# ABSTRACT

KEEBLER, JONATHAN EDWARD MYERS. Spontaneous Mutation Discovery via High-Throughput Sequencing of Pedigrees. (Under the direction of Eric Stone.)

Recent technological advances have made high-throughput DNA sequencing a routine laboratory experiment. This progression in technology has been made possible by the parallel production of millions of short fragments of sequence. The responsibility of garnering biological information from these DNA fragments has shifted from the wet-lab to the bioinformatician. As sequencing technology is applied to a growing number of individual human genomes, entire families are now being sequenced. Information contained within the pedigree of a sequenced family can be leveraged when inferring the donors' genotypes, a task that is not necessarily trivial using high-throughput sequencing reads. A violation of Mendelian inheritance laws observed amid the resequenced genomes of family members can indicate the presence of a *de novo* mutation. A method for locating *de novo* mutations by probabilistically inferring genotypes across a pedigree using high-throughput sequencing is presented and applied to two resequenced nuclear families: one as a collaborative effort within The 1,000 Genomes Project, and the second in an attempt to discover candidate driver and passenger mutations within the genome of an Acute Lymphoblastic Leukemia. The mutation findings within these projects are presented, and the approach is examined in detail, highlighting areas where method improvements may be made. Considering the challenges experienced in these studies within the larger context of the nascent field of Personal Genomics, an honest assessment is presented of developments that must be made before the application of whole-genome sequencing on the scale of an individual human can unequivocally be used to predict, diagnose, or treat human disease.

# Spontaneous Mutation Discovery via High-Throughput Sequencing of Pedigrees

by
Jonathan Edward Myers Keebler

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2010

APPROVED BY:

| | |
|---|---|
| _____ | _____ |
| Eric Stone<br>Committee Chair | Philip Awadalla |
| _____ | _____ |
| Jeffrey Thorne | Ignazio Carbone |
| _____ | |
| Alison Motsinger-Reif | |

# DEDICATION

To Susan and Larry Keebler, Mary and Jason Becker: your unconditional love, encouragement, and benevolent guidance gives me strength and focus when it is needed most.

To Ian Edward Becker: may your heart always be joyful, may your song always be sung, and may you stay forever young.

To Adam Howell and Andy Young: throughout the past twenty years your friendship has been an inspiration in more ways than you can know.

# BIOGRAPHY

Jonathan Keebler was born in August 29, 1980 to Larry and Susan Keebler in Johnson City, TN. Growing up in near-by Elizabethton, TN, Jonathan was fortunate to have been taught by remarkable educators within the Elizabethon public school system. Those most influential were Meleta Kardos, Rick Simerly, Richard Culver, William Church, Lisa-Dawn Smith, Sam Greenwell, Perry Elliott, Reuben Pierce, and Tara Peters. Throughout junior high and high school Jonathan was actively involved in athletics and music, participating in the marching, jazz, and symphonic bands and a member of the soccer, football, and track-and-field teams. Jonathan attended East Tennessee State Univiersity beginning in the fall of 1998 and pursued a Bachelor of Science degree in Computer Science, focusing on software engineering. Jonathan continued also continued his involvment with muic, acquiring a Minor in Music-Performance. During his senior year, he was fortunate to work as an intern at Johnson and Johnson compnay, Ethicon Endo-Surgury, where he was introduced to developing software for use in clinical medicine. Following graduation from ETSU, Jonathan elected to continue to apply his training as a software developer within a scientific context that would contribute to the advancement of health care. He began his graduate work at North Carolina State University in the Genomic Sciences graduate program in the fall of 2004, with the goal of attaining a Doctorate of Philosophy in Bioinformatics. As he fulfulled the course

requirements, Jonathan was attracted to the field of Population Genetics. He was first introduced to the field of genomics during an internship with Richard Gibbs at the Baylor College of Medicine during the summer of 2005. He joined the lab of Philip Awadalla n the fall of 2005 where he developed software for discovering novel genetic variation within Plasmodium falciparum, the parasite responsible for the most lethal form of Malaria in humans, a work which was published in Nature Genetis in 2006. Witness to the rise of high-throughput sequencing, Jonathan focused his efforts on obtaining useful genotype information from these experiments and applying them to fundamental biological questions concerning the evolution of humans and the mutations which contribute to human disease. Under the direction of Eric Stone and Philip Awadalla, Jonathan has worked towards these goals while collaborating with clinical researchers studying cancer at Saint-Justine Research Hospital in Montreal and members of the analysis group within the 1,000 Genomes Project at the Wellcome Trust Sanger Institute and the Broad Institute.

