associations. While rare variants segregate at low frequency in the population, sequencing multiple af-

fected individuals in the same family can be potentially enriched for causal mutations [20] and can increase the statistical power of rare variant analyses [88, 89]. There have been several methods dedicated to variant discovery from next generation sequencing datasets, and the majority of these assume that the samples are unrelated [40, 30]. Modeling Mendelian inheritance when analyzing such datasets can potentially improve the sensitivity and accuracy of results, in particular of non-founder individuals. This is because by modeling the data as a Bayesian network, genotype inference for non-founder individuals is leveraging information from parental samples. Here I present a method called Pgmsnp that incorporates pedigree relationships when assigning SNP genotypes to each member from a family sequencing dataset. The method models the pedigree as a Bayesian network and uses a belief propagation algorithm to compute posterior genotype probabilities of family members. First I describe the basics of Bayesian networks and the belief propagation algorithm used. Next, I present simulation results on a variety of pedigree structures using Pgmsnp and three other SNP calling methods. Finally Pgmsnp results, as well as competing methods, are presented on an empirical sequencing dataset from the Ilumina Platinum genomes collection on a subset of a 17 member pedigree. Pgmsnp genotyping results perform better than using the standard approach of assuming all samples are un-related at lower sequence coverage. Compared to other pedigree aware methods tested in this study, Pgmsnp has comparable sensitivity of detection, but has slightly less genotyping accuracy. Specifically, for non-founder individuals in the Illumina Platinum pedigree, Pgmsnp has a higher sensitivity and better genotyping accuracy than the method GATK, which doesn't incorporate Mendelian relationships. Overall, results suggest that incorporating Mendelian relationships of samples as a Bayesian network improves the sensitivity of SNP detection of non-founder members.

Probabilistic Graphical Models

Bayesian networks are a type of probabilistic graphical model (PGM). Probabilistic graphical models compactly represent a complex distribution using a graph based representation. Random variables are represented as nodes and edges represent probabilistic relationships between random variables [76]. Probabilistic graphical models have the following useful properties: 1) Visualization of a probability model 2) Probablistic dependencies can be inspected from the graph. 3) Complex computations like joint, conditional, and marginal probabilities can be expressed in terms of graphical manipulations [10].

A complex probability distribution can be represented compactly in a graphical way, and using this representation inferences about certain variables can be computed using efficient algorithms. One example is computing posterior probabilities of some variables given observations or evidence about others [76]. These algorithms work directly on the graph structure rather than manipulating the joint distribution algebraically, which can become quite cumbersome and unintuitive to handle, especially if the numbers of variables in the distribution is large [113, 76].

Fundamental to the representation of Bayesian networks is the chain rule of probability:

$$Pr(X_1, \ldots X_k) = Pr(X_1)p(X_2|X_1) \cdots Pr(X_k|X_1, \ldots, X_{k-1}) \tag{5.1}$$

where the left side of the equation represents the joint distribution of a set of random variables $X_1 \ldots X_k$.

Graphical models can represent joint probabilities in a symbolically efficient way by defining *local* relationships amongst variables. Suppose each node has a set of parent nodes (which can be the empty set). Let π_i represent the set of indices of the parent node X_i, such that X_{π_i} refers to the parents of X_i. The parent-child relationships can used in making efficient representations of joint probability distributions:

$$Pr(x_1 \dots x_n) \equiv \prod_{i=1}^{n} Pr(x_i | x_{\pi_i}) \tag{5.2}$$

so the joint probability is a product of the local functions in the graph. Lauritzen and Sheehan [84] refer to this as a *Bayesian network* if the graph is a directed acyclic graph (DAG). Also, for any node, given the values of its parents, are conditionally independent of all nodes which are not descendants. This is the directed local Markov property. Using Equation 5.2, the joint distribution of a Bayesian network is described from the associated DAG and conditional probability distributions of each node, given its parents. The corollary of this is that pedigrees are DAGs and their joint distribution of genotypes can be specified using Equation 5.2.

Representation

Using directed graphs to analyze probability distributions has a long history in genetics, dating back to the work of Sewall Wright and his work on path analysis [143, 144]. Pedigree structure can be represented quite naturally as a Bayesian network. First we introduce the concept of a *factor*. Let $\mathbf{D}$ be a set of random variables. We define a factor ϕ to be a function from Val($\mathbf{D}$) to the set of real numbers. The set of variables in $\mathbf{D}$ is the scope of the factor and is denoted $Scope[\phi]$. Essentially, we can think of factors as (conditional) probability tables. Inference algorithms for Bayesian net-

works manipulate factors to compute entities of interest such as joint and marginal probabilities. Details of the structure and representation of the Bayesian network used in this study is presented in Section 5.4.

Inference

Inference in Bayesian networks involves computing the (posterior) values of some variables, given evidence about others [76]. Efficient exact inference algorithms are an essential feature of Bayesian networks that allow joint, conditional, and marginal probabilities to be computed. The following sections describe the variable elimination and clique tree algorithm for computing marginal posterior probabilities.

Variable Elimination and Exact Inference

The common feature of any inference techniques with Bayesian networks are the manipulation of factors. The underlying operation when computing the probability of some variable in a Bayesian network is marginalizing out variables from a distribution. We can view the as computation on a factor. Let X be a set of variables and $Y \notin X$ be a variable. Next, let $\psi(X, Y)$ be a factor. Marginalizing out Y generates a new factor ψ over X:

$$\psi(X) = \sum_{X} \phi(X, Y) \tag{5.3}$$

A key trick in doing inference on Bayesian networks is exchanging a summation and a product if $X \notin Scope[\phi_1]$:

$$\sum_{X} (\phi_1 \phi_2) = \phi_1 \sum_{X} (\phi_2) \tag{5.4}$$

A marginal probability computation involves taking the product of factors and doing a summation over all the variables except the query variables (the variables you are interested in). So in general, the inference task involves taking a *sum-product* of the form

$$\sum_{Z} \prod_{\phi \in \Phi} \phi \tag{5.5}$$

where Φ is a set of factors.

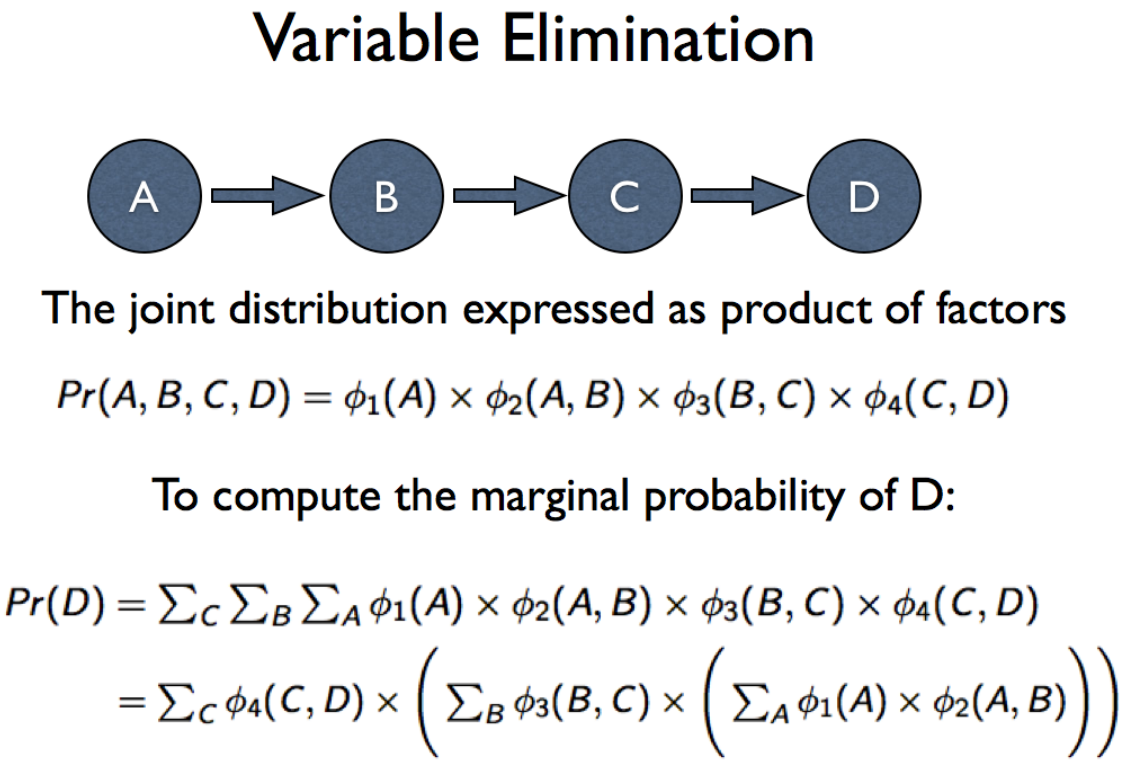


Figure 5.1: **Sum-product variable elimination** - The marginal probability of D is computed by applying equation 5.4

An example of sum-product variable elimination is given in Figure 5.1. To compute the marginal probability of the variable D in the figure, variables A, B, and C are eliminated by applying equation 5.4. When a variable is summed out, all factors that contain that variable in its scope are multiplied, generating a product factor. Then the variable to be eliminated is summed out of this product factor. Again, let X be a set of variables and Φ be a set of factors such that for each $\phi \in \Phi$, $Scope[\phi] \subseteq X$. Let $Y \subset X$ be a set of query variables and the remaining variables be $Z = X - Y$. Then for any elimination ordering of non-query variables, sum-product variable elimination returns a new factor $\phi^*(Y)$:

$$\phi^*(Y) = \sum_Z \prod_{\phi \in \Phi} \phi \tag{5.6}$$

Graph theoretic view of variable elimination

The sum product variable elimination (VE) algorithm is agnostic about the type of graph on which it operates. But the manipulation of factors can be viewed as a series of graph transformations. Let H be an undirected graph whose nodes are variables in the $Scope[\Phi]$ and where there is an edge between nodes if there exists a factor $\phi \in \Phi$ such that X_i and $X_j \in Scope[\phi]$. In other words, the undirected graph H is a fully connected sub-graph over the scope of each factor $\phi \in \Phi$.

In the process of eliminating a variable a new factor ψ is created with X and all the other variables $\mathbf{Y}$ that appear with it in factors. Then X is summed out, creating a new factor τ that contains all the variables $\mathbf{Y}$ but not X. Let Φ_X be the resulting set of factors. When the factor ψ is created, there exist edges between all the variables $Y \in \mathbf{Y}$. Some may have been in the original graph H_Φ, others are introduced as fill edges. When the factor τ is created, X is removed and all its incident edges are

removed. The elimination order is reflected as a series of graphs and every factor that appears in the steps of the VE sum product algorithm is a clique. The set of factors generated in VE is a clique in the induced graph.

The induced graph is the union of all graphs made during the course of variable elimination. Again, let Φ be a set of factors over $\mathbf{X}$ and $\prec$ be an elimination ordering for some subse of variables $\mathbf{X} \subseteq X$. The induced graph $I_{\Phi,\prec}$ is an undirected graph over X where X_i and X_j have an edge between them if they appear in an intermediate factor, *psi*, generated during the course of variable elimination. Each factor ψ used in the course of variable elimination is a *complete subgraph* of the induced subgraph, $I_{\Phi,\prec}$, and is known as a *clique*.

Clique Trees and Exact Inference

In the previous section on variable elimination (VE) we describe the sum product algorithm which sums out variables one at a time. In this section we describe how to use a clique tree as a global data structure to eliminate larger sets of variables.

A cluster graph, U, for a set of factors, Φ, over X, a set of random variables, is an undirected graph whose nodes are associated with a subset $C_i \subseteq X$. The cluster graph must be family preserving such that each factor $\phi \in \Phi$ should be assigned to a cluster such that the scope of of the factor assigned to the cluster should be a subset of the variables in the cluster: $Scope[\phi] \subseteq C_i$. Finally, each edge between a pair of clusters C_i and C_j forms a sepset: $S_{i,j} \subseteq C_i \cap C_j$.

The cluster graph is used as a data structure to help track the factor manipulation process at the heart of inference calculations in Bayesian networks. Each node is a cluster of variables and undirected edges connect clusters that have a non-empty

intersection of variables. Performing variable elimination defines the structure of the cluster graph. In VE, once a variable is eliminated, it doesn't appear in any computations, so the cluster graph induced by variable elimination is a tree. The order of VE defines a direction to the flow of messages between clusters, hence we can define a root. If cluster C_i is on the path from C_j to the root, then C_i is upstream from C_j and C_j is downstream from C_i. Define T be a cluster tree over a set of factors Φ. Its nodes and edges are defined as V_T and E_T. The tree T has the *running intersection property* whenever there is a variable X such that $X \in C_i$ and $X \in C_j$, then X is in every cluster in the (unique) path in T between C_i and C_j.

In variable elimination a variable appears in every factor from the time its first multiplied in (by a factor whose scope contains the variable) till the time is summed out. Let T be a cluster tree induced by a variable elimination ordering over some set of factors Φ. Let C_i and C_j be neighboring clusters such that C_i passes a message τ_i to C_j. The scope of this message is the intersection of variables: $C_i \cap C_j$ So the running intersection property (RIP) is quite helpful. Deriving from the RIP of cluster trees, we define a *clique tree*: Let Ψ be a set of factors over X. A cluster tree over Φ satisfying the *running intersection property* is a clique tree (also called a junction tree or join tree).

Variable Elimination and Clique Trees

Recall again in each step in VE a factor ψ_i is created by multiplying together factors and a variable is eliminated from ψ_i to create a new factor τ_i. This process is continued till the algorithm is finished. The generation of factors can be seen as *message passing* where a factor ψ_i takes incoming message τ_j generated by factors ψ_j, then generates its own message τ_i which in turn is passed onto another factor ψ_l. Each node in

the cluster graph are a set of variables and whose edges have variable scopes with a non-empty intersection.

Sum Product Message Passing

An execution of VE results in a clique tree. But you can start with a clique tree and use it as a data structure to perform variable elimination. The same clique tree can be used multiple times for different executions of VE. So given a tree that satisfies family preservation and the RIP property, you can do can use it in several different ways to do inference with Bayesian networks. The clique tree can be used as a data structure for caching computations so you can do multiple variable eliminations rather than performing VE separately for each variable of interest. Hence the steps to use a clique tree to compute posterior marginal probabilities are as follows:

Step 0: Construct a clique tree given a set of factors Φ

Step 1: Assign each factor to a clique.

Step 2: Calculate initial potentials by multiplying all factors assigned to a clique

Step 3: Denote an arbitrary clique as the root of the tree. Pass messages from the neighbor nodes upwards towards the root. Once complete, pass messages from the root downwards to its neighbors. At this point the clique tree is designated to be *calibrated*, meaning that if a variable appears in more than one clique node, the should agree on the marginal probability of the variables in their sepset.

Step 4: Compute the final beliefs for each clique which means multiplying a nodes initial potential with that of all its incoming messages of its neighbors. Once the

final beliefs are computed, you can extract out the variables of interest to inspect their posterior marginal probability.

The message passing steps described above where messages are passed upwards to the root and downwards towards the leaves is called *sum-product belief propagation*. If c is the cost of message passing, the total cost of of the algorithm is $2c$. If one were to do sum product variable elimination separately for each variable we wish to compute the posterior marginal for, the cost would be nc, where n is the total number of variables. The main advantage of sum-product clique tree calibration algorithm is it computes the posterior probability of all variables using only twice the computation of the upward pass of the same tree. In general, the clique tree algorithm is the best way to calculate posterior probability of multiple query variables [76].

Max product message passing

When constructing a Bayesian network to make inferences about posterior genotype probabilities of samples, rather than computing the marginal posterior probabilities of genotypes, we want to compute the most probably instantiation of genotypes. This is also known as the *maximal a posteriori* (MAP) assignment of genotypes. The same steps outlined in clique tree belief propagation are followed, but instead sums are replaced by maxima. This is called max product belief propagation. The steps of clique tree construction and max product belief propagation are outlined in Figures 5.2 and 5.3 below.

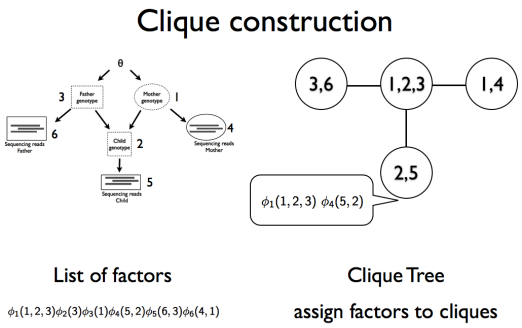


Figure 5.2: A clique tree is constructed from a list of factors. Each factor is assigned to a clique node.

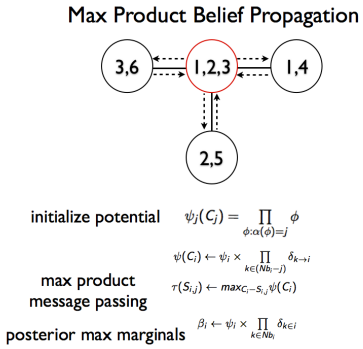


Figure 5.3: Posterior marginals are computed with max-product belief propagation. Once the tree is calibrated, final beliefs and posterior max marginals can be extracted from the tree.

5.2 Results and Discussion

Simulated Pedigrees

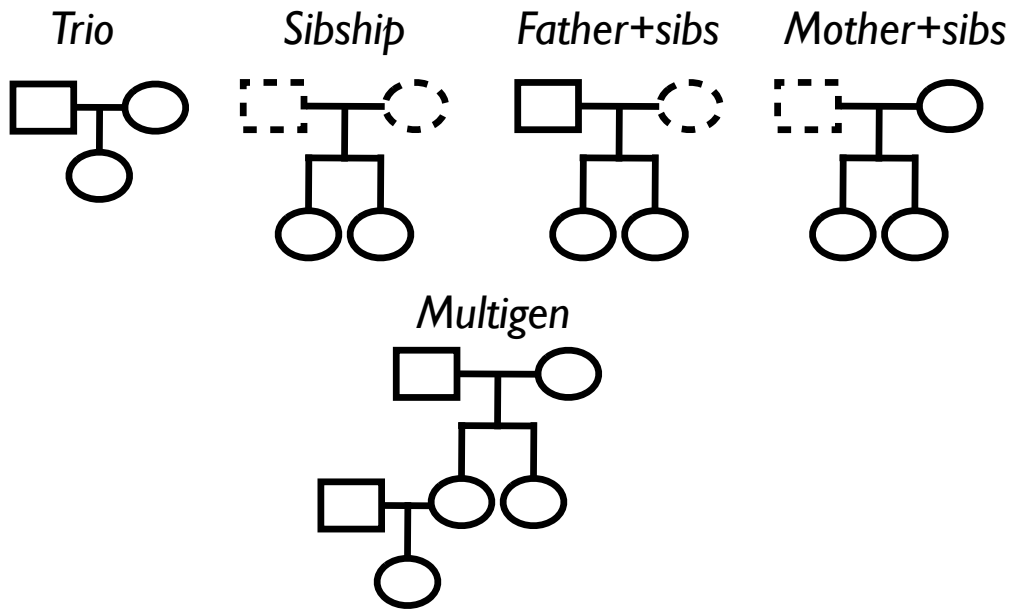


Figure 5.4: **Simulated pedigrees** - Five pedigree structures were simulated.

Pgmsnp was first tested on simulated pedigrees without sequencing or mapping error (see section 5.4 for more details). A set of 5 pedigree structures shown in Figure 5.4 were generated from founder haplotypes and recombinant gametes. The pedigrees in the figure that contain individuals with dashed lines denotes samples whose sequence data was not included as input for variant calling. Each founder individual had a 1 Mbp genome randomly picked from a population of 50 haplotypes simulated via the coalescent with a previously defined demographic model of European ancestry [126]. Non-founder individuals were simulated by modeling recombination with a Poisson distributed number of recombination events to generate recombinant gametes. Paired

end Illumina sequencing reads were generated with the program mason [59]. Each individual's genome was sequenced to 20x coverage and then downsampled to 10x and 5x coverage. Each pedigree structure was examined with Pgmsnp and three other methods: GATK UnifiedGenotyper [30], Famseq [111], and Polymutt [86] at 20x, 10x, and 5x coverage. Famseq is a similar method to Pgmsnp that uses Bayesian networks to model the pedigree sequencing data to compute posterior genotype probabilities. Polymutt is another family aware method that uses the Elston-Stewart algorithm [135] to compute the likelihood of reads in a pedigree. UnifiedGenotyper is a Bayesian variant caller that does not incorporate Mendelian relationships amongst samples.

The two main concordance metrics used to measure the performance of SNP calling of Pgmsnp are non-reference sensitivity (NRS) and non-reference discrepancy (NRD). NRS measures the proportion of sites called variant in the gold standard (comparison) callset that are also called variant in the evaluation callset. Here the evaluation callset are the SNP variant calls returned by Pgmsnp and the three other methods used. Each of these call sets are compared to the gold standard callset, which are genotypes of the samples derived from the coalescent simulation. NRD measures the proportion of differing genotypes between the gold standard and evaluation callsets, at sites called in both data sets, excluding concordant homozygous reference calls. (See section 5.4 for how they are computed).

Trio

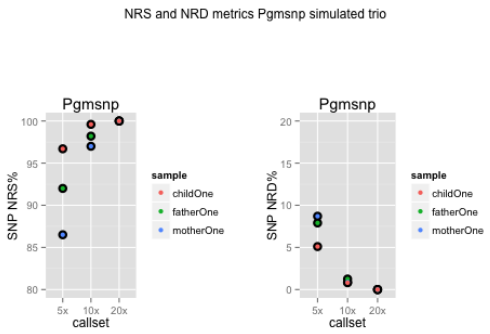


Figure 5.5: NRS and NRD metrics Pgmsnp simulated trio.

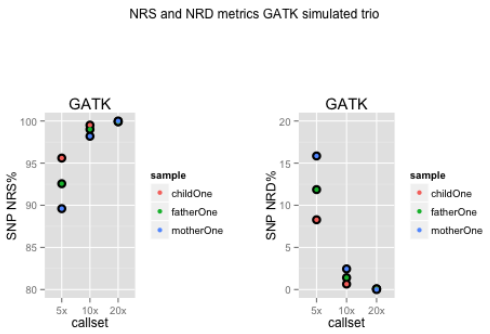


Figure 5.6: NRS and NRD metrics GATK simulated trio.