# ACKNOWLEDGEMENTS

This work was made possible primarily by the generous support and guidance of my graduate advisor, Philip Awadalla. Dr. Awadalla is credited for the initial conception and encouraging impetus for this and many noteworthy projects throughout my five-year tenure with his lab. In addition, the advice and mentorship provided by Eric Stone have been instrumental to my success. Whether over a cup of coffee or hours in front of a white board, his contributions towards my development have been significant. Reed Cartwright played a crucial role in the production of the mathematical model herein described, and he consistantly made himself available for direction as I struggled through the implementation effort. I appreciated working with Jeffrey Thorne, Ignazio Carbone, and Alison Motsinger, and I was fortunate to have their service as members of my graduate committee.

Ferran Casals and Youssef Idaghdour are thanked for their many weeks of work in the wet lab validating mutation predictions. Diego Czul was helpful as well in data management and scripting support. Less tangible but no less significant was the support provided by past and current members of the Awadalla lab including Kate McGee, Rachel Myers, Martine Zilversmit, Julie Hussin, Jackie Quinlan, and Natalie Tishenko.

It has been my pleasure to work with Don Conrad and Matt Hurles as members of the analysis group within the 1,000 Genomes Project. They contributed greatly in the analysis effort and remain exceedingly generous with access to their own findings and

methods, feuling the collaborative effort. Mark DiPresto and Mark Daly are also acknowledged for access to their mutation predictions.

Daniel Sinnet, Ekat Kritkou, and Mathieu Lavrivière are credited for their work in studying the genomics of Acute Lymphoblastic Leukemia.

Thanks are owed to Alison Motsinger, Nick Hardison and Zeke Harris for consistent technical support and generous use of their computational resources at the Bioinformatics Research Center, NCSU.

Deepest appreciation is given to Rachel Myers, Ben Dorshorst, Lisa McFerrin, and Shengdar Tsai, who each have acted admirably as both friends and colleagues throughout my graduate education and thier contributions cannot be measured.

# TABLE OF CONTENTS


LIST OF TABLES ..... x
LIST OF FIGURES..... xi

Chapter I A Review of Personal Genomics ..... 1
Introduction ..... 2
Human population genetics and applications to human disease..... 6
The Human genome and types of genetic variation ..... 6
Introduction to DNA Sequencing ..... 8
DNA sequencing reads and assembly ..... 8
Inferring genotype from read alignm..... 10
Human Genome Sequencing Review ..... 11
The genomes of Craig Venter and James Watson ..... 12
Three personal genomes released in November, 2008 ..... 16
Personal genomics in 2009..... 19
The current generation of personal genomics..... 24
Conclusion – Lingering Challenges in Human Genomics..... 27
References..... 34

Chapter II Probabilistic Inference of Pedigree Genotypes using High-throughput Sequencing..... 41
Introduction..... 42
Methods..... 46
Model ..... 49
Probability of the observed data..... 52
Kernel functions..... 57
Somatic mutation kernel..... 58
Germline mutation kernel..... 58
Error kernel..... 59
Population kernel..... 61
Tree peeling algorithm..... 67
Summary statistics..... 69
Calculating the probability of de novo mutation, $\delta$ ..... 72
Maximum-likelihood estimation of model parameters..... 73
Model extension to larger pedigrees..... 75
Distinguishing sources of mutation using various pedigree structures..... 79
Method Implementation..... 86
Simulation Studies..... 89

Single-site simulations..... 90
Trio simulation study..... 94
Monozygotic twin simulation study..... 106
Conclusion and Future Work..... 110
References..... 113

Chapter III Discovering *de novo* Mutations in Collaboration with The 1,000 Genomes Project Analysis Group..... 116
Introduction..... 117
The 1,000 Genomes Project..... 117
De novo Mutation Discovery in the 1,000 Genomes Project Pilot 2 Trios..... 121
Methods..... 122
Data pre-processing and compression..... 122
Application of Piphits to the 1,000 Genomes Pilot 2 Trios..... 124
Candidate mutation list post-processing..... 126
Experimental validation of the predicted mutations..... 130
Results..... 136
Validated Mutations..... 136
Piphits false positives..... 142
Piphits false negatives..... 148
Piphits true negatives..... 152
Conclusion..... 154
References..... 156

Chapter IV Discovery of Somatic Mutations within the Sequenced Exome of an Acute Lymphoblastic Leukemia..... 158
Introduction..... 159
Methods..... 160
Somatic mutation discovery by testing the inheritance of single nucleotide polymorphisms (SNP-method) ..... 165
Somatic mutation discovery using two applications of Piphits..... 167
Results..... 171
SNP-method mutation predictions for genes listed by the Cancer Gene Census..... 172
SNP-method prediction of nonsense mutations..... 173
Mutations within untranslated regions predicted by the SNP-method..... 175
Mutations at dbSNP loci predicted by the SNP-method..... 177

Specific effects of a coding mutation
predicted exclusively by Piphits-CGC..... 179
Discussion..... 182
Discrepancies between the predicted
mutation lists from each method..... 182
Assessment of SNP-method compared to Piphits
in predicting mutations in the quartet sample..... 185
Conclusion..... 188
References..... 190

Chapter V Concluding Remarks..... 196

APPENDIX..... 201

# LIST OF TABLES


Table 2-1. The data structure used to represent the observed nucleotides at a site for all three members of a trio nuclear family ..... 48

Table 3-1. Custom Data Quality Filters ..... 124

Table 3-2. Maximum likelihood estimates of model parameters ..... 126

Table 3-3. Bowtie alignment results of UdeM validation reads ..... 135

Table 3-4. Validation classification rules based on observations of the putative mutant allele ..... 136

Table 3-5. Base substitution patters for validated *de novo* mutations ..... 142

Table 4-1. Quartet Bowtie alignment results ..... 162

Table 4-2. Prior model parameters used for the applications of Piphits-CGC and Piphits-EW ..... 169

Table 4-3. SNP-method predicted mutations at sites within genes within The Cancer Gene Census list. ..... 173

Table 4-4. Putative nonsense mutations predicted by the SNP-method ..... 175

Table 4-5. Mutations predicted by the SNP-method within untranslated regions ..... 176

Table 4-6. Predicted mutations at sites discordant with dbSNP genotypes ..... 178

Table 4-7. Observed read data supporting a mutation call at 5:55283865, an E508* nonsense mutation in *IL6ST* ..... 182

Table A-1. Coding somatic mutations predicted by the Chapter IV SNP-method ..... 202

Table A-2. Coding somatic mutations predicted by Piphits-CGC in Chapter IV ..... 212

Table A-3. Coding somatic mutations predicted by Piphits-EW in Chapter IV ..... 212

Table A-4. Glossary of Genes Referenced ..... 214

# LIST OF FIGURES


Figure 2-1. Trio model for a single site ..... 52

Figure 2-2. Possible genealogical trees for a sample of four alleles from a single population ..... 63

Figure 2-3. Extended pedigree model ..... 76

Figure 2-4. Identifiable and non-identifiable un-rooted pedigrees ..... 80

Figure 2-5. Family resequencing study designs with IBD allele pedigrees ..... 82

Figure 2-6. Mendelian inheritance of the Two-sib, Three-sib, and Grandparent study designs ..... 85

Figure 2-7. Transition points in model inferences ..... 92

Figure 2-8. Trio simulation estimation results at 20x coverage ..... 96

Figure 2-9. Variable coverage trio simulation parameter estimation results ..... 99

Figure 2-10. Trio simulation parameter results with variable coverage and no alignment error ..... 100

Figure 2-11. Trio simulation ROC curves ..... 103

Figure 2-12. True positive, false positive, and false negative *de novo* mutation predictions in trio simulations ..... 105

Figure 2-13. Rate estimation results for the first set of quartet simulation replicates ..... 108

Figure 2-14. Rate estimation results for the second set of quartet simulation replicates ..... 109

Figure 3-1. The 1,000 Genomes Project Pilot 2 Deep Sequencing Trios ..... 120

Figure 3-2. Generation of the Validation Set for the CEU Family ..... 128

Figure 3-3. Generation of the Validation Set for the YRI Family ..... 129

Figure 3-4. Piphits confidence measure distribution for final candidate site list ..... 130

Figure 3-5. Overlap of CEU and YRI candidate sites with WTSI and BI groups ..... 131

Figure 3-6. UdeM validation capture procedure used for each sample ..... 133

Figure 3-7. Validation pedigree for CEU family ..... 134

Figure 3-8. Validation classification results across both families for all prediction lists ..... 137

Figure 3-9. Illustration of genomic locations of validated *de novo* mutations for the YRI and CEU families ..... 138

Figure 3-10. Relative contribution to each validation class for each family by each group ..... 140

Figure 3-11. Proportion of Piphits predicted mutation sites for each validation result ..... 143