The first simulated pedigree structure examined was the trio, with coverages at 20, 10, and 5x. At 20x coverage all the methods analyzed have 100 percent sensitivity and zero genotyping discrepancy, as shown in figures 5.5, 5.6, 5.7, 5.8. The performance of each of the pedigree aware methods is indistinguishable to that of GATK. This was a broad pattern seen across all simulated pedigree designs. Things get more interesting at lower coverages. At 5x coverage, each of the pedigree aware methods have slightly higher NRS values (96.7 Pgmsnp), (96.7 Famseq), (96.1 Polymutt) than GATK (95.6) for childOne. The corresponding NRD for GATK childOne 5x calls is 8.28%. The NRD values for Pgmsnp, Famseq, and Polymutt are 5.1, 5.9, and 4.0%. Pgmsnp performs comparable to Famseq and Polymutt in detection sensitivity, but has a 1 % greater genotyping discrepancy. If we look at the genotype concordance matrices for each of the four methods for childOne at 5x coverage, as shown in figures 5.9, 5.10, 5.11, 5.12, incorporating Mendelian inheritance in the genotype priors makes the greatest difference in detecting heterozygotes. While Pgmsnp performs comparably to Polymutt in terms of NRS, the increase in genotype discrepancies in

Pgmsnp can be attributed to incorrectly calling 24 sites as AB heterozygotes (where B is the non-reference allele), when the gold genotype was BB homozygote alternate.

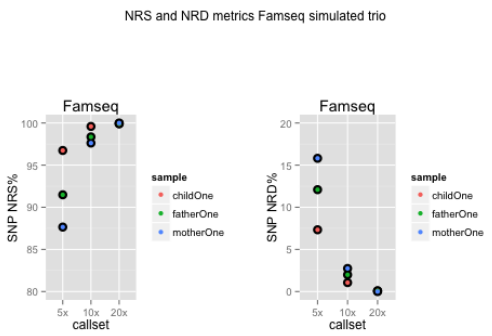


Figure 5.7: NRS and NRD metrics Famseq simulated trio.

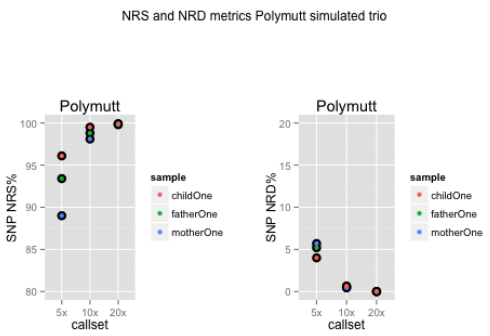


Figure 5.8: NRS and NRD metrics Polymutt simulated trio.

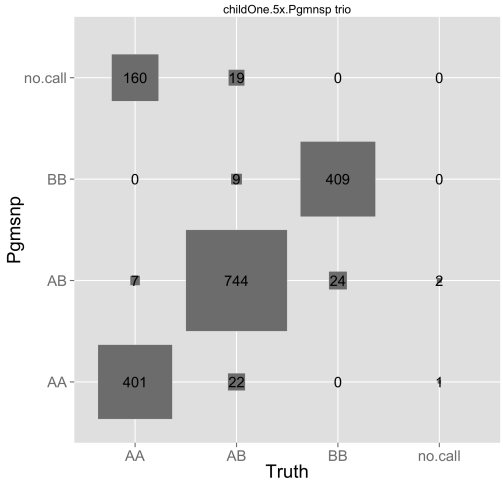


Figure 5.9: Genotype matrix child one, 5x coverage Pgmsnp.

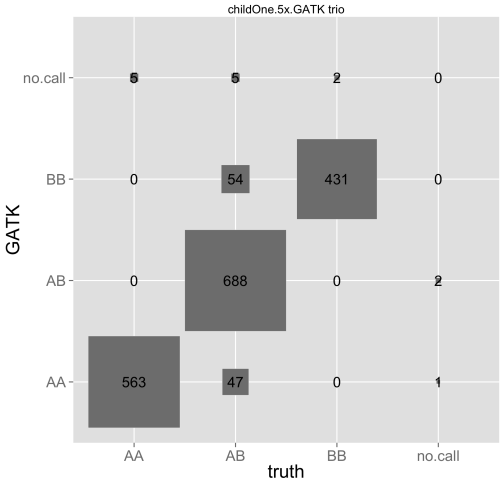


Figure 5.10: Genotype matrix child one, 5x coverage GATK.

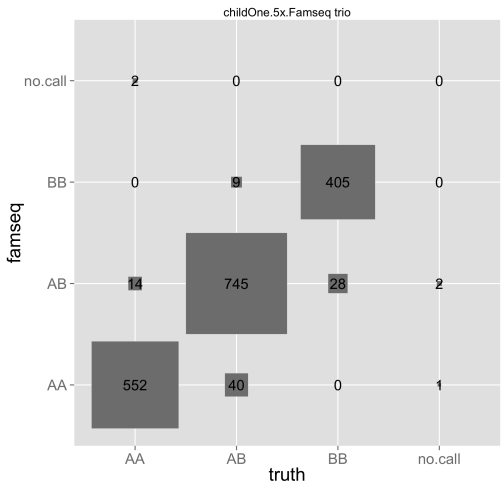


Figure 5.11: Genotype matrix child one, 5x coverage Famseq.

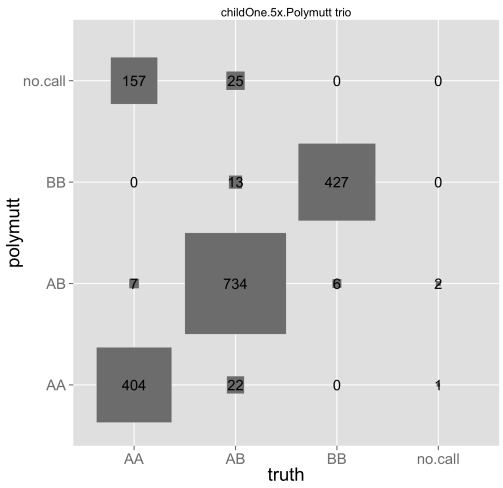


Figure 5.12: Genotype matrix child one, 5x coverage Polymutt.

Sibship

The next simulated pedigree design tested with Pgmsnp and associated methods was a sibship. The parents of the two sibs are the same parents in the simulated trio. The child from the trio has a simulated sibling, referred to as child3 in the preceding figures. The NRS and NRD metrics for all four methods are shown in figures 5.13, 5.14, 5.15, 5.16. As with the trio results, the 20x simulations of the pedigree aware, and standard calling method are identical, with 100 percent sensitivity and zero percent genotype discrepancy. At 10x coverage, the sample denoted as child3, has genotyping discrepancy notably higher than its sibling for both Pgmsnp (10%) and Famseq (12%) derived callsets. Polymutt derived genotype discrepancies are essentially zero. This pattern is interesting because both Pgmsnp and Famseq have identical posterior genotype inference algorithms. At 5x coverage, the NRS values for Pgmsnp, Famseq, and Polymutt are very similar to GATK derived calls for both siblings, but Polymutt's genotyping accuracy is remarkably higher than all three methods. Pgmsnp and Famseq's NRD metrics at 5x coverage parallel each other, with values ranging between 20-45%, nearly 8 times higher than Polymutt. The genotype concordance matrices of all four methods for child3 at 5x, are shown in figures 5.17, 5.18, 5.19, 5.20. The matrices for Pgmsnp and Famseq are nearly identical. Comparing the pedigree aware matrices to GATK, each of the pedigree aware methods correctly call more heterozygotes. The reason for the difference in NRD metrics between Polymutt and each Pgmsnp and Famseq lies in correctly genotyping homozygous non-reference BB genotypes.

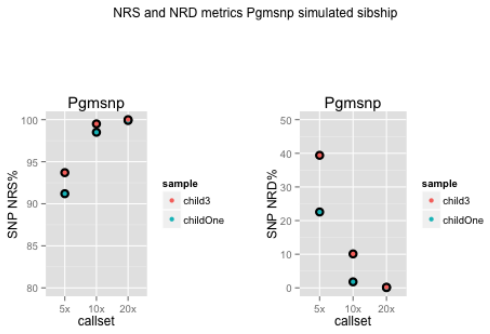


Figure 5.13: NRS and NRD metrics Pgmsnp simulated sibship.

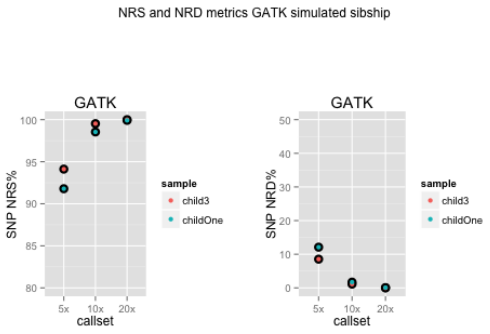


Figure 5.14: NRS and NRD metrics GATK simulated sibship.

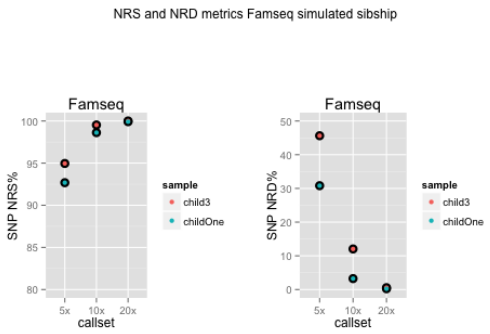


Figure 5.15: NRS and NRD metrics Famseq simulated sibship.

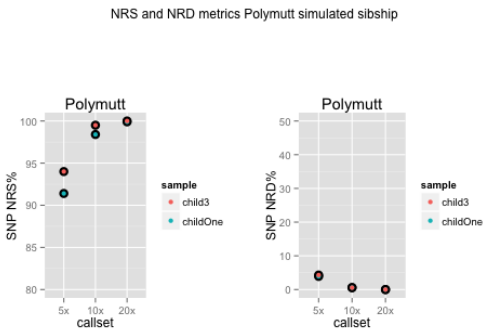


Figure 5.16: NRS and NRD metrics Polymutt simulated sibship.

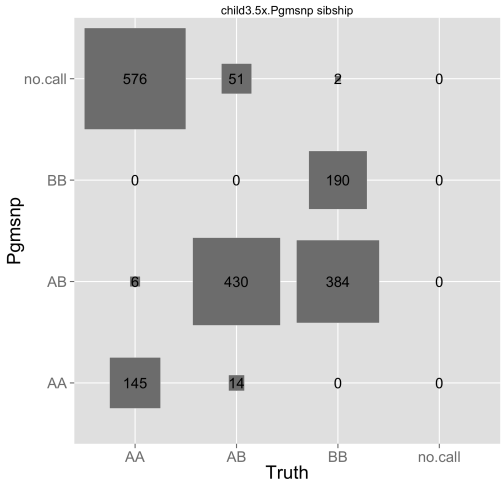


Figure 5.17: Genotype matrix child3 sibship, 5x coverage Pgmsnp.

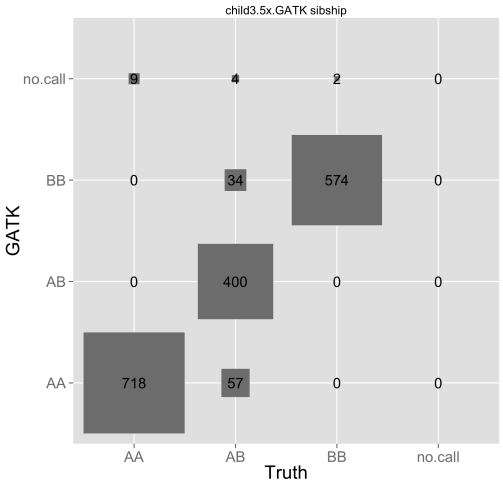


Figure 5.18: Genotype matrix child3 sibship, 5x coverage GATK.

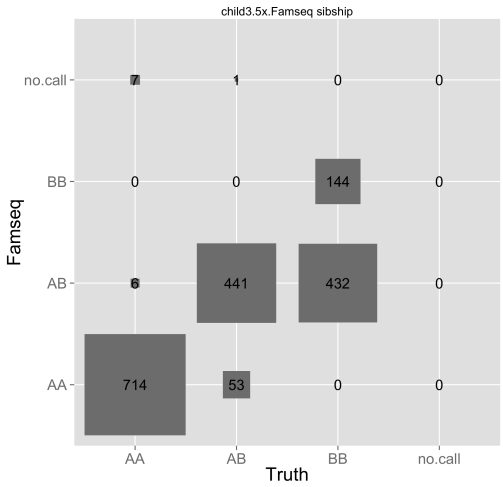


Figure 5.19: Genotype matrix child3 sibship, 5x coverage Famseq.

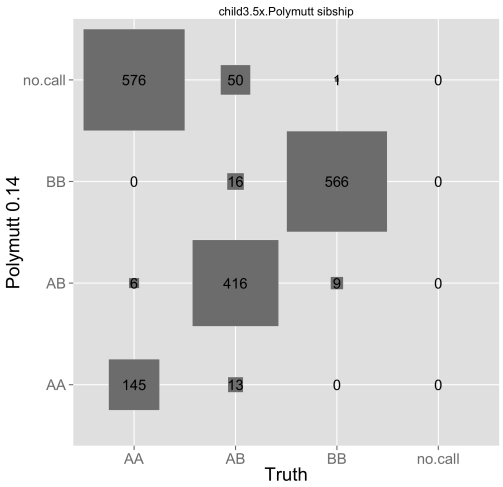


Figure 5.20: Genotype matrix child3 sibship, 5x coverage Polymutt.

Father plus sibs

This pedigree structure took the sibs and added back the father's sequencing data. The NRS and NRD metrics are shown in figures 5.21, 5.22, 5.23, 5.24. Incorporating pedigree awareness doesn't make a difference at high coverage, as both NRS and NRD measures for all 3 pedigree methods are indistinguishable to the results derived from GATK. NRS values at 10x for all methods are very similar to each other, but as with the sibship, child3 has a pronounced increase in NRD for both Pgmsnp (9.3) and Famseq (8.3) compared to Polymutt derived calls (0.65). At 5x coverage, NRD values for Pgmsnp and Famseq derived calls have very similar values, ranging from 6% for fatherOne to 39% for child3. NRS values for each of the pedigree methods is slightly higher that the GATK derived calls for each of the sibs at 5x, but Polymutt's NRD values are the lowest of all methods at 5x coverage. Looking at the genotype concordance matrices of child3 at 5x coverage for all four methods, as shown in figures 5.25, 5.26, 5.27, 5.28, shows a similar pattern to the sibship results. All the pedigree aware methods have better detection power for heterozygote genotypes when compared to GATK. The major stumbling block for both Pgmsnp and Famseq are correctly genotyping BB homozygous non-reference sites.

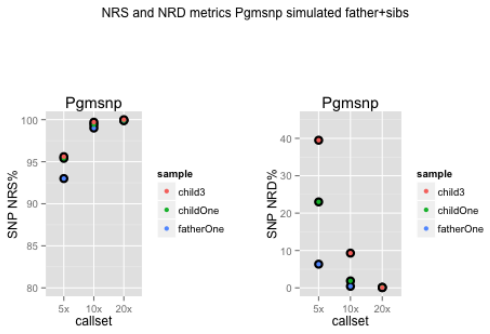


Figure 5.21: NRS and NRD metrics Pgmsnp simulated father+sibs.

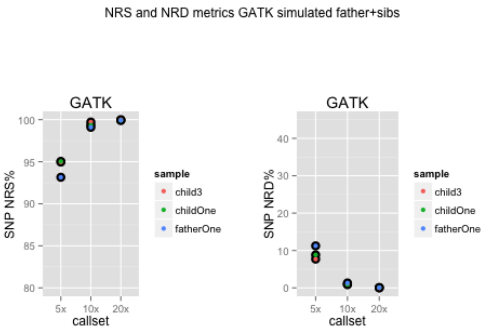


Figure 5.22: NRS and NRD metrics GATK simulated father+sibs.

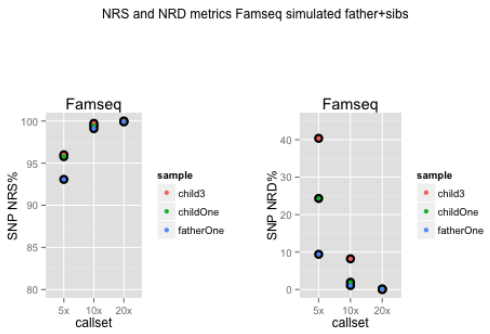


Figure 5.23: NRS and NRD metrics Famseq simulated father+sibs.

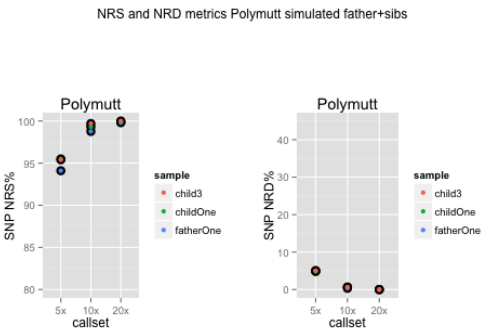


Figure 5.24: NRS and NRD metrics Polymutt simulated father+sibs.

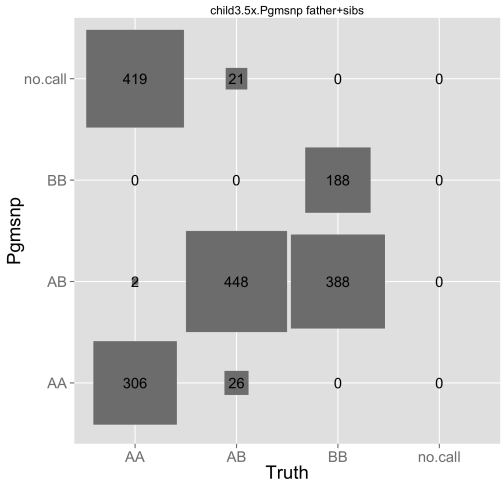


Figure 5.25: Genotype matrix child three father+sibs, 5x coverage Pgmsnp.

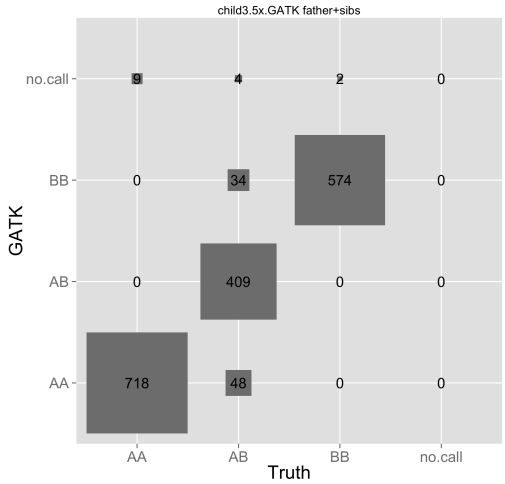


Figure 5.26: Genotype matrix child three father+sibs, 5x coverage GATK.

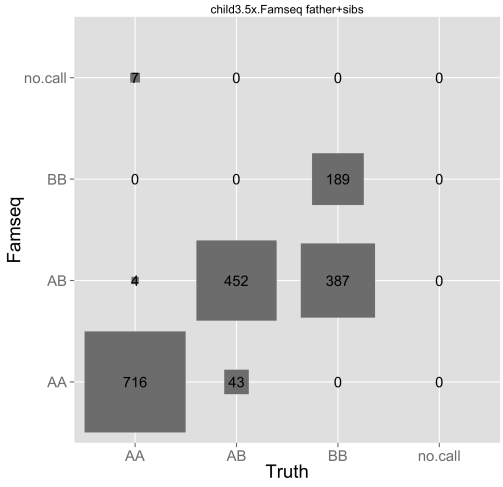


Figure 5.27: Genotype matrix child three father+sibs, 5x coverage Famseq.

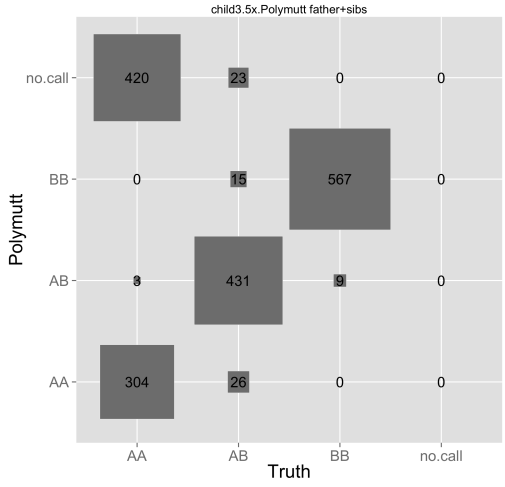


Figure 5.28: Genotype matrix child three father+sibs, 5x coverage Polymutt.

Mother plus sibs

Results of analyzing the simulated sequence data of motherOne with both offspring are very similar to the previous section. The child3 NRD values are considerably higher at 10x for both Pgmsnp and Famseq, as shown in figures 5.29, 5.30, 5.31, 5.32. NRS values are very similar for all methods, suggesting that modeling Mendelian relationships doesn't have as large of an impact as one would expect. The genotype matrices for child3 at 5x are shown in figures 5.33, 5.34, 5.35, 5.36. Again, where Polymutt beats out both Pgmsnp and Famseq is in correctly genotyping homozygous non-reference sites. Many of these sites are incorrectly called as heterozygotes in Pgmsnp and Famseq.

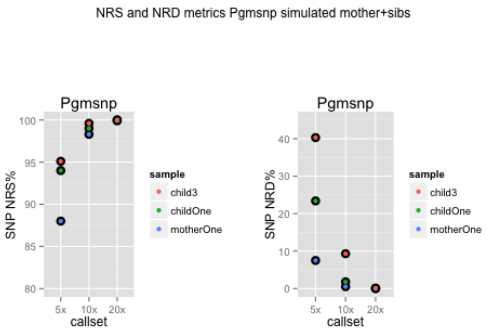


Figure 5.29: NRS and NRD metrics Pgmsnp simulated mother+sibs.

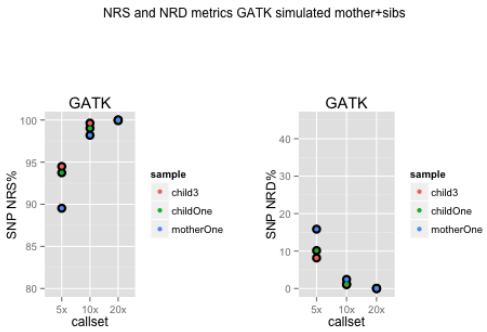


Figure 5.30: NRS and NRD metrics GATK simulated mother+sibs.

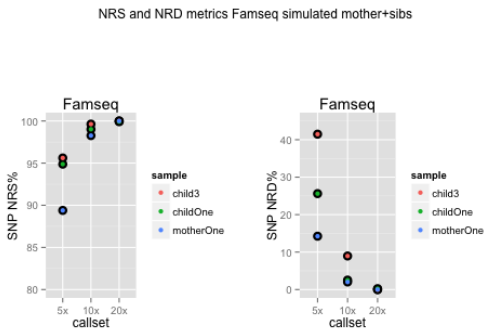


Figure 5.31: NRS and NRD metrics Famseq simulated mother+sibs.