Figure 3-12. Correlation of validation results across both families with the Piphits confidence measure $\delta$ ..... 144

Figure 3-13. Validation classifications of low- ($\delta$<0.90) and high- ($\delta$>0.90) confidence mutation predictions made by Piphits ..... 145

Figure 3-14. Integrative Genomics Viewer (IGV) multiple sequence alignment for the 1kGP reads at falsely-predicted site 8:77023179 for YRI individual 40 ..... 147

Figure 3-15. Mapping and Base qualities for reads carrying the reference and mutant allele at site 8:77023179 in YRI individual 40 ..... 148

Figure 3-16. IGV multiple sequence alignment for the 1kGP reads mapping to false-negative site 1:110099347 for individual 78, child of the CEU trio ..... 150

Figure 3-17. Original mapping and base qualities for reads carrying the reference and mutant allele at site 1:110099347 In CEU individual 78 ..... 151

Figure 3-18. Mutaion falsely predicted by BI and WTSI but not by Piphits ..... 153

Figure 4-1. Pedigree relating the quartet blood samples acquired ..... 160

Figure 4-2. Pedigree model for the four blood samples ..... 164

Figure 4-3. Procedure for culling somatic mutatins from the leukemia SNP list ..... 167

Figure 4-4. SNP-quality and Piphits probability of mutation plotted against the percent coverage in the leukemia sample ..... 187

# Chapter I

# A Review of Personal Genomics

# Introduction

## *Human population genetics and applications to human disease*

Genetic diversity underlies the biodiversity found on Earth. Similarly, the modifications in biochemical circuitry prompted by variations in the primary nucleotide sequence of deoxyribonucleic acid (DNA) have shaped many human characteristics. The human genome, however, is not a single sequence of nucleic acids, but is better described as a catalog of many variations of a common theme. Linking the differences in measurable human traits, henceforth referred to as phenotype, to a specific molecular genetic variation, or genotype, is a central aim of human genetics. Gaining a comprehensive understanding of human DNA sequence variation, including the types, or alleles, present and the frequency at which they occur, is a major component of this pursuit. Population genetics is the study of this genetic variation, including the natural mechanisms that create, maintain, and destroy allelic diversity between and within biological populations. The resulting DNA sequence variation is not entirely random, but has been formed by natural forces acting in concert: the sampling of alleles from one generation to the next (genetic drift), population demographics such as size changes and migration, stochastic replication errors and environmental factors prompting spontaneous mutation, and natural selection acting primarily to purge deleterious alleles and promote the propagation of advantageous ones. Population genetics provides a critical evolutionary context for the modern human genome just as history and anthropology can provide context for human social and political relationships[1]. In this manner population genetics has provided powerful insights not only

to how genotype-phenotype correlations can be uncovered, but into our history as a species as well.

The phenotypes which we are perhaps most interested in linking to a genetic basis is that of disease and disease susceptibility. Perhaps the most striking modern example of the contribution population genetics has had to this end can be seen in the modern debate of the significance of rare variations[2,3]. Rare genetic variants, distinguished from 'common' variants, have a population minor allele frequency (MAF) of less than 1%[4]. The debate is centered on the frequency at which the variations of functional importance can be expected to occur underlying complex traits, specifically common human diseases. Such traits are characterized as having continuously distributed phenotypes resulting from the independent action of many genes, as well as being influenced by the environment and gene-by-environment interactions[4]. The 'common disease, common variant' (CDCV)[5,6] hypothesis expects such variants to have a more ancient origin and occur at higher population frequencies, where as the 'common disease, rare variant' (CDRV)[7] argues they are more likely to have a more recent origin, be population specific, and be found at lower frequencies[3,8]. By applying population genetic principles to the debate, a model put forward by Jonathan Pritchard in 2001 suggests rare variants can play a significant role, largely due to the forces of weak purifying selection. More specifically, given the human population expansion over the past few hundred of years, many mutations have been able to accumulate that are mildly deleterious, since damaging mutations occur at a higher rate than repairing or protective mutations, whereas because common mutations are older, they

have been subjected to natural selection for longer periods of time, and are then less likely to have negative impacts[7]. This hypothesis has seen some support in the inability of very large studies based on CDCV theory[9] to describe more than 10% of the heritable risk in complex diseases[10,11], and the finding of many more novel alleles in resequenced personal genomes than expected[12]. In addition, population genetic theory predicts that less frequent mutations are more likely to have functional effects, a postulation that has recently been seen support in empirical findings[2]. In truth, there are likely some diseases within some populations that are governed by the CDCV paradigm and others by CDRV, and the strategies used for studying their respective genetic bases should take into account the relevant evidence for each and characteristics of the population in question.