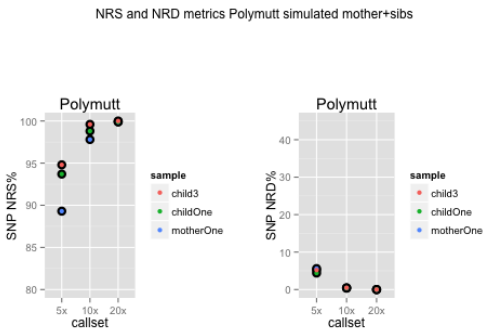


Figure 5.32: NRS and NRD metrics Polymutt simulated mother+sibs.

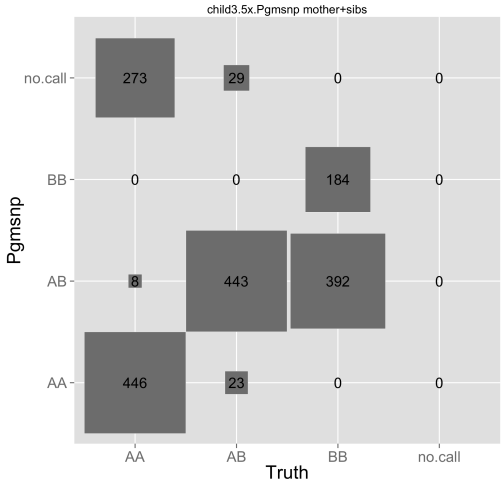


Figure 5.33: Genotype matrix child three mother+sibs, 5x coverage Pgmsnp.

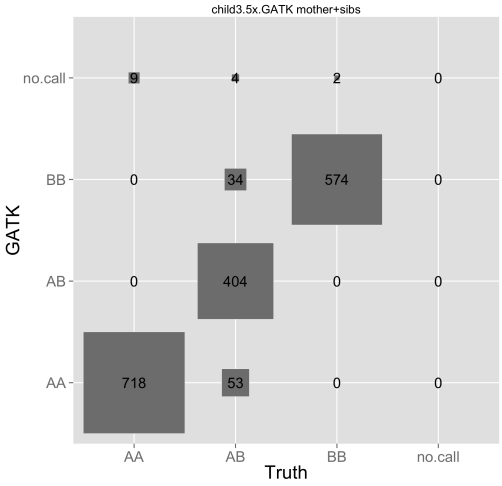


Figure 5.34: Genotype matrix child three mother+sibs, 5x coverage GATK.

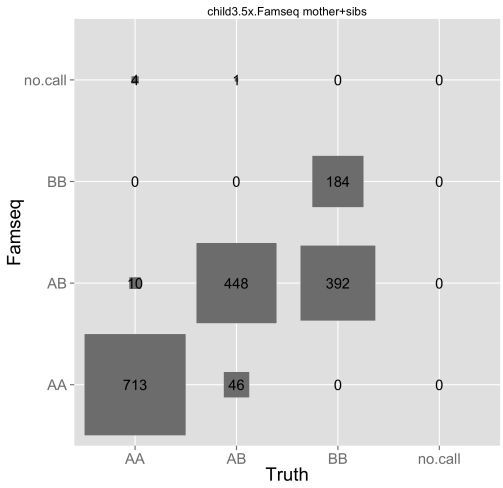


Figure 5.35: Genotype matrix child three mother+sibs, 5x coverage Famseq.

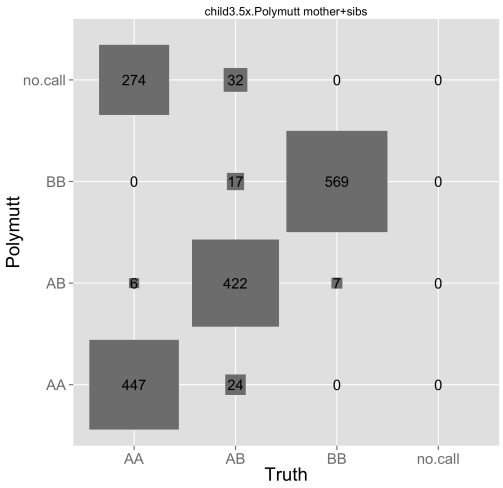


Figure 5.36: Genotype matrix child three mother+sibs, 5x coverage Polymutt.

Multigeneration

The final simulated pedigree structure tested was a three generation pedigree comprised of the founders and sibs of the previous pedigrees, with an additional marry in (marryinOne) who had a simulated offspring (grandchildOne) with childOne. The total size of the pedigree is six individuals. The NRS and NRD metrics for each of the four methods are shown in figures 5.37, 5.38, 5.39, 5.40. The biggest difference that modeling Mendelian inheritance makes is in non-founder individuals at 5x coverage. Each of the 3 non-founder individuals (childOne, child3, and grandchildOne) had higher NRS values in each of the pedigree aware methods than GATK. Comparing Pgmsnp's NRS values at 10x of childOne, child3, and grandchildOne (98.05, 95.6, 95.4) to GATK's for the same samples (95.3, 95.8, 94.0), Pgmsnp has higher sensitivity for two of the three. In terms of genotyping discrepancy, childOne, child3, and grandchildOne have lower genotype discrepancy percentages in Pgmsnp derived calls (.57, .65, .47) than GATK (.84, .73, 1.18). Polymutt's genotype accuracy at 10x is even better for these samples with NRD values of .33,.37, and .63. This NRD differences between GATK and Pgmsnp are even more pronounced at 5x for the the three non-founders, with Pgmsnp's NRD values of 5.4,7.8, and 8.07%, compared to GATK's of 7.32, 8.01, and 10.72%. Polymutt's genotyping accuracy at 5x is lowest of all methods with values of 4.35, 2.52, and 7.0%. The sample grandchildOne genotype matrices from the four methods are show in figures 5.41, 5.42, 5.43, 5.44. Each of the pedigree methods wins out in correctly identifying more heterozygote genotypes. The Pgmsnp and Famseq matrices are nearly identical. Again, as with previous pedigree structures, the reason why Polymutt has better genotyping accuracy is because it accurately distinguishes between AB heterozygotes and BB homozygous non-reference genotypes.

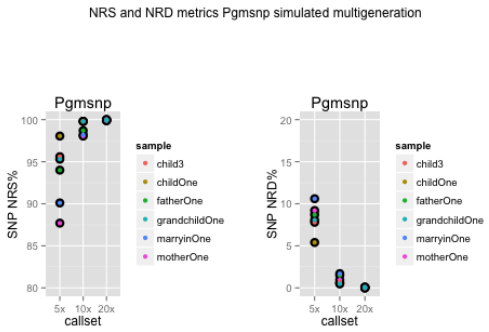


Figure 5.37: NRS and NRD metrics Pgmsnp simulated mutigeneration.

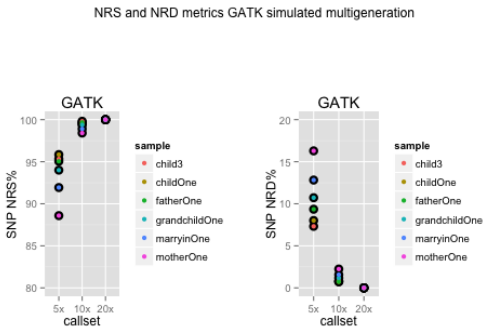


Figure 5.38: NRS and NRD metrics GATK simulated mutigeneration.

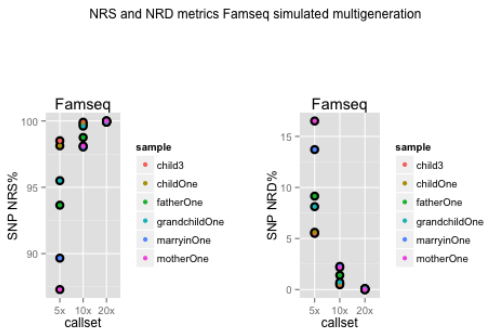


Figure 5.39: NRS and NRD metrics Famseq simulated mutigeneration.

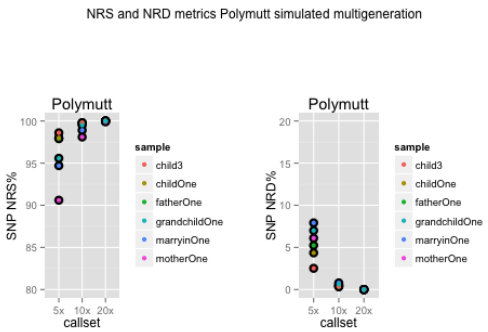


Figure 5.40: NRS and NRD metrics Polymutt simulated mutigeneration.

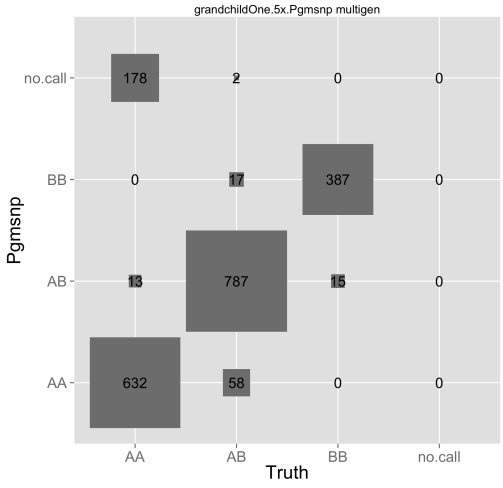


Figure 5.41: Genotype matrix grandchild multigen, 5x coverage Pgmsnp.

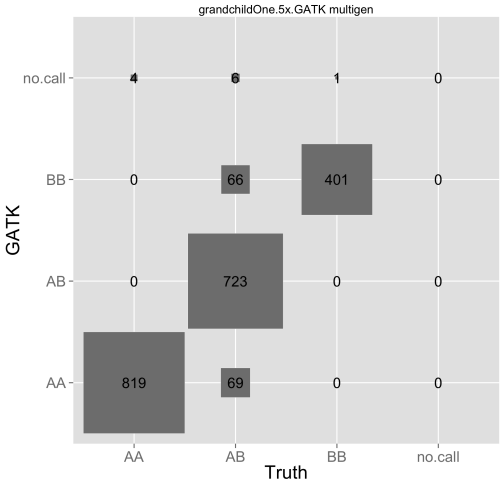


Figure 5.42: Genotype matrix grandchild multigen, 5x coverage GATK.

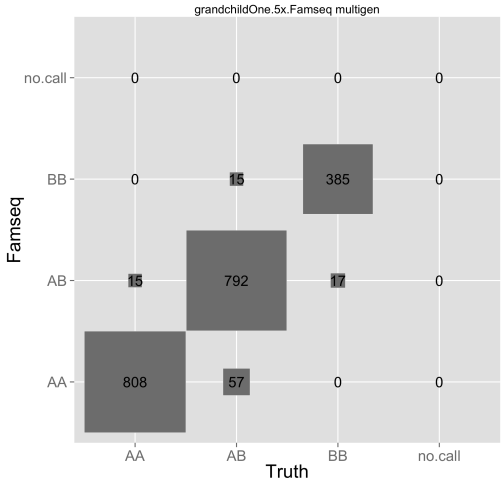


Figure 5.43: Genotype matrix grandchild multigen, 5x coverage Famseq.

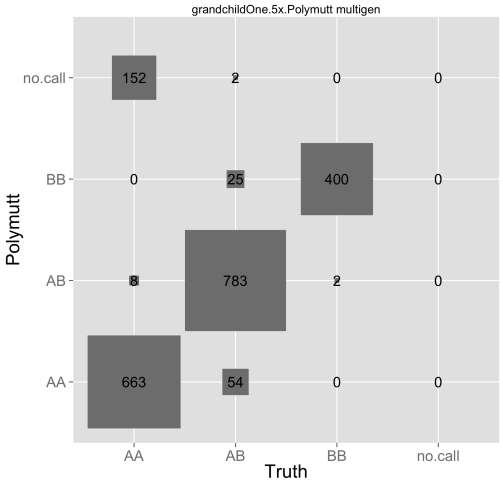


Figure 5.44: Genotype matrix grandchild multigen, 5x coverage Polymutt.

Illumina Platinum Genomes

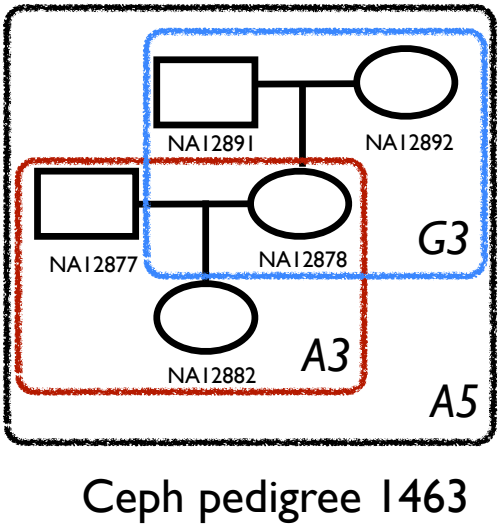


Figure 5.45: **Ceph pedigree 1463** - Three pedigrees examined from the Illumiina Platinum genomes dataset

Pgmsnp was tested on empirical data from the the Illumina Platinum Genomes dataset. Illumina sequenced the 17-member Ceph 1463 pedigree to 50x coverage and released the data to the genomics community as a resource [62]. These 50x genomes were aligned with BWA [30]. Additionally, SNP variant calls were made with GATK [30] on a single sample basis, meaning that the variant calls were not made jointly with all 17 members of the full pedigree. A 5 member subset of the 17 member pedigree was used to test Pgmsnp and is shown in Figure 5.45. The 5 member pedigree is referred to as *A5* and is comprised of individuals NA12891, NA12892, NA12878, NA12887, and NA12882. The two founders and their daughter is referred to as pedigree *G3*, and the marry in to NA12878 and their offspring

is denoted as pedigree *A3*. Each individual's BAM file [90] was downloaded from European Nucleotide Archive. Based on the simulation experiments with Pgmsnp and associated methods, data at high coverage (≥20x), pedigree aware methods perform the same as the standard approach that do not incorporate Mendelian inheritance. Hence, each of the 5 Ceph BAM files were downsampled to 5x and 10x coverage.

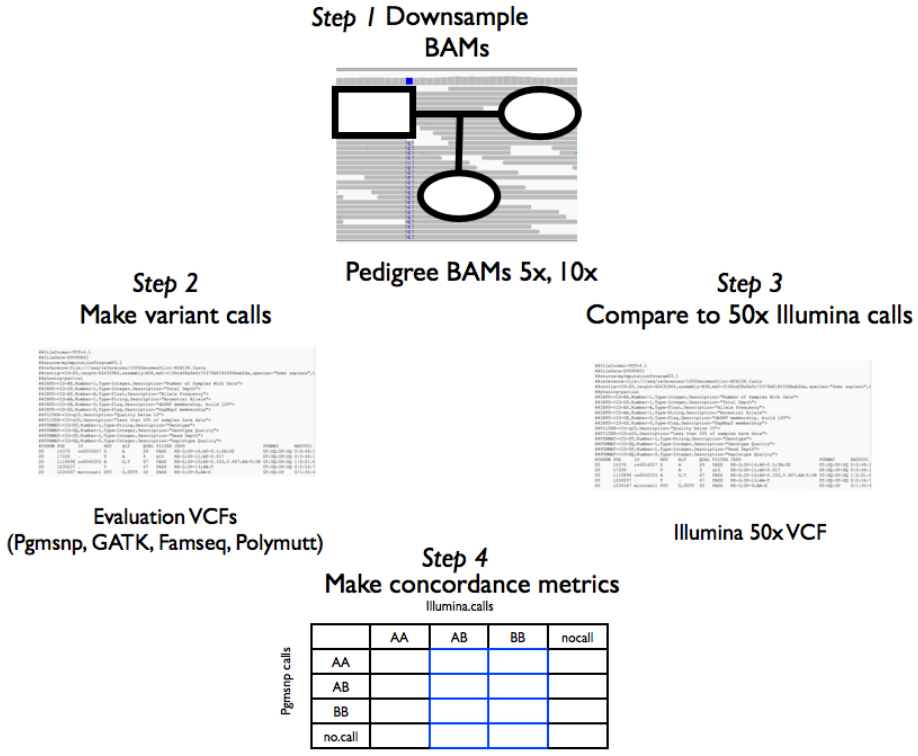


Figure 5.46: **Analysis steps to compare call sets** - Each of the evaluation call sets, Pgmsnp, GATK, Famseq, and Polymutt were compared to the 50x Illumiina Platinum genomes dataset

Each of the downsampled pedigrees (A5, G3, A3) SNP variant calls were made with Pgmsnp, GATK, Famseq, and Polymutt, and then compared to the callset derived from the original 50x Illumina Platinum genomes for chr20 only. This process is shown

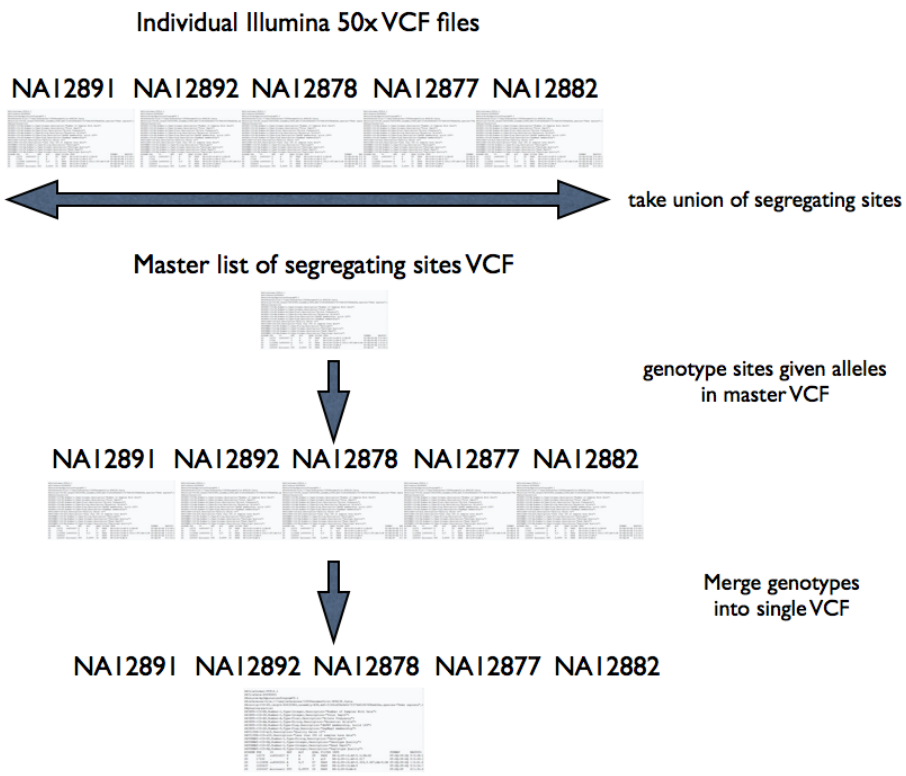


Figure 5.47: **Process used to merge single sample Illumina 50x VCF files** - A master list of sites to genotype in the individual BAM files is created from the union of segregating sites. The sites are genotyped and then merged into a single VCF containing all members of the pedigree.

in Figure 5.46. The reason for analyzing chr20 is that the speed of Pgmsnp, since its implemented in Python, is much slower to run. Performing whole-genome calls would not have been practical. I address this issue in section 5.3. The steps to generate the Illumina 50x derived callset is a bit nuanced. Since the original VCF files derived from the 50x Illumina BAM files were called individually, it was necessary to merge them into a single VCF. The process to do this involves three steps. First is to take the union of polymorphic sites in each individual VCF and create a master list VCF containing their positions and alleles. Next step is to genotype each of the individual samples at the sites contained in the master list using GATK [30]. Finally, the last step is to merge each of genotyped samples into a single VCF containing all members

of the pedigree. This process is depicted in Figure 5.47. This process was applied to the A5, G3, and A3 pedigrees to create the gold comparison callset to which the evaluation call sets derived from Pgmsnp and other methods were compared to.

Venn Analysis of evaluation callsets

Venn analysis looks at the site level concordance between the evaluation and gold comparison call sets. A site is either in the intersection of calls (meaning that the site was in both the evaluation and gold callset) or the unique fraction (meaning that the site was called by one method, but not the other). Table 5.1 show the numbers of SNPs in the unique fraction of Pgmsnp calls, intersection, and unique fraction of Illumina 50x calls. The numbers in parentheses indicate the transition-transversion (TsTv) ratio of the callset. Salient points to take away from the table are that the TsTv values are higher for sites in the intersection than the unique fraction of Pgmsnp. The size of the unique fraction of the Illumina calls is much smaller for the 10x coverage call sets, which is most likely attributable to the higher number of reads in the BAM file. The size of unique fraction of the Pgmsnp calls is quite large, relative to the unique fraction of Illumina and the TsTv values are much lower, indicating these are potentially low quality calls. Closer examination reveals that approximately 80% of Pgmsnp unique fraction sites across all experiments were called in the Illumina 50x VCF, but were filtered out Illumina when applying GATK's Variant Quality Score Recalibration (VQSR) algorithm. VQSR was not applied to the Illumina callset(s) for each of the pedigrees analyzed after merging individual call sets (see Figure 5.47).

Table 5.2 shows the Venn analysis of GATK applied to the 5x and 10x BAMs to the three pedigrees analyzed. The intersection calls are a bit higher than Pgmsnp in table 5.1. Again, the unique fraction of the GATK calls are quite sizeable, but like in the Pgmsnp Venn results, many of these sites are present in the Illumina callset, but were

Pedigree	Pgmsnp unique	Intersection	Ilumina unique	Coverage
A5	27995 (1.04)	101757 (2.27)	7725 (2.25)	5x
A5	42340 (1.0)	108098 (2.27)	1384 (2.33)	10x
A3	19941 (1.13)	86580 (2.27)	8427 (2.25)	5x
A3	31621 (1)	93543(2.26)	1464 (2.25)	10x
G3	19550 (1.0)	85007 (2.26)	8412 (2.27)	5x
G3	31308 (1.01)	91978 (2.26)	1441 (2.20)	10x

Table 5.1: Pgmsnp site level Venn analysis. TsTv ratios are shown in parentheses.

Pedigree	GATK unique	Intersection	Illumina unique	Coverage
A5	54562 (1.08)	103028 (2.27)	6454 (2.22)	5x
A5	67462 (1.05)	108405 (2.27)	1077 (2.34)	10x
A3	42547 (1.13)	88650 (2.27)	6357 (2.20)	5x
A3	53257 (1.09)	93914 (2.27)	1093 (2.20)	10x
G3	43232 (1.12)	87085 (2.27)	6334 (2.24)	5x
G3	54260 (1.07)	92330 (2.26)	1089 (2.34)	10x

Table 5.2: GATK site level Venn analysis. TsTv ratios are shown in parentheses.

filtered out by VQSR. The Famseq Venn results shown in table 5.3 are essentially the same as the GATK results. Famseq takes as input a GATK derived VCF, records the genotype likelihoods in the file, and adjusts the genotypes, taking into account Mendelian inheritance. Since Venn analysis is site based, it would be expected that the numbers would be relatively unchanged. Finally, 5.4 shows the Venn analysis results for Polymutt. The TsTv ratios in the intersection are all very similar to the previous methods analyzed, but the TsTv values for the unique fraction of Ilumina calls is slightly lower in this analysis then the others. Overall, the intersection fractions of all methods when compared to the Illumina 50x calls are similar in size and TsTv

Pedigree	Famseq unique	Intersection	Illumina unique	coverage
A5	54424 (1.09)	103028 (2.27)	6454 (2.22)	5x
A5	67227 (1.04)	108404 (2.27)	1078 (2.34)	10x
A3	42478 (1.13)	88648 (2.27)	6359 (2.20)	5x
A3	53134 (1.09)	93913 (2.27)	1094 (2.20)	10x
G3	43166 (1.13)	87085 (2.27)	6334 (2.24)	5x
G3	54150 (1.07)	92330 (2.26)	1089 (2.34)	10x

Table 5.3: Famseq site level Venn analysis. TsTv ratios are shown in parentheses.

Pedigree	Polymutt unique	Intersection	Illumina unique	coverage
A5	42359 (1.0)	102368 (2.28)	7114 (2.05)	5x
A5	50742 (1.0)	107970 (2.27)	1512 (1.81)	10x
A3	32089 (1.04)	87554 (2.28)	7453 (2.05)	5x
A3	40102 (1.32)	93572 (2.27)	1435 (1.78)	10x
G3	32439 (1.40)	86002 (2.28)	7417 (2.06)	5x
G3	40540 (1.30)	91981 (2.27)	1438 (1.86)	10x

Table 5.4: Polymutt site level Venn analysis. TsTv ratios are shown in parentheses.

ratio values. The same broad pattern can be said about the unique fraction of the Illumina 50x calls. The sizes of the unique fractions of the evaluation 5x and 10x call sets are inflated due to not applying the same VQSR filters which Illumina did when generating the single-sample variant calls.

Ceph A5 genotype concordance

The previous section detailed the results of site level concordance, here we discuss genotypic concordance of the A5, A3, and G3 pedigrees for Pgmsnp and associated methods. Genotypic concordance is measured by the NRS and NRD metrics, described in Figure 5.72 in section 5.4 of the chapter. When the genotype concordance matrix is constructed, there is an underlying empirical distribution of site quality scores, which is denoted in the VCF file in the QUAL column. The number is the Phred scaled probability that the site is not a variant. High QUAL values indicate high confidence calls. NRS and NRD metrics can be computed at various QUAL cutoffs and plotted. (See section 5.4 for more details). Figure 5.48 shows the NRS vs. NRD as a function of QUAL plotted for the A5 pedigree for the Pgmsnp derived calls at 5x and 10x coverages. Similar plots for other methods for the A5 and remaining pedigrees are shown in the Appendix.

Table 5.5 shows the maximum NRS values and associated NRD and QUAL values. A similar table for 10x coverage results is shown in the Appendix. Salient features

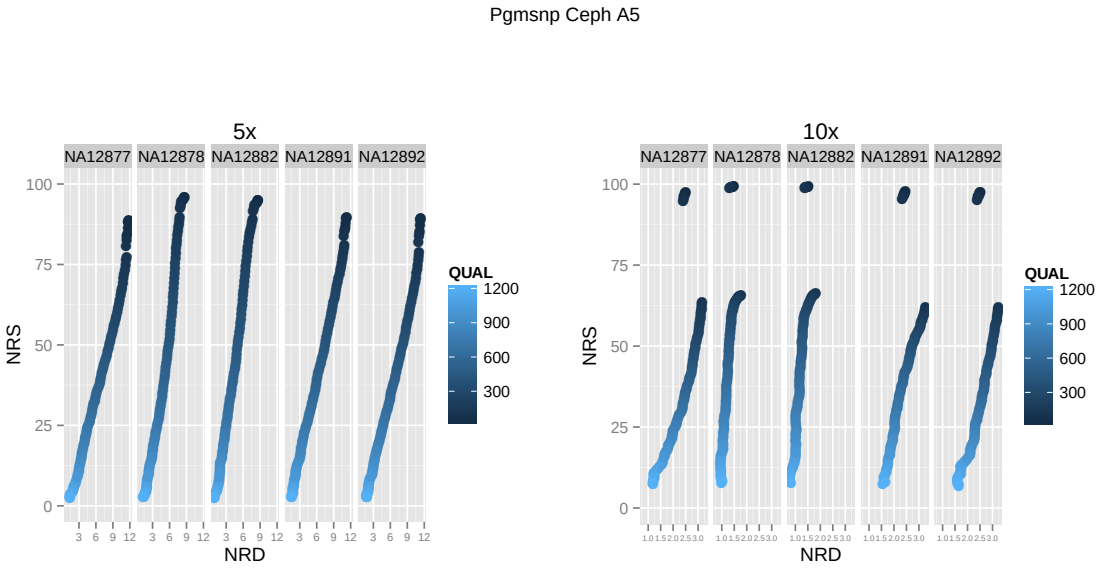


Figure 5.48: **Pgmsnp metrics Ceph A5** - Pgmsnp NRD and NRS metrics as a function of QUAL from Ceph-A5 pedigree

of the table are that Pgmsnp has a higher NRS and lower NRD value than GATK derived calls for the two non-founder individuals (NA12882 and NA12878). Famseq has the highest NRS values overall for these individuals, but when also considering NRD, Polymutt derived calls have the lowest genotype discrepancy percentages. The NRS and NRD values for the founder individuals are highly correlated between Pgmsnp and Famseq, and have lower NRS and higher NRD values, when compared to GATK derived calls. In contrast, Polymutt has higher sensitivity and lower genotype discrepancies for non-founder individuals when compared to GATK.

The genotype concordance matrices for the two non-founders (NA12882 and NA12892) are show starting in Figure 5.49 for NA12878 and in Figure 5.53 for NA12882. Similar to the simulation results, the biggest gain in modeling Mendelian

	Metric	NA12882	NA12877	NA12878	NA12891	NA12892	QUAL
Pgmsnp	NRS	95.15	88.67	95.94	89.68	89.37	10
	NRD	8.66	11.78	8.65	11.32	11.37	10
GATK	NRS	93.57	91.77	93.42	91.85	91.73	10
	NRD	10.17	9.30	10.13	9.60	9.35	10
Famseq	NRS	96.11	88.58	96.95	89.44	89.08	10
	NRD	7.4	12.02	7.58	11.62	11.71	10
Polymutt	NRS	95.92	93.11	96.54	93.83	93.30	10
	NRD	6.03	6.80	4.98	6.57	7.02	10

Table 5.5: Ceph A5 5x callset metrics

inheritance with Pgmsnp is in correctly identifying 3000 more heterozygotes than using the standard approach of GATK, which assumes samples are unrelated. All four methods have a large number of sites not called by Illumina. The majority of these sites are ones that were VQSR filtered by Illumina. Comparing the matrices of Pgmsnp and Polymutt, the biggest difference is that Pgmsnp has nearly 4x-6x greater number of incorrectly called AB heterozygotes that were called homozygous non-reference BB in Illumina. To investigate the incorrectly called NA12878 AB genotypes further, the genotypes of her parents were examined at these sites. Approximately 40% of these sites had an incorrectly called paternal genotype, 40% had an incorrectly called maternal genotype, and the remaining 20% were evenly split in either both parental genotypes being incorrect, or both parents being called correctly. When either parent's genotype wasn't called correctly at these sites, the vast majority were incorrectly called as AA homozygous reference genotypes, when the truth genotype was AB heterozygote. Famseq also had the same pattern of calling BB sites incorrectly as AB for both samples. Nearly 90% of these sites in the Famseq callset overlap the same category of incorrectly called sites in the Pgmsnp callset. The parental genotypes at sites in the Polymutt callset in the same error class for NA12882 and NA12878 were called correctly at 55% of the sites in NA12882

and 46% of the sites in NA12878. GATK does a much better job at correctly calling BB homozygous non-reference genotypes than all the pedigree aware methods. The differences between how the genotype posterior marginal probabilities in the two Bayesian network algorithms (Pgmsnp and Famseq) and Polymutt computes these values needs to be investigated further.

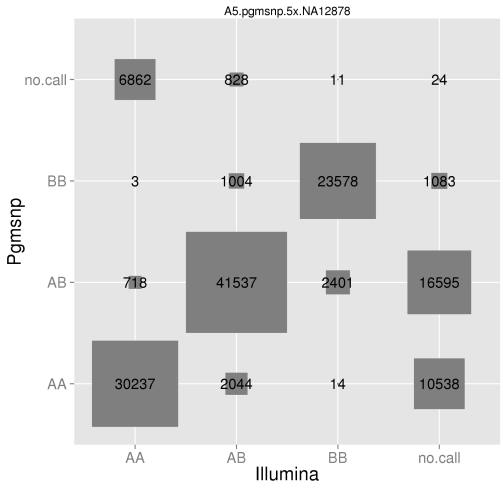


Figure 5.49: Pgmsnp NA12878 genotype concordance matrix A5 pedigree 5x coverage

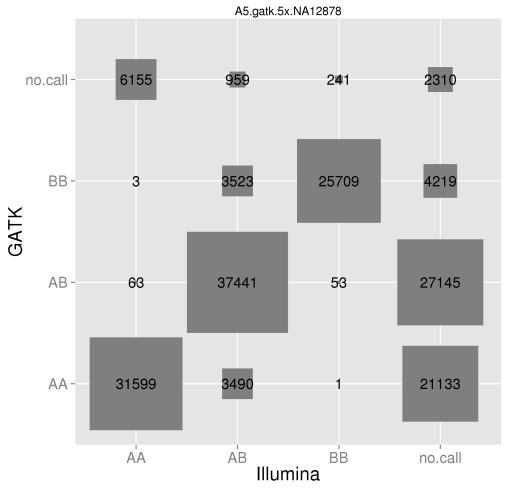


Figure 5.50: GATK NA12878 genotype concordance matrix A5 pedigree 5x coverage

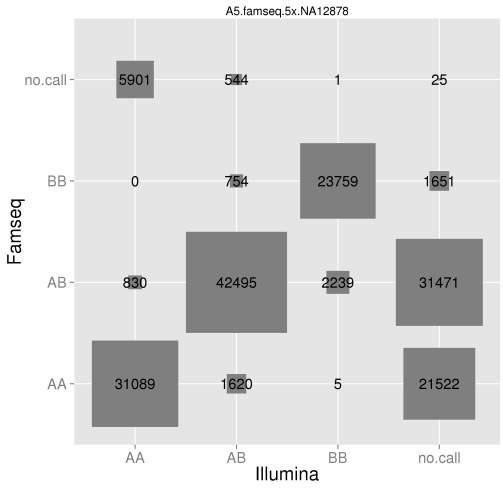


Figure 5.51: Famseq NA12878 genotype concordance matrix A5 pedigree 5x coverage

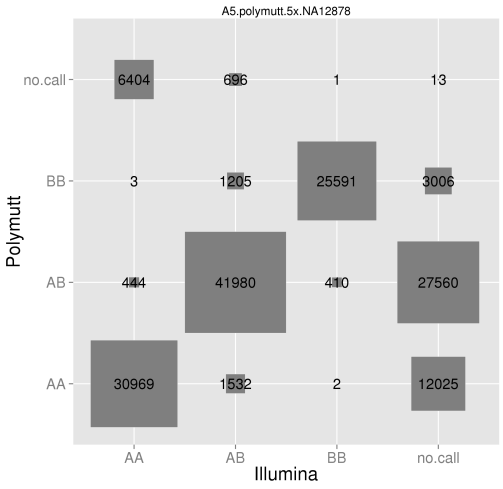


Figure 5.52: Polymutt NA12878 genotype concordance matrix A5 pedigree 5x coverage

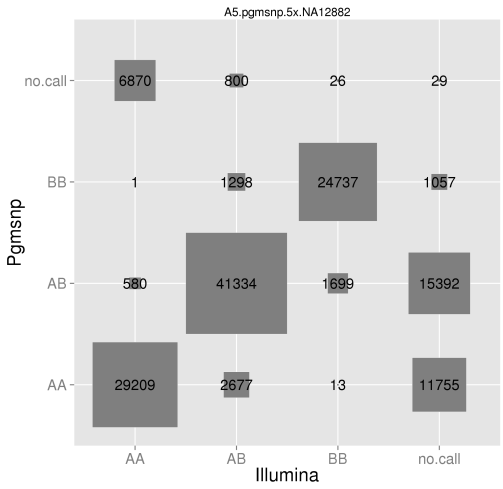


Figure 5.53: Pgmsnp NA12882 genotype concordance matrix A5 pedigree 5x coverage

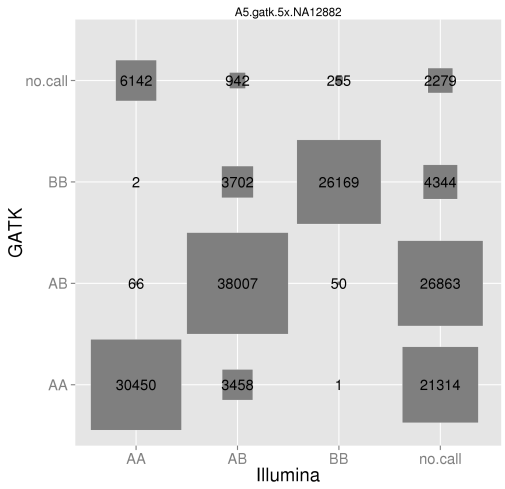


Figure 5.54: GATK NA12882 genotype concordance matrix A5 pedigree 5x coverage

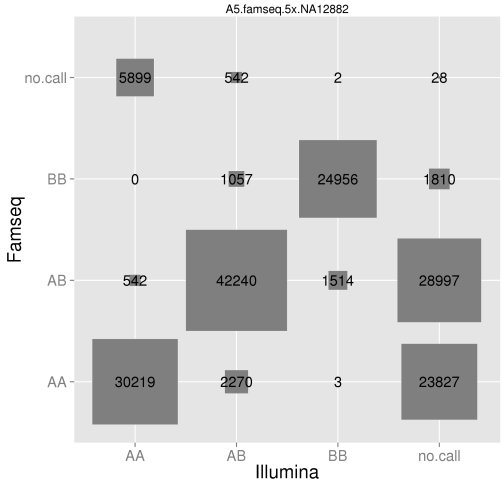


Figure 5.55: Famseq NA12882 genotype concordance matrix A5 pedigree 5x coverage

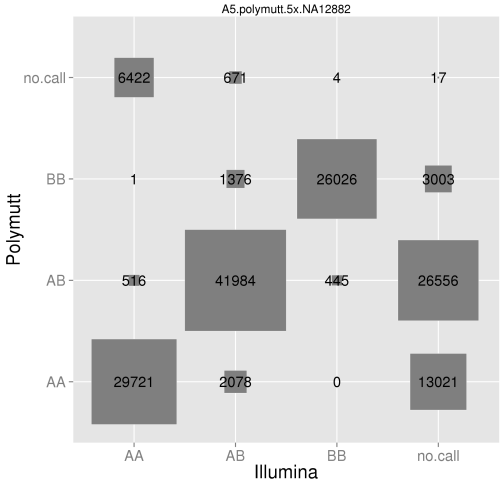


Figure 5.56: Polymutt NA12882 genotype concordance matrix A5 pedigree 5x coverage

Ceph A3 genotype concordance

The difference between the A5 and A3 pedigrees is that the two grandparental founders are removed, and their daughter, NA12878, is treated as a founder in the A3 structure, along with marry in NA12877. NRS vs. NRD values as a function of QUAL values are shown in 5.57. Its clear from looking at the graphs for both 5x and 10x coverage, the offspring NA12882 achieves a higher maximum NRS value than either of the the parents for Pgmsnp. Similar plots for the other methods tested are shown in the Appendix for the A3 pedigree.

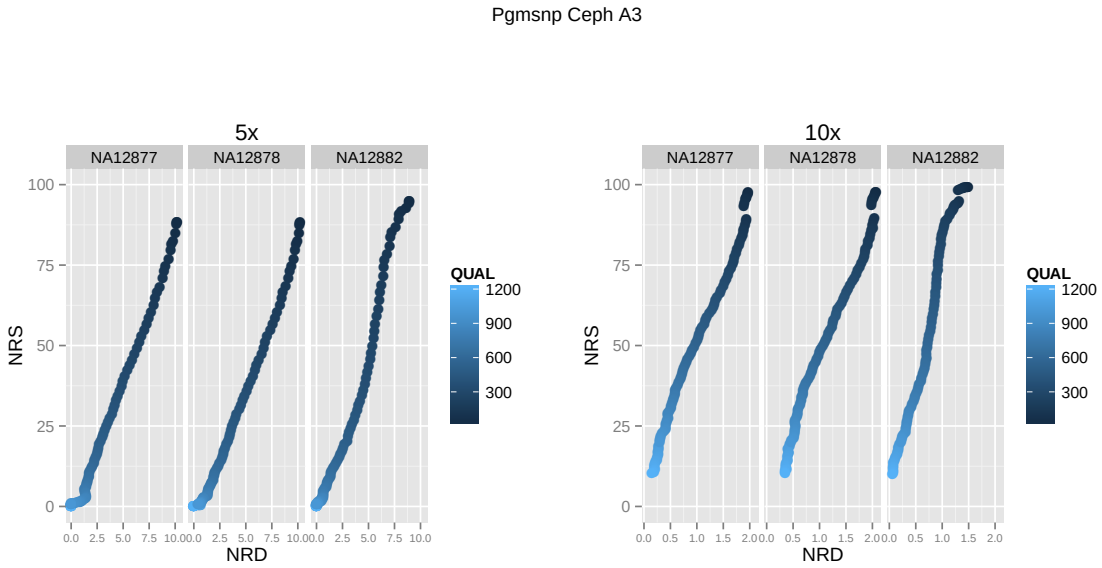


Figure 5.57: **Pgmsnp metrics Ceph A3** - Pgmsnp NRD and NRS metrics as a function of QUAL from Ceph-A3 pedigree

Table 5.6 show the maximum NRS values achieved with all four methods, along with associated NRD and QUAL values. As with the A5 pedigree, the offspring individual has better sensitivity and genotype discrepancy metrics than the parents

	Metric	NA12882	NA12877	NA12878	QUAL
Pgmsnp	NRS	94.82	88.41	88.32	10
	NRD	8.94	10.18	10.22	10
GATK	NRS	93.31	90.8	90.65	10
	NRD	9.61	8.68	8.76	10
Famseq	NRS	95.87	88.89	88.67	10
	NRD	8.05	10.17	10.42	10
Polymutt	NRS	95.58	92.47	92.42	10
	NRD	5.92	6.36	6.47	10

Table 5.6: Ceph A3 5x callset metrics

for Pgmsnp. Famseq achieves the highest NRS value in the child NA12882, but Polymutt has better genotyping accuracy than Pgmsnp and Famseq. Pgmsnp and Famseq both use the same Bayesian network framework for calculating posterior genotype marginals, but results suggest modeling the data as a Bayesian network doesn't improve sensitivity or accuracy for founder individuals. This is clearly shown if we compare the Pgmsnp NRS values of NA12878 in the A5 pedigree, which is shown in Table 5.5 , and has a value of 95.94%, compared to 88.32% in the A3 pedigree, where its treated as a founder. The same pattern is seen in Polymutt results for NA12878 in the A5 pedigree, where NA12878 has an NRS of 96.54% compared to 92.42%. in A3.

Inspecting the genotype concordance matrix of the child NA12882 of the A3 pedigree, starting in Figure 5.58 for Pgmsnp, we see again that modeling Mendelian inheritance with Pgmsnp makes the biggest gain in identifying AB heterozygotes correctly when you compare it to GATK derived calls. The differences in genotype accuracy between Pgmsnp and Polymutt can again be attributed to incorrectly called AB heterozygotes in Pgmsnp that were correctly called as BB homozygous non-reference in Polymutt. Still, GATK beats out the three other pedigree aware methods in this category, as it

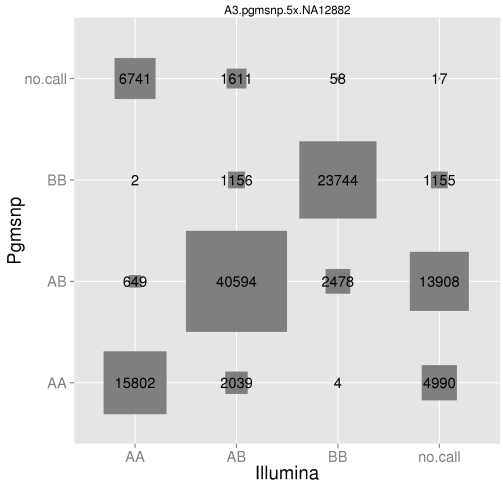


Figure 5.58: Pgmsnp NA12882 genotype concordance matrix A3 pedigree 5x coverage

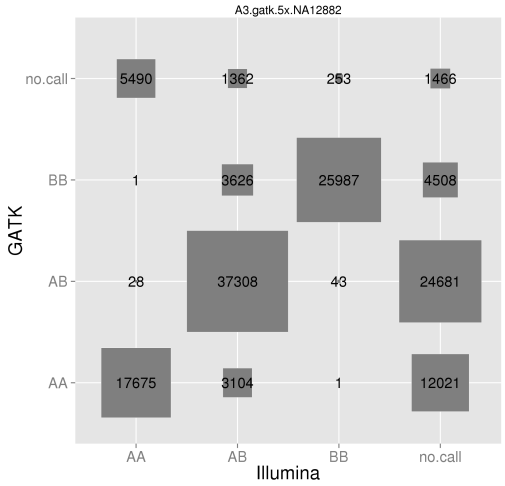


Figure 5.59: GATK NA12882 genotype concordance matrix A3 pedigree 5x coverage

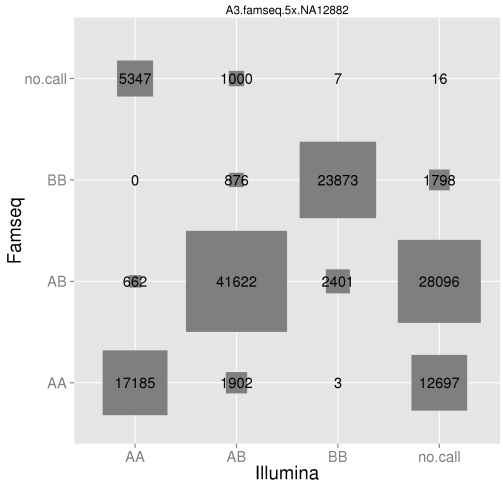


Figure 5.60: Famseq NA12882 genotype concordance matrix A3 pedigree 5x coverage

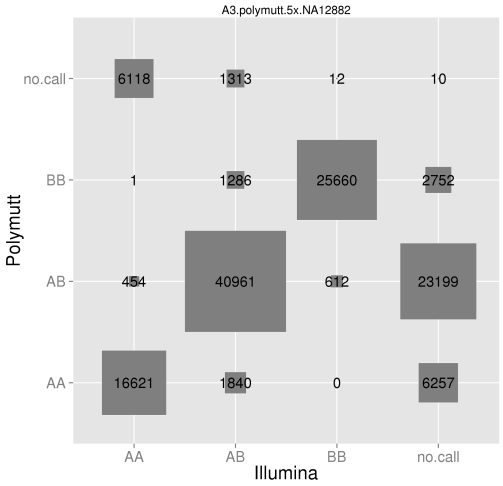


Figure 5.61: Polymutt NA12882 genotype concordance matrix A3 pedigree 5x coverage

incorrectly called only 43 BB sites as AB heterozygotes. Again, as with the A5 genotypes, a similar pattern emerges. The parental genotypes of NA12882 are incorrectly called at 80% of these sites in the Pgmsnp callset (split evenly between maternal and paternal genotypes). The majority of these incorrectly called genotypes are misclassified as homozygote AA genotypes, when in fact they are AB heterozygotes. 84% of the sites in the Famseq callset incorrectly called as AB are shared by the same

error class in the Pgmsnp callset. Examining the parental genotypes at the sites in the Polymutt callset incorrectly called as AB (BB truth), 40% of them are correctly called as either heterozygous or homozygous non-reference.

Ceph G3

The G3 pedigree comprises the two founder individuals NA12891 and NA12892 along with their daughter NA12878. Figure 5.62 shows NRS and NRD values plotted as a function of QUAL values for 5x and 10x coverage. Again, the offspring NA12878 achieves higher maximal NRS with a lower NRD value than compared to either of its parents. Table 5.7 shows the maximal NRS values achieved by all four methods, along with NRD and associated QUAL values at 5x coverage. Pgmsnp achieves higher sensitivity for NA12878 and better genotyping accuracy than GATK, but this is not the case for the two founders. Famseq achieves the highest NRS value for NA12878 with a value of 95.73%, but its NRD percentage is similar to Pgmsnp. Polymutt achieves the best balance between sensitivity and genotyping accuracy. Treating NA12878 as a non-founder with its parents included achieves better sensitivity and genotype accuracy in the G3 and A5 pedigree (see Table 5.5) than treating it as a founder in the A3 pedigree for Pgmsnp and Famseq call sets. Polymutt calls for NA12878 had slightly better metrics in the A3 pedigree (see Table 5.6).

Figures 5.63 through 5.66 show the genotype concordance matrix for the G3 5x NA12878 calls. Incorporating Mendelian inheritance makes a difference in correctly calling heterozygote AB sites for Pgmsnp when compared to GATK. This is true as well for Famseq and Polymutt. A similar pattern emerges again when comparing the differences between Pgmsnp and Polymutt derived calls where Pgmsnp has 4x more incorrectly called AB heterozygotes, whose truth genotype is BB, when compared to Polymutt. The same can be said when comparing Famseq to Polymutt calls.

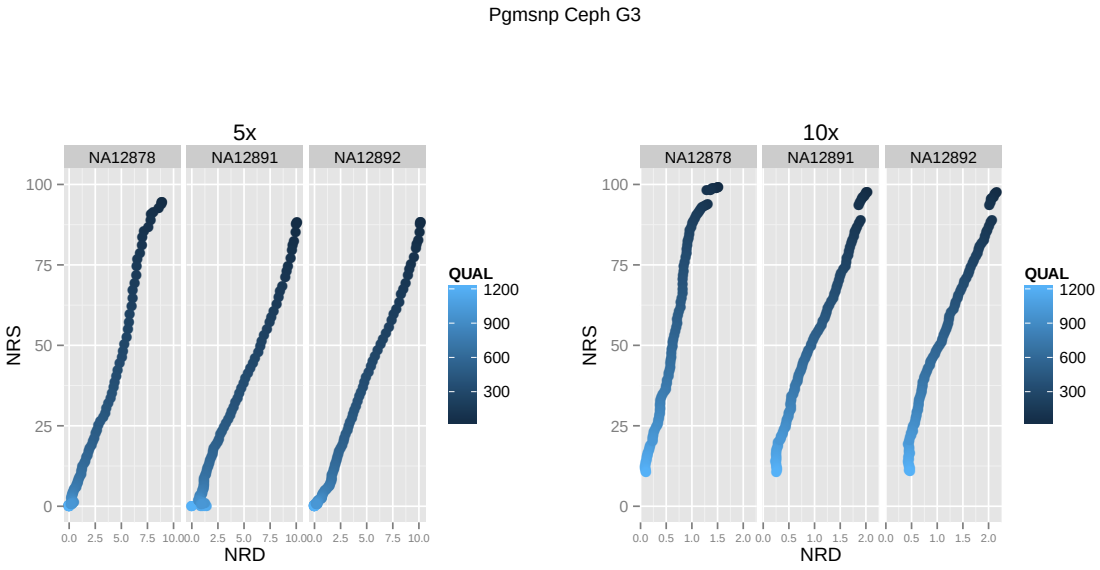


Figure 5.62: **Pgmsnp metrics Ceph G3** - GATK NRD and NRS metrics as a function of QUAL from Ceph-G3 pedigree

	Metric	NA12878	NA12891	NA12892	QUAL
Pgmsnp	NRS	94.60	88.32	88.41	10
	NRD	8.91	10.09	10.81	10
GATK	NRS	93.13	90.54	90.71	10
	NRD	9.57	8.73	8.81	10
Famseq	NRS	95.73	88.91	88.74	10
	NRD	8.17	10.16	10.4	10
Polymutt	NRS	94.87	91.27	91.54	10
	NRD	5.83	6.38	6.51	10

Table 5.7: Ceph G3 5x callset metrics

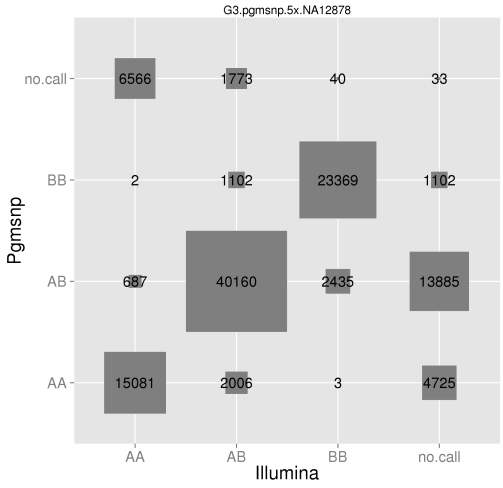


Figure 5.63: Pgmsnp NA12878 genotype concordance matrix G3 pedigree 5x coverage

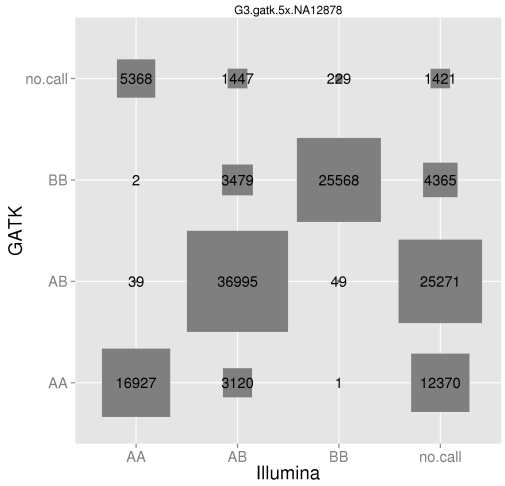


Figure 5.64: GATK NA12878 genotype concordance matrix G3 pedigree 5x coverage

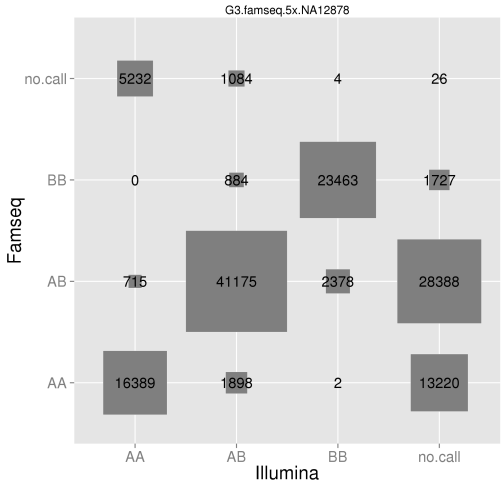


Figure 5.65: Famseq NA12878 genotype concordance matrix G3 pedigree 5x coverage

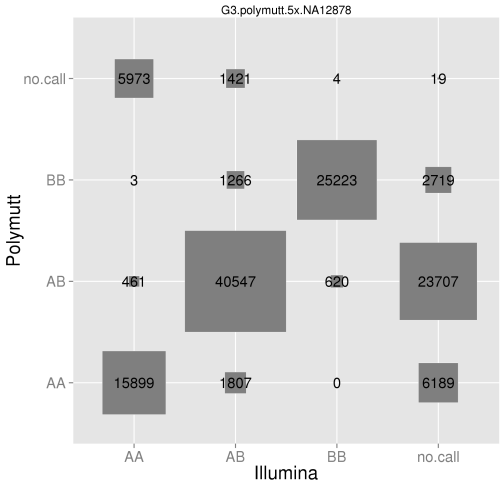


Figure 5.66: Polymutt NA12878 genotype concordance matrix G3 pedigree 5x coverage

Again, nearly 85% of the Famseq sites in this error class are shared with Pgmsnp. This suggests a significant structural difference in how prior information and marginal posteriors are computed in the Bayesian network algorithm (Pgmsnp and Famseq) versus the Elston-Stewart algorithm used by Polymutt. However, GATK does a better job than any of the 3 pedigree methods in correctly calling BB genotypes.

5.3 Conclusions

Here, I presented a novel genotyping algorithm, Pgmsnp, that models a family sequencing dataset as a Bayesian network. The work presented here gives a detailed overview of how Bayesian networks are represented, and how the belief propagation algorithm makes inferences about the marginal posterior genotype probabilities. The results of Pgmsnp was compared to three other methods. The first is Polymutt, a pedigree aware variant caller that uses the Elston-Stewart algorithm in computing the likelihood of reads in a pedigree. The second method is Famseq, which also uses the same Bayesian network framework to model pedigree sequencing data. Finally, the last method is the UnifiedGenotyper algorithm from GATK which uses the standard approach of not incorporating Mendelian inheritance amongst samples because it assumes that all samples are un-related.

Pgmsnp and its competing methods were first tested on different simulated pedigrees and sequencing datasets. At high coverage (greater than 20x coverage) the performance by all methods is equally good, and little is gained by modeling pedigree relations. At low coverage (5x), the non-reference sensitivity of Pgmsnp in non-founder, offspring individuals is higher compared to GATK. This suggests that modeling Mendelian inheritance in the priors is more informative. The genotype accuracy of Pgmsnp at low coverage is not as good when compared to Polymutt. The performance of Pgmsnp compared to Famseq is fairly similar.

In addition to simulated data, Pgmsnp was tested on an empirical dataset of Illumina sequencing reads from a subset of the Ceph 1463 pedigree. The pedigree is comprised of 5 individuals spanning three generations. Three different cuts of the pedigree were examined, all 5 individuals, and two trios from the first and second generation. The original sequencing data was generated at 50x coverage by Illumina and released as

part of their Platinum Genomes data resource. SNP calling was performed on chr20 which is 65 Mbp in total size on downsampled alignments at 5x and 10x coverage, respectively. Overall patterns from the Pgmsnp results show that it does a better job at correctly calling heterozygous sites in offspring individuals. In founder individuals, Pgmsnp has a lower sensitivity of variant detection than GATK and Polymutt. The non-reference sensitivity and non-reference discrepancy values of Pgmsnp and Famseq are very similar. This is to be expected, as they both employ a Bayesian network based genotype inference algorithm. In particular, both Pgmsnp and Famseq have an increased number of genotyping errors compared to Polymutt when incorrectly calling BB homozygous non-reference sites as AB heterozygous. Polymutt does a much better job of correctly calling these sites. GATK outperforms all three pedigree methods at these sites. Potential reasons for why Pgmsnp performs this way is that the genotyping prior places more weight on heterozygous genotypes. But Polymutt computes its priors in the same way, so there is some structural difference in how Pgmsnp and Famseq are computing posterior genotype probabilities when compared to Polymutt.

There are several ways to improve and expand the features of Pgmsnp. The program is implemented in Python, and performs at reasonable speed for moderately sized genomic intervals, but certainly can be improved. One way is to write the core functions in C++. The Cython programming language is a superset of the Python programming language and provides an interface for invoking C and C++ routines in a Python program. Pgmsnp doesn't genotype indel sites in its current implementation. The way both Polymutt and Pgmsnp handle indel genotyping is that it takes in indel data likelihoods calculated by GATK [30] or samtools [92] which are read from a VCF file, and then models Mendelian relationships of samples to emit genotypes. Current implementation of Pgmsnp requires BAM files as input,

and calculates genotype likelihoods then makes posterior genotype calls. It can be modified easily to take as input VCF or GLF (genotype likelihood files) which contain the data likelihoods of samples, and then just carry out posterior marginal computations. This would also speed up the performance of Pgmsnp. Finally, when trios are sequenced to high coverage (greater than 30x) , this can enable the detection of de-novo mutations (DNM) in the offspring. To modify the structure of the Pgmsnp Bayesian network to make inferences about DNMs would involve adding in a factor to represent the germline mutation rate. Cartwright et. al. [16] have implemented method using a graphical model to discover DNMs similar in structure to Pgmsnp.

5.4 Methods

Graphical model used

Bayesian networks are comprised of a list of factors. Figure 5.67 shows the general structure of the Bayesian network used in this study. It can be generalized to any pedigree structure. The unobserved nodes are enclosed by dashed lines representing unobserved genotypes. The observed data are enclosed by solid lines and represent sequencing reads. Figure 5.68 shows the particular factors used in the study. The three core factors are the genotype prior of the non-founder individual(s), the genotype prior of founder individuals, and the data likelihood factor of the sequencing reads. The genotype prior factor represents the conditional probability of the child genotype given its two parents. Essentially, this is a Punnett square. The genotype prior of the founders represents the conditional probability of a founder genotype given θ, which is the population scaled mutation rate [52]. For this study θ's value is set to .001. The data likelihood factor represents the likelihood function the probability of the basecall given the genotype of the individual. Likelihood functions are not proper

probability distribution functions, and their values do not necessarily sum to one. More details on the genotype data likelihood factor is given in the next section.

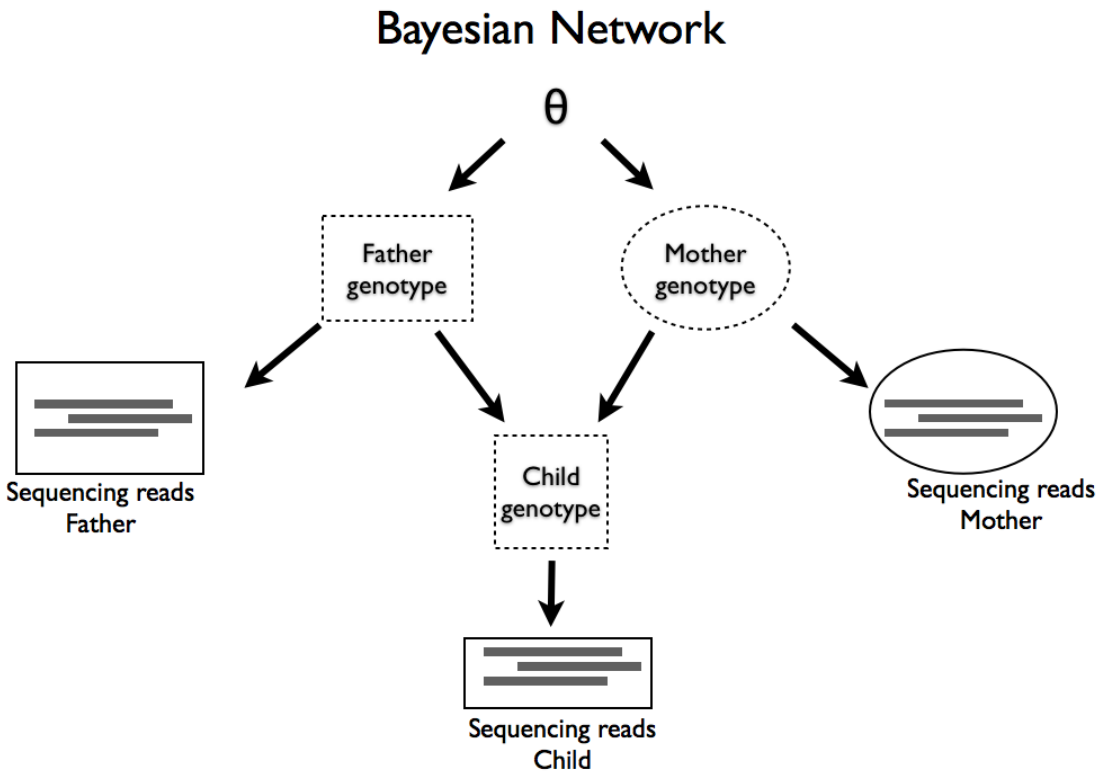


Figure 5.67: **Pgmsnp Network** - The Bayesian network used in this study

Genotype Likelihood Factor

Figure 5.69 shows a graphical representation of the genotype likelihood table. The likelihood function described is taken from [87]. At a given position in the genome let there be N aligned bases consisting of A's, C's, G's, and T's: $N = N_A + N_C + N_G + N_T$ Each aligned base also has an associated Phred quality score. A Phred quality score, Q, is logarithmically related to the base calling error probability, P:

$$Q = -10log_{10}P$$

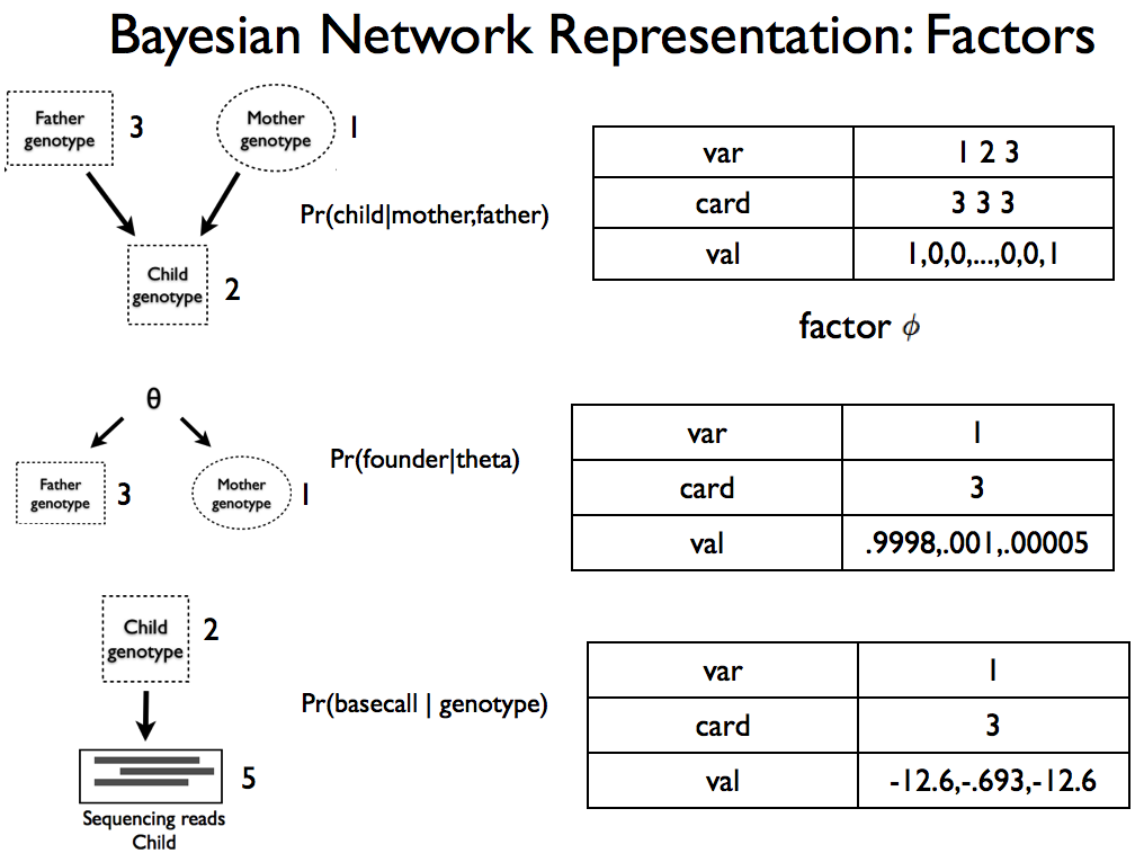


Figure 5.68: **Individual factors of network** - There are three core factors used in the Bayesian network, shown in the figure starting from the top.

$$P = 10^{\frac{-Q}{10}}$$

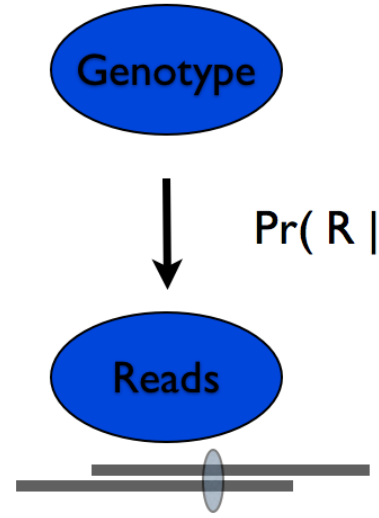
If we let R be all basecalls for a particular position across all aligned reads for an individual, then we can calculate the data likelihood of reads, given a particular genotype:

$$Pr(R|G_i), i = 1 \ldots 10$$

For example, if the assumed genotype was AA, the likelihood function would be:

$$Pr(R|AA) = \prod_{j=1}^{N_A}(1 - e_j) \prod_{k=1}^{N-N_A} \frac{e_k}{3} \tag{5.7}$$

If the assumed genotype was heterozygous AC, the likelihood function is:



Genotype	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT
LL	-12.6	-0.693	-12.6	-12.6	-12.6	-1.00E-05	-0.693	-0.693	-12.6	-12.6

Figure 5.69: **Genotype Likelihood Factor** - A graphical representation of the genotype likelihood

$$Pr(R|AC) = \prod_{j=1}^{N_A+N_C} 0.5(1-\frac{2e_j}{3}) \prod_{k=1}^{N-N_A-N_C} \frac{e_k}{3} \tag{5.8}$$

Note, that equation 5.8 was obtained in the following way. If a basecall was A, with associated error probability e, then

$$.5(P(A|A) + P(A|C)) = .5((1-e) + e/3) = 0.5(1-\frac{2e_j}{3})$$

since we have equal chance of sampling either chromosome (assuming diploidy). The likelihood function(s) for the remaining 8 genotypes would be similar to equations 5.7 and 5.8. For each individual with aligned reads R, there would be 10 genotype data likelihood values.

Pgmsnp algorithm overview

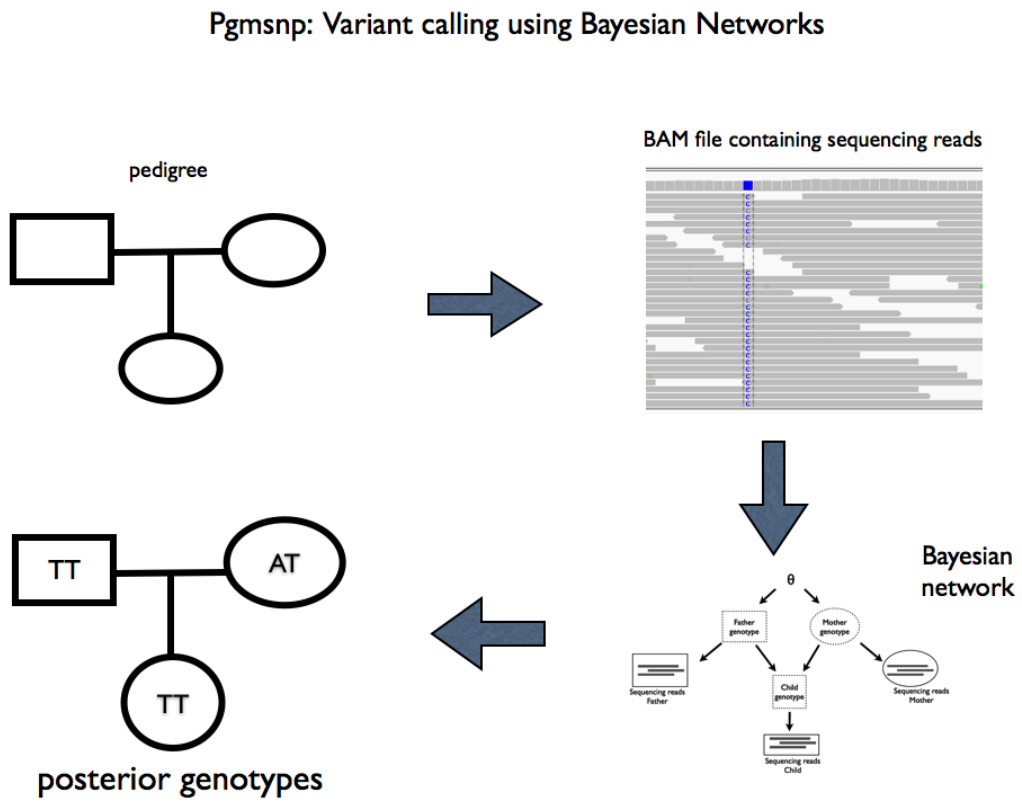


Figure 5.70: **Pgmsnp overview** - Pgmsnp takes in as input the pedigree information of samples and the BAM file(s) containing sequencing reads. It constructs a Bayesian network at every position in the genome, performs inference, and posterior genotypes are reported.

Figure 5.70 gives a high level overview of how the Pgmsnp works. Inputs are pedigree information and a merged BAM file containing sequencing reads of the samples. At each position of the genome a Bayesian network is constructed and max-product belief propagation inference is performed. The output of the program are genotypes of samples reported in a Variant Call Format (VCF) file.

Computing QUAL values of sites

The Variant Call Format (VCF) specification [27] defines the QUAL column to the Phred scaled probability that there is no variant. Higher QUAL values indicate higher confidence that the site is segregating. To compute this value, Pgmsnp needs to return the probability that all samples are homozygous reference. The clique tree data structure is used to compute marginal posteriors of variables representing genotypes in the Bayesian network. We can use this data structure as well to compute the joint distribution. Recall, that a calibrated clique tree holds the results of probabilities of all cliques in the tree, but it is also an alternative representation of the joint distribution, which is denoted as P_Φ. If we denote $\mathbf{X}$ to be the set of random variables in a Bayesian network, then a calibrated clique tree provides an alternative measurement of the joint distribution by the following formula:

$$P_\Phi = \frac{\prod_{i \in V_T} \beta_i(C_i)}{\prod_{(i-j) \in \epsilon_T} \mu_{i,j}(S_{i,j})} \tag{5.9}$$

The numerator represents the product of the final beliefs of each clique node in the tree and the denominator represents the sepset beliefs of the edges between nodes. The proof as to how equation 5.9 is an alternate representation of the joint distribution is described in [76]. Once the joint distribution is computed in Pgmsnp by implementing the formula, the value of the instantiation of all variables having homozygous reference genotypes is retrieved and the QUAL value is computed.

Data Simulation

As proof of concept, Pgmsnp was tested on simulated data free of sequencing and mapping error. Figure 5.71 show the steps taken to generate simulated data for initial testing of method. Haplotypes were simulated with the program cosi [126]

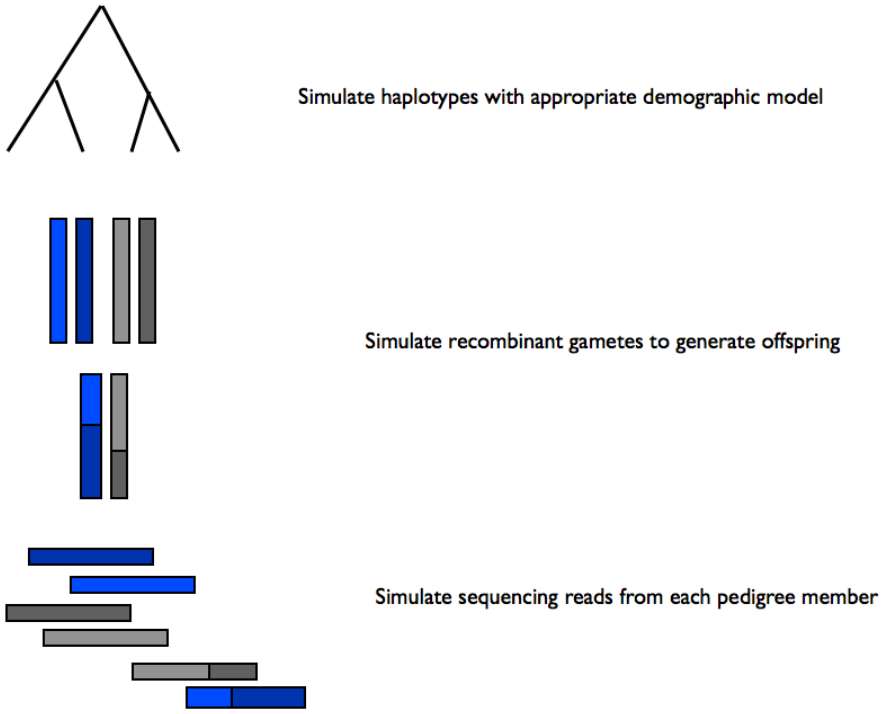


Figure 5.71: **Data simulation** - Simulation pipeline used to generate simulated data

which is a coalescent simulator that generates haplotypes. The program was used to generate 50 founder 1 Mbp haplotypes. Parameters given to the program were as follows: mutation rate of 1.5e-8 per base per generation, effective population size of 10000, recombination events based on the deCode genetic map for autosomes. The demographic model used was an Out-of-Africa model with a European bottleneck. A total of 3159 segregating (polymorphic) sites was observed from a single run of the program. All of the sites are in Hardy-Weinberg equilibrium which was checked with the genetic analysis program PLINK [118]. Non-founder haplotypes were formed by simulating a Poisson number of recombination events to generate gametes in each parent. Each gamete had to at least have at least one crossover event.

Illumina sequencing reads of 101 basepairs were simulated with the program mason [59] without the introduction of any sequencing errors. Each individual had an average 20x coverage of its 1 Mbp genome, based on the equation $C = \frac{R \times N}{G}$ where R is the read length, N is the number of reads, G is the size of the genome, and C is the coverage.

The founders in each of the five pedigrees shown in Figure 5.4 are the same and are referred to as motherOne and fatherOne. The child in the trio pedigree is referred to as childOne and its sibling is referred to as child3. The marryin (marryinOne) in the multi generation pedigree married childOne to produce the grandchild referred to as grandchildOne.

Ceph Pedigree

Illumina has provided the genomics community with a set of high coverage 50x genomes deemed the Illumina Platinum dataset [62]. This dataset comprises of the Ceph 1463 pedigree which is made of 17 people of European descent in Utah. The aligned BAM files of 5 of the 17 individuals were examined in this study. They are NA12891, NA12892, NA12878, NA12877, and NA12882. The aligned reads of chr20 were downsampled to 5x and 10x coverage and then examined by Pgmnsp and associated methods.

Genotype concordance metrics

The two concordance metrics used to evaluate the performance of all methods tested in the study were non-reference sensitivity and non-reference discrepancy (NRS and NRD). The genotype concordance matrix is used to calculate these values, and they can be thought of as summary statistics of the raw genotype concordance metrics.

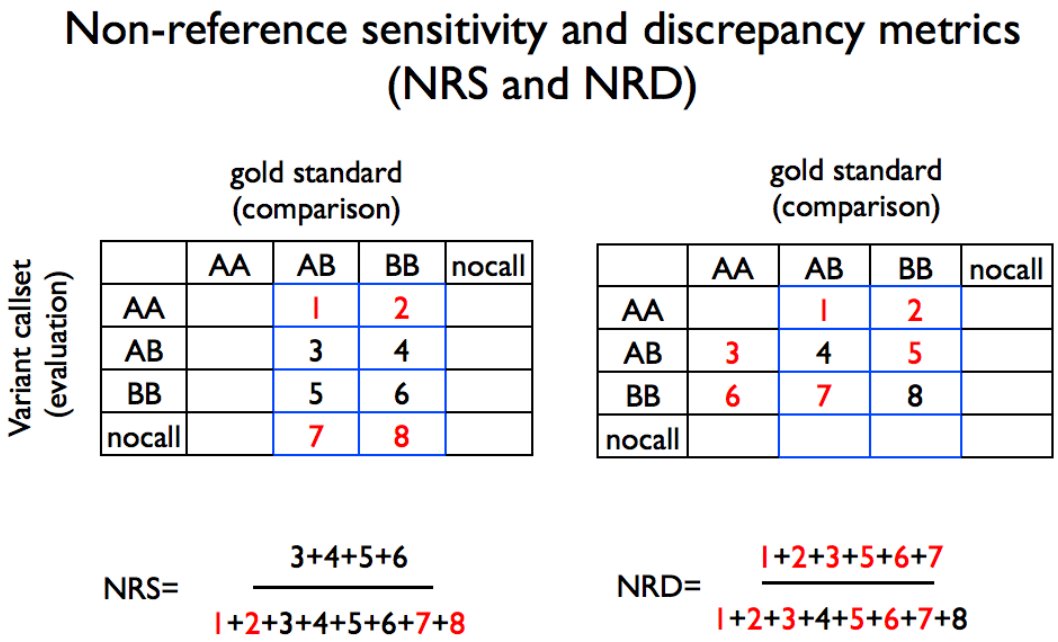


Figure 5.72: **Genotype concordance metrics calculation** - The genotype matrix of the gold and evaluation call sets is used to compute NRS and NRD values.

NRS and NRD concordance metrics as a function of QUAL

NRS and NRD metrics can be plotted at various QUAL values. To investigate the relationship between QUAL values and NRS and NRD values, an evaluation dataset's variant records, prior to be compared to the gold standard dataset, can be binned according to QUAL. For each of the four methods examined in the study the empirical cumulative distribution function (ecdf) of QUAL values was calculated. For the methods Pgmsnp, GATK, and Famseq the distribution ranged from 10 to 1200; for Polymutt the QUAL values ranged from 1 to 100. Based on this, records from each method were binned into 100 bins evenly spaced between the numbers 10-1200 and 1-100. Once binned, the evaluation callset is merged with the gold standard comparison callset, and concordance metrics are calculated. The results can be visualized easily

and a QUAL cutoff for a desired non-reference sensitivity and genotyping accuracy can be empirically determined based on the results. Moreover, variant call sets derived from different methods on the same input data can be directly compared to empirically derive what the maximum NRS is obtained by a method.

Chapter 6

Conclusions

6.1 Summary of work

Next generation sequencing and genomic enrichment technologies has had a profound impact on the field of human genetics by enabling researchers to fully characterize the complete frequency spectrum of genetic variation segregating in the human population. My thesis focused on data analysis capture sequencing datasets and the development of a novel genotyping algorithm to discover SNPs in family sequencing datasets. In Chapter 2 I describe my contribution to the analysis of capture sequencing data from the exon sequencing pilot of the 1000 Genomes Project [100, 33]. This was one of the first studies to characterize rare coding variation using genomic enrichment technologies. The allele frequency spectrum of the data clearly showed an excess of singleton class variants. This is due to rapid, explosive population growth in the human population over the past 10,000 years [70] and similar patterns of an excess of singletons have been observed in other large exome studies like the Exon Sequencing Project [134]. Chapter 3 continues the focus on capture sequencing, but in a different context. The work presented investigates the applicability of using whole genome amplified (WGA) DNA in capture sequencing. Although its a small scale

study, it demonstrates that WGA DNA can be used effectively in such experiments. The work has been recently published in BMC Genomics [63] and has been noted by the journal as a highly accessed paper. Chapter 4 describes the bioinformatics steps to identify a potential causative mutation in a Mendelian form of hearing loss. Through a discrete filtering approach with a minor allele frequency cutoff, a non-synonymous variant in the *TMTC2* gene was identified as a putative causative mutation. The genotype segregates perfectly with the phenotype in the family (both in the sequenced individuals of the pedigree, as well members not chosen for sequencing). In addition, the variant is enriched in a cohort of unrelated individuals which share the same phenotype. Finally in Chapter 5 I describe a novel pedigree aware genotyping algorithm to discover SNPs in family sequencing datasets. The method uses a Bayesian network inference algorithm, called belief propagation, to compute posterior marginal genotype probabilities. Pgmsnp, has comparable detection sensitivities to other pedigree callers, but has a slightly higher genotype discordance rate.

I gained expertise in a variety of subjects including population genetics, human genetics, probability theory, and machine learning during the course of my Ph.D. The underlying theme that threads each chapter together is using data from next generation sequencing experiments to discover rare genetic variation. The field of human genomics is rapidly evolving and below I address future directions of various aspects of the field that I find of interest.

6.2 Future directions

Large scale sequencing

Without next generation sequencing and genomic enrichment technology large scale re-sequencing projects like the 1000 Genomes Project would not have been possible.

Re-sequencing large samples of individuals is the only way to characterize the full spectrum of human genetic variation. Besides being a data resource for the human genetics community, the more important contribution of the 1000 Genomes Project was the development of algorithms and computer software to analyze data from next generation sequencing experiments. Methods of alignment and variant calling are continually improving, and to a certain extent have reached a maturation level. This has made re-sequencing of large datasets more common, whereas 10 years ago such experiments were the realm of large scale genome sequencing centers. Besides driving algorithm development, the 1000 Genomes Project has ushered in the era of rare variant GWAS. Many of the variants discovered in the 1000 Genomes Project have been added to commercial genotyping arrays and it is now possible to design a GWAS with both rare and common variants being genotyped [25]. Whether or not the inclusion of rare variants in GWAS designs will result in new associations explaining unaccounted for heritability remains to be seen, and the end results will probably have to drawn by a disease by disease basis. Depending on the effect size of variants and the underlying biological complexity of the phenotype, GWAS sample sizes will have to increase into the several hundred thousand. Building on the tool development driven by the 1000 Genomes Project, future analysis of large scale sequencing experiments will focus more on interpretation [46], as well as sharing of results from different projects and researchers. Building an effective bioinformatics infrastructure for data sharing and visualization will take on increased importance. To that end, there are several active projects in the Marth laboratory to develop effective tools to visualize genetic variation datasets.

Defining causality for Mendelian phenotypes

Over the past 3 years there have been several successful examples of exome sequencing uncovering causative mutations and more than 180 novel genes have been discovered

[14]. The work presented here that putatively identified a casual candidate mutation builds on the successful methodologies of using a filtering approach to narrowing down candidate mutations [131]. While ad hoc filtering approaches are the predominant approach to analyzing Mendelian exome datasets, it lacks the statistical rigor and conventions of linkage mapping (LOD scores) or GWAS (p-values). Both linkage and association indirectly identify causal mutations, while sequencing directly attempts to uncover them [46]. There have been no studies to my knowledge that have attempted to assign p-values to causal variants uncovered in exome sequencing and this remains an open area of research.

Methods to better articulate causality in a statistical way (p-values, LOD scores, etc) may not be the way most biologists think about the issue. Rather, causality is defined by physical interactions of proteins, DNA, and other cellular structures [125]. Once a variant is uncovered via sequencing, understanding the functional consequences of the mutation across different tissue types is quite difficult. High-throughput cellular assays are being developed [110, 5] to test the functional consequences of mutations uncovered in sequencing. Testing the effects of mutations that are potentially incompletely penetrant, like the *TMTC2* variant in the hearing loss study has the added challenge of controlling for genetic background and gene-gene interactions.

Pedigree aware haplotyping and genotyping

The Pgmsnp method described in Chapter 5 treats each site independently. A natural extension of this method would be to incorporate linkage to extend Pgmsnp into being a pedigree aware haplotype caller. The steps involved are closely related to haplotype phasing, which is the process of inferring haplotype phase from genotype data. Inferring haplotype phase from next generation sequencing data is an active area of research [149]. Extending the graphical model of Pgmsnp to make it haplotype

aware would involved implementing a dynamic Bayesian network (DBN). DBN's are a more general form of a Hidden Markov Model (HMM) [76]. The inference algorithms for DBNs and HMMs are closely related. The main difference between a Bayesian network and DBN is that a DBN relates variables that are adjacent over position or time. A recent publication by Zhang has implemented a DBN to both infer haplotypes and genotypes from NGS data [149].

Genotype imputation of large pedigrees

Pgmsnp uses an exact method to compute posterior probabilities. For larger pedigrees for multiple generations and more individuals, the computational burden becomes intractable, due to the size and cardinality of intermediate factors generated during belief propagation [76]. As family based designs become more popular, it is quite feasible that large pedigrees of individuals will be sequenced. One way to workaround the computational roadblock of genotype calling in large pedigrees is to perform imputation. Imputation refers to inferring missing data by borrowing information from full observations on related subjects [19]. Genotype imputation is an active area of research where genotypes are inferred in a set of un-related individuals using a reference panel of densely genotyped samples, leveraging linkage disequilibrium patterns in the data [97]. Recently, a publication by Ellen Wijsman and colleagues [19] described an imputation algorithm for large pedigrees called GIGI that imputes genotypes derived from individuals sharing genomic intervals that are identical by descent. Hence for a genetic study of a large pedigree, one practical option would be to genotype a subset of individuals with Pgmsnp and impute genotypes of unsequenced individuals in the pedigree with GIGI. The design of large family studies utilizing genomic sequencing is a balance between the statistical issues of family data with the cost of and infrastructure of sequencing large numbers of genomes.

Sequencing is only the first step

As algorithms for the analysis of high throughput sequencing datasets mature and sequencing technologies become more accessible to the larger human genetics community, many of the open questions and challenges that remain to fully take advantage of these datasets do not necessarily have to do with bioinformatics. Prior to NGS sequencing, variant discovery was the rate limiting step in comprehensively describing the genetic variation present in a sample of individuals [128]. NGS has removed this step, so finding enough properly consented samples is a critical issue [75]. In particular, as clinical applications of sequencing increase, proper phenotyping of patients, is of increased importance. To this end, there have been new online tools to help record and share precise phenotypic data on subjects, such as PhenoDB [50] and PhenoTips [44]. Another challenge is the biological interpretation of findings from sequencing experiments. The human genetics community is transitioning from figuring out what's the best way to identify variants from NGS data to trying to identify which variants discovered in NGS experiments are implicated in novel biology [5]. The marriage of high throughput sequencing with novel high throughput functional genomic assays will be the critical step needed to assess the function of variants uncovered through sequencing.

Appendix A

A.1 Additional tables for Chapter 3

This portion of the Appendix contains results of the statistical anlaysis of allele bias in the WGA capture sequencing dataset.

A.1.1 Computed p-values of allele bias results whole-exome capture

```
AC SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 0.77 1 -
Concordant 1 1 0.56


AG SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 1 1 -
Concordant 0.95 1 1

AT SNPs WGA.uniq Genomic.uniq NRD
```

Genomic.uniq 1 --

NRD 0.99 1 -

Concordant 1 1 0.75

CG SNPs WGA.uniq Genomic.uniq NRD

Genomic.uniq 1 --

NRD 1 1 -

Concordant 0.01 1.70E-04 2.40E-04

CT SNPs WGA.uniq Genomic.uniq NRD

Genomic.uniq 1 --

NRD 1 1 -

Concordant 0.13 0.33 1

GT SNPs WGA.uniq Genomic.uniq NRD

Genomic.uniq 0.756 --

NRD 0.759 1 -

Concordant 0.033 1 1

A.1.2 Computed p-values of allele bias results chr12 capture

AC SNPs WGA.uniq Genomic.uniq NRD

Genomic.uniq 1 --

NRD 1 1 -

Concordant 1 1 0.074

AG SNPs WGA.uniq Genomic.uniq NRD

```
Genomic.uniq 1 --
NRD 1 1 -
Concordant 1 1 1

AT SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 1 1 -
Concordant 1 1 1

CG SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 1 1 -
Concordant 1 1.00E+00 1.00E+00

CT SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 1 1 -
Concordant 1 0.42 1

GT SNPs WGA.uniq Genomic.uniq NRD
Genomic.uniq 1 --
NRD 1 1 -
Concordant 1 1 1
```

A.1.3 Computed p-values of allele bias results Affymetrix whole-exome capture

```
Genomic calls comparisons to Affy genotypes comparisons
```

AC SNPs Concordant NRS

NRS 0.546 -

NRD 1 1

AG SNPs Concordant NRS

NRS 1 -

NRD 1 1

AT SNPs Concordant NRS

NRS 0.006 -

NRD 1 1

CG SNPs Concordant NRS

NRS 1 NA

NRD 1 1

CT SNPs Concordant NRS

NRS 0.22 -

NRD 1 0.69

GT SNPs Concordant NRS

NRS 1 -

NRD 0.88 0.82

WGA calls comparisons to Affy genotypes

AC SNPs Concordant NRS

NRS 0.124 -

NRD 1 0.53

AG SNPs Concordant NRS

NRS 0.4 -

NRD 1 1

AT SNPs Concordant NRS

NRS 0.03 -

NRD 1 0.84

CG SNPs Concordant NRS

NRS 1 NA

NRD 1 1

CT SNPs Concordant NRS

NRS 0.22 -

NRD 1 0.69

GT SNPs Concordant NRS

NRS 1 -

NRD 0.5 0.8

A.1.4 Computed p-values of allele bias results Affymetrix chr12 capture

Genomic calls comparisons to Affy genotypes comparisons

AC SNPs Concordant NRS

NRS 0.546 -

NRD 1 1

AG SNPs Concordant NRS

NRS 1 -

NRD 1 1

AT SNPs Concordant NRS

NRS 0.006 -

NRD 1 1

CG SNPs Concordant NRS

NRS 1 NA

NRD 1 1

CT SNPs Concordant NRS

NRS 0.22 -

NRD 1 0.69

GT SNPs Concordant NRS

NRS 1 -

NRD 0.88 0.82

WGA calls comparisons to Affy genotypes

AC SNPs Concordant NRS

NRS 0.124 -

NRD 1 0.53

AG SNPs Concordant NRS

NRS 0.4 -

NRD 1 1

AT SNPs Concordant NRS

NRS 0.03 -

NRD 1 0.84

CG SNPs Concordant NRS

NRS 1 NA

NRD 1 1

CT SNPs Concordant NRS

NRS 0.22 -

NRD 1 0.69

GT SNPs Concordant NRS

NRS 1 -

NRD 0.5 0.8

A.2 Additional figures for Chapter 5

This portion of the Appendix contains genotype concordance results from Pgmsnp and related methods for pedigree aware SNP calling.

A.2.1 Maximum NRS values and associated NRD values at 10x coverage

Listed below are additional tables showing the maximum NRS values and associated NRD values for each of the Ceph pedigrees studied.

	Metric	NA12882	NA12877	NA12878	NA12891	NA12892	QUAL
Pgmsnp	NRS	99.25	97.45	99.33	97.7	97.56	10
	NRD	1.52	2.51	1.48	2.46	2.51	10
GATK	NRS	99.27	98.84	99.24	98.88	98.82	10
	NRD	1.40	1.36	1.52	1.43	1.51	10
Famseq	NRS	99.48	97.83	99.56	97.99	97.83	10
	NRD	1.06	2.16	1.12	2.13	2.24	10
Polymutt	NRS	99.43	98.59	99.48	98.75	98.55	10
	NRD	0.87	0.98	0.78	1.05	1.16	10

Table A.1: Ceph A5 10x callset metrics

	Metric	NA12882	NA12877	NA12878	QUAL
Pgmsnp	NRS	99.31	97.61	97.62	10
	NRD	1.49	1.97	2.07	10
GATK	NRS	99.28	98.68	98.72	10
	NRD	1.24	1.16	1.20	10
Famseq	NRS	99.50	97.98	97.95	10
	NRD	1.08	1.70	1.83	10
Polymutt	NRS	99.45	98.69	98.74	10
	NRD	0.8	0.83	0.89	10

Table A.2: Ceph A3 10x callset metrics

A.2.2 NRS and NRD values as a function of QUAL

Listed below are additional figures for GATK, Famseq, and Polymutt showing the concordance metrics non-reference sensitivity (NRS) and non-reference discrepancy (NRD) as a function of quality (QUAL) score.

	Metric	NA12878	NA12891	NA12892	QUAL
Pgmsnp	NRS NRD	99.16 1.51	97.65 2.03	97.51 2.16	10 10
GATK	NRS NRD	99.25 1.39	98.66 1.24	98.61 1.33	10 10
Famseq	NRS NRD	99.45 1.14	98.00 1.75	97.85 1.90	10 10
Polymutt	NRS NRD	99.30 .86	98.26 .94	98.18 1.0	10 10

Table A.3: Ceph G3 10x callset metrics

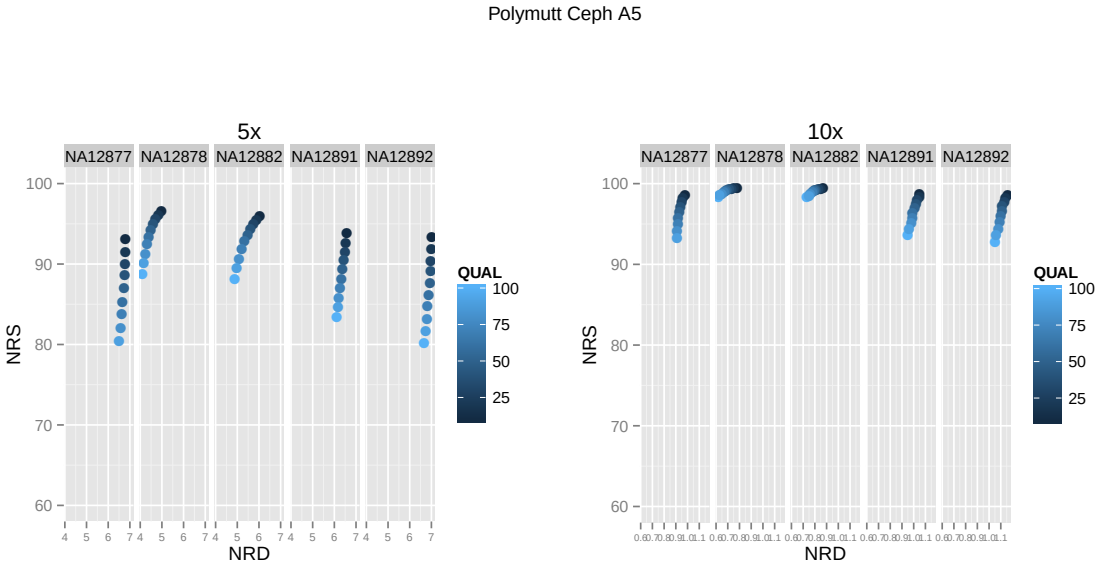


Figure A.1: **Polymutt metrics Ceph A5** - Polymutt NRD and NRS metrics as a function of QUAL from Ceph-A5 pedigree

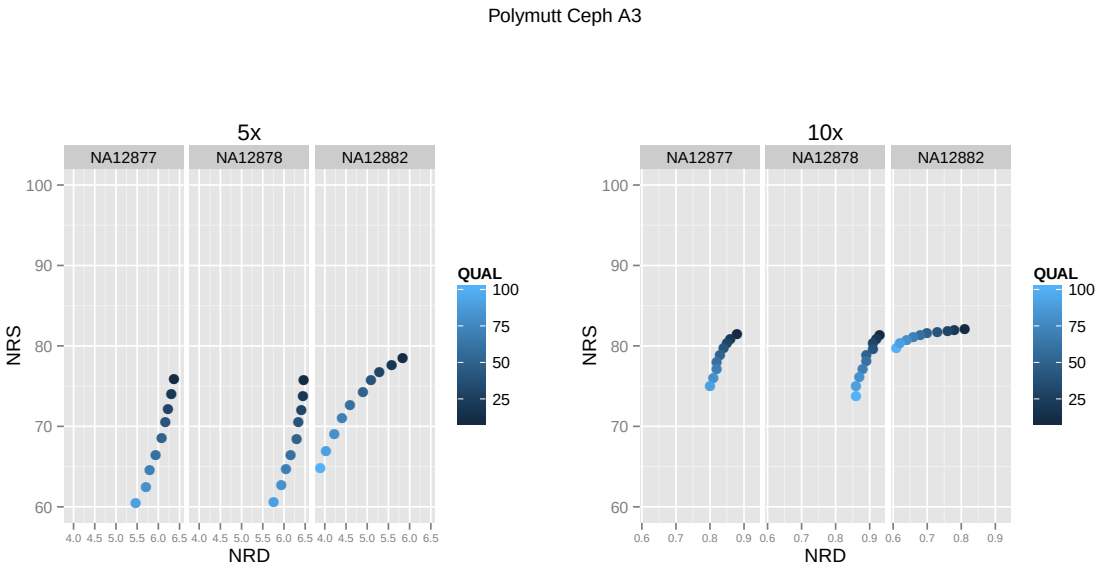


Figure A.2: **Polymutt metrics Ceph A3** - Polymutt NRD and NRS metrics as a function of QUAL from Ceph-A3 pedigree

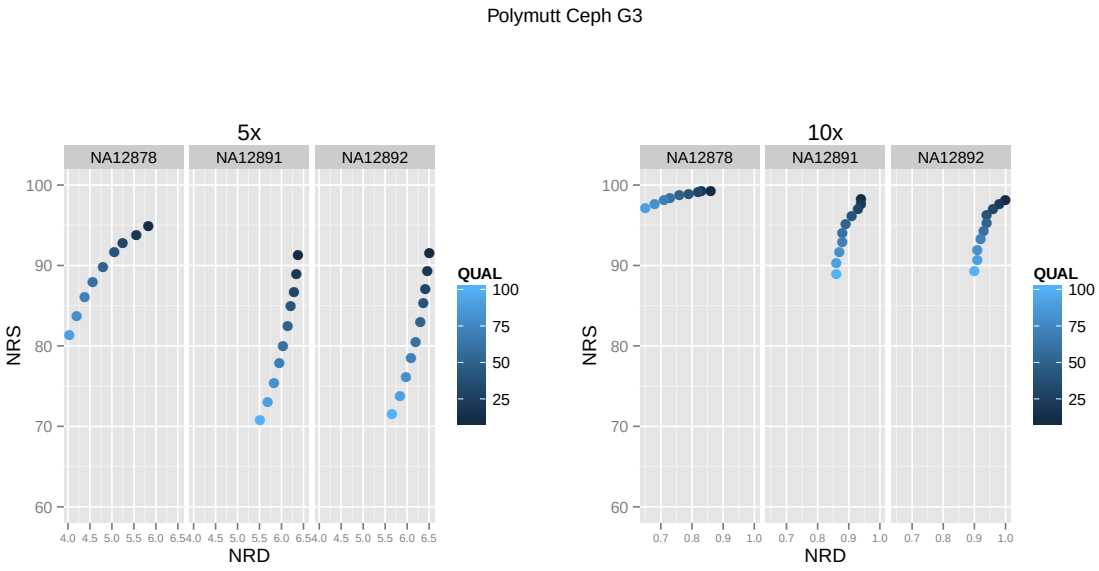


Figure A.3: **Polymutt metrics Ceph G3** - Polymutt NRD and NRS metrics as a function of QUAL from Ceph-G3 pedigree

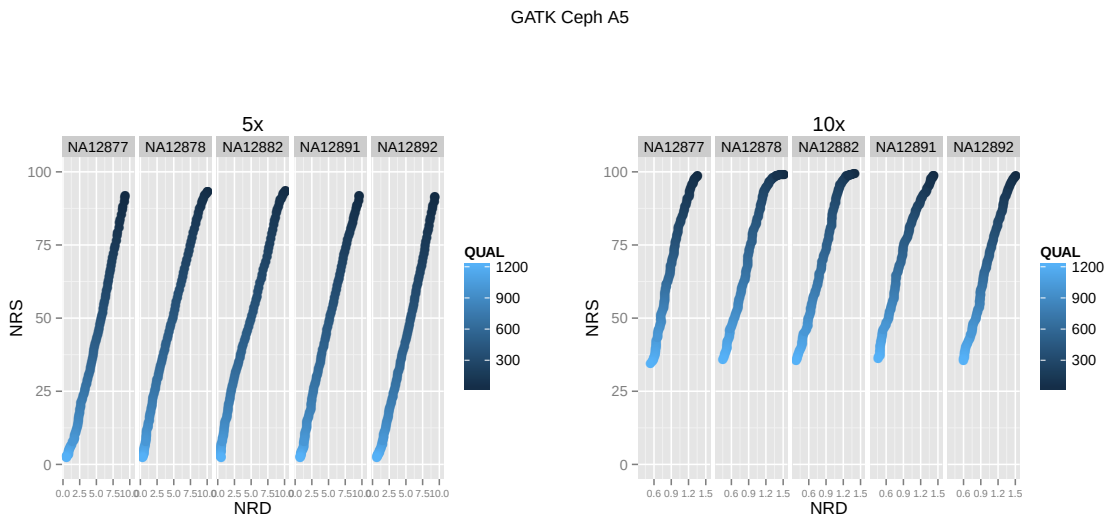


Figure A.4: **GATK metrics Ceph A5** - GATK NRD and NRS metrics as a function of QUAL from Ceph-A5 pedigree

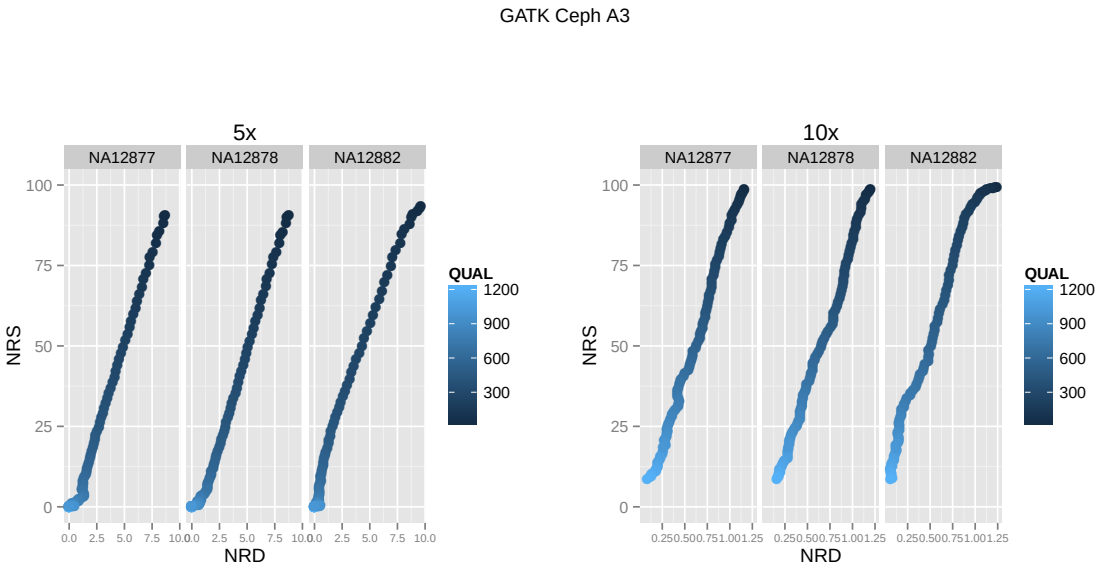


Figure A.5: **GATK metrics Ceph A3** - GATK NRD and NRS metrics as a function of QUAL from Ceph-A3 pedigree

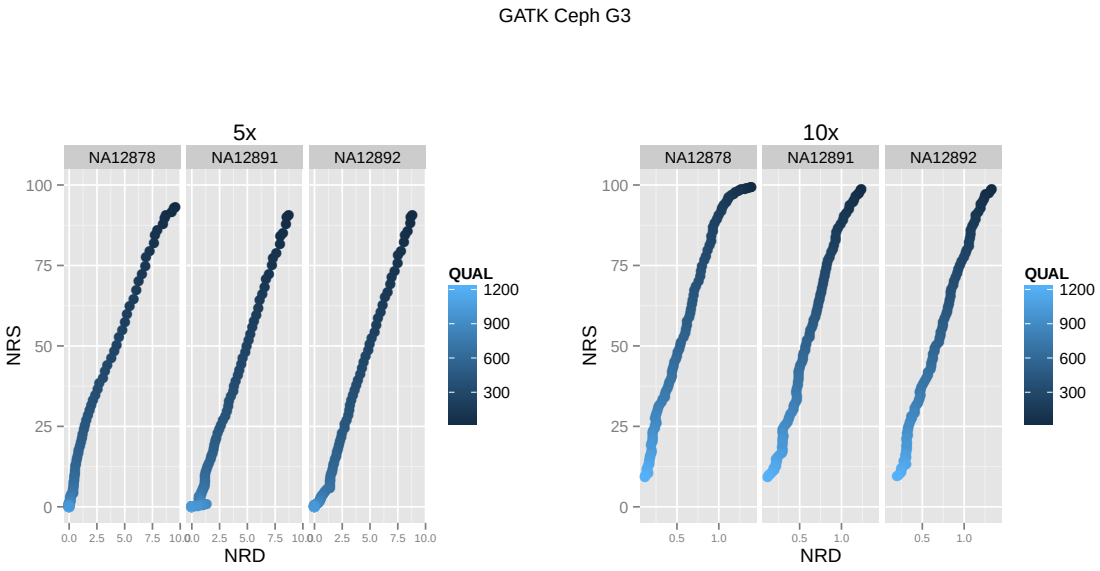


Figure A.6: **GATK metrics Ceph G3** - GATK NRD and NRS metrics as a function of QUAL from Ceph-G3 pedigree

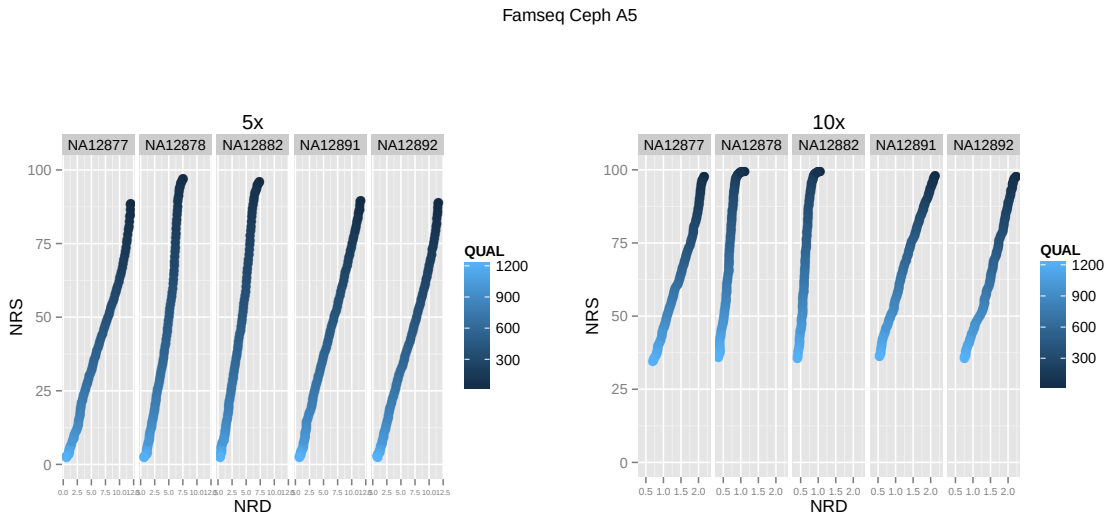


Figure A.7: **Famseq metrics Ceph A5** - Famseq NRD and NRS metrics as a function of QUAL from Ceph-A5 pedigree

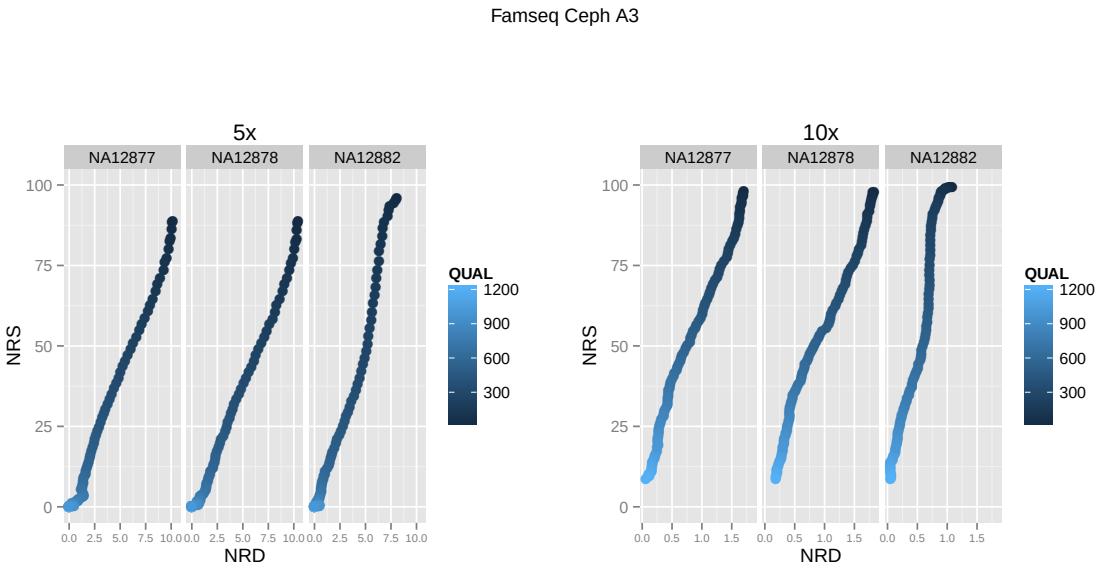


Figure A.8: **Famseq metrics Ceph A3** - Famseq NRD and NRS metrics as a function of QUAL from Ceph-A3 pedigree

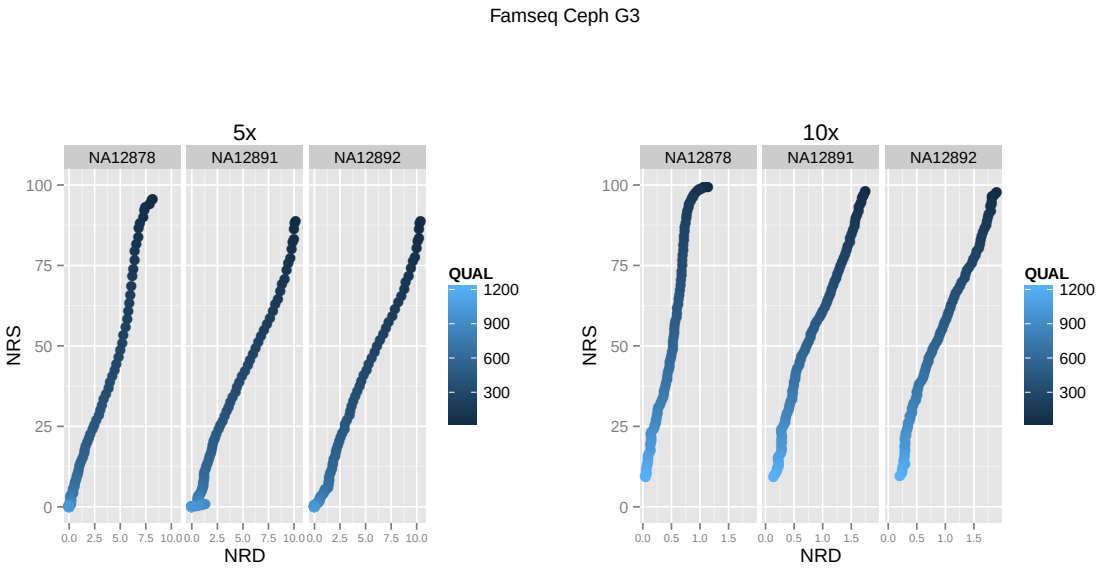


Figure A.9: **Famseq metrics Ceph G3** - Famseq NRD and NRS metrics as a function of QUAL from Ceph-G3 pedigree

Bibliography

[1] G. R. Abecasis, B. M. Yashar, Y. Zhao, N. M. Ghiasvand, S. Zareparsi, K. E. Branham, A. C. Reddick, E. H. Trager, S. Yoshida, J. Bahling, E. Filippova, S. Elner, M. W. Johnson, A. K. Vine, P. A. Sieving, S. G. Jacobson, J. E. Richards, and A. Swaroop. Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease. *Am. J. Hum. Genet.*, 74(3):482–494, Mar 2004.

[2] M. Abney. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, 25(12):1561–1563, Jun 2009.

[3] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations., 2010.

[4] Stylianos E Antonarakis, Aravinda Chakravarti, Jonathan C Cohen, and John Hardy. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nature reviews. Genetics*, 11(5):380–4, May 2010.

[5] M. Baker. Functional genomics: The changes that count. *Nature*, 482(7384):259–262, Feb 2012.

[6] D. J. Balding. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7(10):781–791, Oct 2006.

[7] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[8] David R Bentley and et. al. Balasubramanian. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[9] Y. Berthier-Schaad, W. H. Kao, J. Coresh, L. Zhang, R. G. Ingersoll, R. Stephens, and M. W. Smith. Reliability of high-throughput genotyping of whole genome amplified DNA in SNP genotyping studies. *Electrophoresis*, 28(16):2812–2817, Aug 2007.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] G. L. Blatch and M. Lassle. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays*, 21(11):932–939, Nov 1999.

[12] W. F. Bodmer. Human genetics: the molecular challenge. *Cold Spring Harb. Symp. Quant. Biol.*, 51 Pt 1:1–13, 1986.

[13] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331, 1980.

[14] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. Mackenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.*, Sep 2013.

[15] Elof Axel Carlson. *Mendel's Legacy: The Origin of Classical Genetics*. Cold Spring Harbor Laboratory Press, 2004.

[16] R. A. Cartwright, J. Hussin, J. E. Keebler, E. A. Stone, and P. Awadalla. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. *Stat Appl Genet Mol Biol*, 11(2), 2012.

[17] A. Chakravarti and A. Kapoor. Genetics. Mendelian puzzles. *Science*, 335(6071):930–931, Feb 2012.

[18] C. H. Chandler, S. Chari, and I. Dworkin. Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends Genet.*, 29(6):358–366, Jun 2013.

[19] C. Y. Cheung, E. A. Thompson, and E. M. Wijsman. GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am. J. Hum. Genet.*, 92(4):504–516, Apr 2013.

[20] E. T. Cirulli and D. B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.*, 11(6):415–425, Jun 2010.

[21] Michael J Clark, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, 29(10):908–914, 2011.

[22] 1000 Genomes Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

[23] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, Oct 2005.

[24] Gregory M Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640, 2011.

[25] A. Cortes and M. A. Brown. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.*, 13(1):101, 2011.

[26] A. Coventry, L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, D. A. Wheeler, A. Sabo, C. Lusk, K. G. Weiss, H. Akbar, A. Cree, A. C. Hawes, I. Newsham, R. T. Varghese, D. Villasana, S. Gross, V. Joshi, J. Santibanez, M. Morgan, K. Chang, W. H. Iv, A. R. Templeton, E. Boerwinkle, R. Gibbs, and C. F. Sing. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, 1:131, 2010.

[27] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and R. Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug 2011.

[28] P. I. de Bakker. Selection and evaluation of Tag-SNPs using Tagger and HapMap. *Cold Spring Harb Protoc*, 2009(6):pdb.ip67, Jun 2009.

[29] Y. J. de Kok, G. F. Merkx, S. M. van der Maarel, I. Huber, S. Malcolm, H. H. Ropers, and F. P. Cremers. A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene. *Hum. Mol. Genet.*, 4(11):2145–2150, Nov 1995.

[30] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo Del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, March 2011.

[31] J C Dohm, C Lottaz, T Borodina, and H Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36:e105, 2008.

[32] H. Donis-Keller, P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, T. P. Keith, D. W. Bowden, D. R. Smith, and E. S. Lander. A genetic linkage map of the human genome. *Cell*, 51(2):319–337, Oct 1987.

[33] Richard M Durbin, Gonçalo R Abecasis, David L Altshuler, Adam Auton, Lisa D Brooks, Richard A Gibbs, Matt E Hurles, Gil A McVean, and The

1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[34] A. Elsharawy, J. Warner, J. Olson, M. Forster, M. B. Schilhabel, D. R. Link, S. Rose-John, S. Schreiber, P. Rosenstiel, J. Brayer, and A. Franke. Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing. *BMC Genomics*, 13:500, 2012.

[35] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, 8(3):175–185, Mar 1998.

[36] MediaWiki Foundation.

[37] Thomas B. Friedman and Andrew J. Griffith. Human nonsyndromic sensorineural deafness. *Annu Rev Genomics Hum Genet*, 4:341–402, September 2003.

[38] W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, J. M. Akey, and the NHBLI Exome Sequencing Project. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, Jan 2013.

[39] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, Jun 2002.

[40] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. Technical report, Boston College, http://arxiv.org/abs/1207.3907, 2012.

[41] G. Gibson. Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, 13(2):135–145, Feb 2011.

[42] C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman. Unlocking Mendelian disease using exome sequencing. *Genome Biol.*, 12(9):228, 2011.

[43] John H. Gillespie. *Population Genetics: A Precise Guide*. Johns Hopkins University Press, 2004.

[44] M. Girdea, S. Dumitriu, M. Fiume, S. Bowdin, K. M. Boycott, S. Chenier, D. Chitayat, H. Faghfoury, M. S. Meyn, P. N. Ray, J. So, D. J. Stavropoulos, and M. Brudno. PhenoTips: patient phenotyping software for clinical and research use. *Hum. Mutat.*, 34(8):1057–1065, Aug 2013.

[45] Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, Stacey Gabriel, David B Jaffe, Eric S Lander, and Chad Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–189, 2009.

[46] D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.*, 14(7):460–470, Jul 2013.

[47] S. Gravel, B.M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.*, 108(29):11983–11988, Jul 2011.

[48] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, 5(10):e1000695, Oct 2009.

[49] G. S. Hageman, D. H. Anderson, L. V. Johnson, L. S. Hancox, A. J. Taiber, L. I. Hardisty, J. L. Hageman, H. A. Stockman, J. D. Borchardt, K. M. Gehrs, R. J. Smith, G. Silvestri, S. R. Russell, C. C. Klaver, I. Barbazetto, S. Chang, L. A. Yannuzzi, G. R. Barile, J. C. Merriam, R. T. Smith, A. K. Olsh, J. Bergeron, J. Zernant, J. E. Merriam, B. Gold, M. Dean, and R. Allikmets. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl. Acad. Sci. U.S.A.*, 102(20):7227–7232, May 2005.

[50] A. Hamosh, N. Sobreira, J. Hoover-Fong, V. R. Sutton, C. Boehm, F. Schiettecatte, and D. Valle. PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. *Hum. Mutat.*, 34(4):566–571, Apr 2013.

[51] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C. K. Chen, J. Chrast, J. Lagarde, J. G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S. E. Antonarakis, and R. Guigo. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, 7 Suppl 1:1–9, 2006.

[52] Daniel C. Hartl and Andrew G. Clark. *Principles of Population Genetics*. Sinauer and Associates, 2007.

[53] A. Hatzimanolis, J. A. McGrath, R. Wang, T. Li, P. C. Wong, G. Nestadt, P. S. Wolyniec, D. Valle, A. E. Pulver, and D. Avramopoulos. Multiple variants aggregate in the neuregulin signaling pathway in a subset of schizophrenia patients. *Transl Psychiatry*, 3:e264, 2013.

[54] Y. J. He, A. D. Misher, W. Irvin, A. Motsinger-Reif, H. L. McLeod, and J. M. Hoskins. Assessing the utility of whole genome amplified DNA as a template for DMET Plus array. *Clin. Chem. Lab. Med.*, 50(8):1329–1334, Aug 2012.

[55] W. G. Hill. Understanding and using quantitative genetic variation. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1537):73–85, Jan 2010.

[56] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, 106(23):9362–9367, Jun 2009.

[57] J. N. Hirschhorn. Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, 360(17):1699–1701, Apr 2009.

[58] Emily Hodges, Zhenyu Xuan, Vivekanand Balija, Melissa Kramer, Michael N Molla, Steven W Smith, Christina M Middle, Matthew J Rodesch, Thomas J Albert, Gregory J Hannon, and W Richard McCombie. Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, 39(12):1522–1527, 2007.

[59] Manuel Holtgrewe. Mason – a read simulator for second generation sequencing data. Technical report, Freie University, Math Department, 2010.

[60] R. H. Houwen, S. Baharloo, K. Blankenship, P. Raeymaekers, J. Juyn, L. A. Sandkuijl, and N. B. Freimer. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.*, 8(4):380–386, Dec 1994.

[61] World Health Organization http://goo.gl/HmLBT. Millions have hearing loss that can be improved or prevented.

[62] Illumina. Platinum genomes, August 2013.

[63] A. R. Indap, R. Cole, C. L. Runge, G. T. Marth, and M. Olivier. Variant discovery in targeted resequencing using whole genome amplified DNA. *BMC Genomics*, 14:468, 2013.

[64] A. R. Indap, G. T. Marth, C. A. Struble, P. Tonellato, and M. Olivier. Analysis of concordance of different haplotype block partitioning algorithms. *BMC Bioinformatics*, 6:303, 2005.

[65] National Center Biotechnology Information. Tmtc2 transmembrane and tetratricopeptide repeat containing 2 [homo sapiens (human)] http://www.ncbi.nlm.nih.gov/gene/160335, July 2013.

[66] I. Ionita-Laza, V. Makarov, S. Yoon, B. Raby, J. Buxbaum, D. L. Nicolae, and X. Lin. Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am. J. Hum. Genet.*, 89(6):701–712, Dec 2011.

[67] T. Jiang, L. Yang, H. Jiang, G. Tian, and X. Zhang. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. *Sci China Life Sci*, 54(10):945–952, Oct 2011.

[68] Y. W. Kan and A. M. Dozy. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl. Acad. Sci. U.S.A.*, 75(11):5631–5635, Nov 1978.

[69] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32(Database issue):D493–496, Jan 2004.

[70] A. Keinan and A. G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, May 2012.

[71] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, Sep 1989.

[72] A. Kiezun, K. Garimella, R. Do, N. O. Stitziel, B. M. Neale, P. J. McLaren, N. Gupta, P. Sklar, P. F. Sullivan, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, T. Lehner, Y. Y. Shugart, A. L. Price, P. I. de Bakker, S. M. Purcell, and S. R. Sunyaev. Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, 44(6):623–630, Jun 2012.

[73] A. H. Kissebah, G. E. Sonnenberg, J. Myklebust, M. Goldstein, K. Broman, R. G. James, J. A. Marks, G. R. Krakower, H. J. Jacob, J. Weber, L. Martin, J. Blangero, and A. G. Comuzzie. Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the Metabolic Syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 97(26):14478–14483, Dec 2000.

[74] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, Apr 2005.

[75] Dan Koboldt, August 2013.

[76] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[77] Jonas Korlach, Patrick J. Marks, Ronald L. Cicero, Jeremy J. Gray, Devon L. Murphy, Daniel B. Roitman, Thang T. Pham, Geoff A. Otto, Mathieu Foquet, and Stephen W. Turner. Selective aluminum passivation for targeted immobilization of single dna polymerase molecules in zero-mode waveguide nanostructures. *PNAS*, 105(4):1176–1181, 2008.

[78] L. Kruglyak. The road to genome-wide association studies. *Nat. Rev. Genet.*, 9(4):314–318, Apr 2008.

[79] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, 2009.

[80] Gerstein Lab. Vat - Variant Annotation Tool http://vat.gersteinlab.org.

[81] E. S. Lander and D. Botstein. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb. Symp. Quant. Biol.*, 51 Pt 1:49–62, 1986.

[82] H. Lango Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, and et. al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, Oct 2010.

[83] R S Lasken and M Egholm. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends in Biotechnology*, 21(12):531–535, 2003.

[84] S L Lauritzen and Nuala A Sheehan. Graphical models for genetic analyses. *Statistical Science*, 18(4):489–514, 2007.

[85] Wan-Ping Lee. Mosaik homepage. `http://bioinformatics.bc.edu/marthlab/Mosaik`.

[86] B. Li, W. Chen, X. Zhan, F. Busonero, S. Sanna, C. Sidore, F. Cucca, H. M. Kang, and G. R. Abecasis. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, 8(10):e1002944, 2012.

[87] B. Li, W. Chen, X. Zhan, F. Busonero, S. Sanna, C. Sidore, F. Cucca, H. M. Kang, and G. R. Abecasis. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, 8(10):e1002944, Oct 2012.

[88] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, 83(3):311–321, Sep 2008.

[89] B. Li and S. M. Leal. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.*, 5(5):e1000481, May 2009.

[90] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[91] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, Nov 2008.

[92] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):1–7, 2011.

[93] Kirk E Lohmueller, Amit R Indap, Steffen Schmidt, Adam R Boyko, Ryan D Hernandez, Melissa J Hubisz, John J Sninsky, Thomas J White, Shamil R Sunyaev, Rasmus Nielsen, Andrew G Clark, and Carlos D Bustamante. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181):994–997, 2008.

[94] Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer and Associates, 1998.

[95] Brendan Maher. Personal genomes: The case of the missing heritability., 2008.

[96] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.

[97] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, 11(7):499–511, Jul 2010.

[98] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, A Lisa, Jan Berka, Michael S Braverman, Yi-ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Goodwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E Mcdade, Michael P Mckenna, Eugene W Myers, Elizabeth Nickerson, R John, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, 437(7057):376–380, 2005.

[99] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, 23(4):452–456, Dec 1999.

[100] Gabor T Marth, Fuli Yu, Amit R Indap, Kiran Garimella, Simon Gravel, Wen Fung Leong, Chris Tyler-Smith, Matthew Bainbridge, Thomas Blackwell, Xiangqun Zheng-Bradley, Yuan Chen, Danny Challis, Laura Clarke, Edward V Ball, Kristian Cibulskis, David N Cooper, Bob Fulton, Chris Hartl, Dan Koboldt, Donna Muzny, Richard Smith, Carrie Sougnez, Chip Stewart, Alistair Ward, Jin Yu, Yali Xue, David Altshuler, Carlos D Bustamante, Andrew G Clark, Mark Daly, Mark Depristo, Paul Flicek, Stacey Gabriel, Elaine Mardis, Aarno Palotie, and Richard A Gibbs. The functional spectrum of low-frequency coding variation. *Genome Biology*, 12(9):R84, 2011.

[101] J. M. McClellan, E. Susser, and M. C. King. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry*, 190:194–199, Mar 2007.

[102] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, Jan 2010.

[103] S. J. Murphy, J. C. Cheville, S. Zarei, S. H. Johnson, R. A. Sikkink, F. Kosari, A. L. Feldman, B. W. Eckloff, R. J. Karnes, and G. Vasmatzis. Mate pair sequencing of whole-genome-amplified DNA following laser capture microdissection of prostate cancer. *DNA Res.*, 19(5):395–406, Oct 2012.

[104] Minato Nakazawa. Cran package fmsb. `http://cran.r-project.org/web/packages/fmsb/index.html`.

[105] Sarah B Ng, Abigail W Bigham, Kati J Buckingham, Mark C Hannibal, Margaret J McMillin, Heidi I Gildersleeve, Anita E Beck, Holly K Tabor, Gregory M Cooper, Heather C Mefford, Choli Lee, Emily H Turner, Joshua D Smith, Mark J Rieder, Koh-Ichiro Yoshiura, Naomichi Matsumoto, Tohru Ohta, Norio Niikawa, Deborah A Nickerson, Michael J Bamshad, and Jay Shendure. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics*, 42(9):790–793, 2010.

[106] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics*, 42(1):30–35, 2010.

[107] Z. Ning, A. J. Cox, and J. C. Mullikin. SSAHA: a fast search method for large DNA databases. *Genome Res.*, 11(10):1725–1729, Oct 2001.

[108] The Molecular Otolaryngology and Renal Research Lab University of Iowa. The deafness variation database - http://deafnessvariationdatabase.org/.

[109] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks

of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, Nov 2001.

[110] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S. I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, 30(3):265–270, Mar 2012.

[111] G. Peng, Y. Fan, T. B. Palculict, P. Shen, E. C. Ruteshouser, A. K. Chi, R. W. Davis, V. Huff, C. Scharfe, and W. Wang. Rare variant detection using family-based sequencing analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 110(10):3985–3990, Mar 2013.

[112] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, Jan 2010.

[113] Olivier Pourret, Patrick Naim, and Bruce Marco. *Bayesian Networks A Practical Guide to Applications*. John Wiley and Sons, 2008.

[114] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.*, 11(20):2417–2423, Oct 2002.

[115] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, 19(7):1316–1323, Jul 2009.

[116] E. G. Puffenberger, E. R. Kauffman, S. Bolk, T. C. Matise, S. S. Washington, M. Angrist, J. Weissenbach, K. L. Garver, M. Mascari, and R. Ladda. Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.*, 3(8):1217–1225, Aug 1994.

[117] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Res.*, 40(Database issue):290–301, Jan 2012.

[118] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool

set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, Sep 2007.

[119] A R Quinlan and I M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[120] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[121] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17(9):502–510, Sep 2001.

[122] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[123] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, Feb 2001.

[124] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, Dec 1977.

[125] E.E. Schadt. *New methods and new technologies for preclinical and clinical neurobiology*, chapter Network methods for elucidating the complexity of common human diseases. Oxford University Press, 2013.

[126] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, 15(11):1576–1583, Nov 2005.

[127] M. Schraders, J. Oostrik, P. L. Huygen, T. M. Strom, E. van Wijk, H. P. Kunst, L. H. Hoefsloot, C. W. Cremers, R. J. Admiraal, and H. Kremer. Mutations in PTPRQ are a cause of autosomal-recessive nonsyndromic hearing impairment DFNB84 and associated with vestibular dysfunction. *Am. J. Hum. Genet.*, 86(4):604–610, Apr 2010.

[128] Jay Shendure. Next-generation human genetics. *Genome Biology*, 12(9):408, 2011.

[129] A. Sirmaci, Y. J. Edwards, H. Akay, and M. Tekin. Challenges in whole exome sequencing: an example from hereditary deafness. *PLoS ONE*, 7(2):e32000, 2012.

[130] A. B. Skvorak Giersch and C. C. Morton. Genetic causes of nonsyndromic hearing loss. *Curr. Opin. Pediatr.*, 11(6):551–557, Dec 1999.

[131] Nathan O Stitziel, Adam Kiezun, and Shamil Sunyaev. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology*, 12(9):227, 2011.

[132] Picard Development Team. Picard homepage. `http://picard.sourceforge.net`.

[133] Life Technologies. Ion torrent.

[134] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, and J. M. Akey. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, Jul 2012.

[135] Duncan C. Thomas. *Statistical Methods in Genetic Epidemiology*. Oxford University Press, 2004.

[136] K. R. Thornton, A. J. Foran, and A. D. Long. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS Genet.*, 9(2):e1003258, Feb 2013.

[137] E. H. Turner, S. B. Ng, D. A. Nickerson, and J. Shendure. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet*, 10:263–284, 2009.

[138] TMHMM Server v2.0. Tmhmm server v2.0 http://www.cbs.dtu.dk/services/tmhmm/.

[139] John Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, 2009.

[140] Tom Walsh, Hashem Shahin, Tal Elkan-Miller, Ming K. Lee, Anne M. Thornton, Wendy Roeb, Amal Abu Rayyan, Suheir Loulus, Karen B. Avraham, Mary-Claire King, and Moien Kanaan. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein {GPSM2} as the cause of nonsyndromic hearing loss {DFNB82}. *The American Journal of Human Genetics*, 87(1):90 – 94, 2010.

[141] M. N. Weedon, H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans, M. Mangino, R. M. Freathy, J. R. Perry, S. Stevens, A. S. Hall, N. J. Samani, B. Shields, I. Prokopenko, M. Farrall, A. Dominiczak, T. Johnson, S. Bergmann,

J. S. Beckmann, P. Vollenweider, D. M. Waterworth, V. Mooser, C. N. Palmer, A. D. Morris, W. H. Ouwehand, J. H. Zhao, S. Li, R. J. Loos, I. Barroso, P. Deloukas, M. S. Sandhu, E. Wheeler, N. Soranzo, M. Inouye, N. J. Wareham, M. Caulfield, P. B. Munroe, A. T. Hattersley, M. I. McCarthy, and T. M. Frayling. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, 40(5):575–583, May 2008.

[142] J. Weissenbach, G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop. A second-generation linkage map of the human genome. *Nature*, 359(6398):794–801, Oct 1992.

[143] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 1921.

[144] Sewall Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 1934.

[145] J. Xing, W. S. Watkins, Y. Zhang, D. J. Witherspoon, and L. B. Jorde. High fidelity of whole-genome amplified DNA on high-density single nucleotide polymorphism arrays. *Genomics*, 92(6):452–456, Dec 2008.

[146] D. Yan, M. Tekin, S. H. Blanton, and X. Z. Liu. Next-Generation Sequencing in Genetic Hearing Loss. *Genet Test Mol Biomarkers*, Jun 2013.

[147] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, Jul 2010.

[148] KemalO. Yariz, Duygu Duman, CeliaZazo Seco, Julia Dallman, Mingqian Huang, TheoA. Peters, Asli Sirmaci, Na Lu, Margit Schraders, Isaac Skromne, Jaap Oostrik, Oscar Diaz-Horta, JuanI. Young, Suna Tokgoz-Yilmaz, Ozlem Konukseven, Hashem Shahin, Lisette Hetterschijt, Moien Kanaan, AnneM.M. Oonk, YvonneJ.K. Edwards, Huawei Li, Semra Atalay, Susan Blanton, AlexandraA. DeSmidt, Xue-Zhong Liu, RonaldJ.E. Pennings, Zhongmin Lu, Zheng-Yi Chen, Hannie Kremer, and Mustafa Tekin. Mutations in otogl, encoding the inner ear protein otogelin-like, cause moderate sensorineural hearing loss. *The American Journal of Human Genetics*, 91(5):872 – 882, 2012.

[149] Y. Zhang. A dynamic Bayesian Markov model for phasing and characterizing haplotypes in next-generation sequencing. *Bioinformatics*, 29(7):878–885, Apr 2013.