The primary method for locating the common alleles responsible for complex human diseases has been to conduct genome wide association studies (GWAS) based on common single nucleotide polymorphisms (SNPs) that show a statistically significant link with a particular disease[11]. While a GWAS does not require an identified SNP to directly contribute disease phenotype, it requires that linkage disequilibrium (LD) exist between the queried SNP and the causal variants. LD is a term derived from population genetics which quantifies the probability that two alleles at two DNA sequence positions, or loci, are inherited together and is usually indicative of the physical DNA distance between the two loci or the frequency recombination occurs between them. In order for a common SNP, which by having a higher population frequency is likely to have an ancient origin, to be in strong LD with a disease-causing allele, the causal allele is also likely to be ancient and

common. To date hundreds of GWA studies have been conducted upon many kinds of diseases, but only marginal success has been achieved at locating the heritable risk to these diseases, leaving 90% to 95% of the heritable component of a disease unexplained[3,11,13].

Due to the requirement of the disease-causing allele to be in LD with a common SNP, GWAS are unable to detect rare variants acting in concert, each contributing a small amount to the overall inherited susceptibility of a disease. These variants are instead detectable through the process of DNA resequencing[3]. Sequencing technology has advanced such that obtaining the primary nucleotide sequence for many candidate genes is now affordable on an unprecedented scale[14]. Specifically, an investigator could proceed by sequencing candidate genes in individuals who are symptomatic and have a family history for a particular disease, as well as in control individuals who are healthy and have no affected relatives. After giving a consideration for any population stratification effects that may exist, the sequences can be directly compared to identify functional, rare variants. Such variations can readily be screened for functional consequences, including occurrence within a conserved region and conveying a protein charge or structure modification. The positive results can be singular or occur as a group, affecting the same gene or set of genes with related functions. Rather than limiting investigation to a small set of gene candidates, an argument exists for exploring diverse regions of the genome that may have functional consequences for the disease phenotype in question. Parts of the genome once considered 'junk DNA' may contain undiscovered functional elements[13]. While such a study has yet

to be presented, with the ability to sequence entire genomes, it is becoming plausible as the number of published personal genomes continues to increase.

Following a brief introduction to some well understood characteristics of the human genome and the types of variation found, the remainder of this review will consist of an overview of DNA sequencing technology, followed by a specific report on each of the personal genomes reported to date, with remarks centered on their contributions to the knowledge of human genetic variation. Finally current limitations of the whole-genome sequencing will be highlighted with comments on areas the field will need to see improvement before becoming fully applicable to human medical genetics.

## *The Human genome and types of genetic variation*

Each human genome is composed of two sets of 23 chromosomes, one inherited from each parent[15]. In females, all 23 chromosomes are present in homologous pairs while males have 22 homologous chromosomes, with a single copy each of the X- and Y-chromosomes[16]. Variation will exist between the two homologous chromosomes in proportion to the antiquity of the most recent common ancestor between the parents of diploid genome in question. A fully sequenced diploid human genome will include characterization loci where between-chromosome variation exists. Such sites are termed heterozygous, while sites at which both chromosomes bear the same allele are homozygous. Haploid genomes only report a single allele at each heterozygous site.

In addition to the frequency at which genetic variants can be observed presented earlier as the distinction between 'common' and 'rare,' genetic variation can also be classified by degrees of functional consequence and by nucleotide composition. Since most new mutations will be mildly deleterious, and therefore removed by the action of purifying selection, most of the observed genetic variation will be selectively neutral, having no functional consequence. This is concept is a tenet of population genetic theory as part of the Neutral Theory of Molecular evolution put forth by Motoo Kimura[17]. Molecular data largely supports this theory, and evidence of positive selection, which occurs when a mutation grants a functional improvement, is generally rare[1]. However, empirical values have yet to be determined for the absolute relative percentage of neutral, near-neutral, and non-neutral variants[4]. In terms of nucleotide composition, in the broadest sense genetic variation is broken into single nucleotide polymorphisms and structural variations (SVs). Structural variations range in size from one to two bases up to several megabases in length and include insertions or deletions (indels), inversions, copy number variations (CNVs), and translocations. While most of the focus within human genetics has been on SNPs, much more attention has recently been given to SVs as they can account for a substantial fraction of genetic diversity overall, and have the propensity to be causative in disease[18].

# Introduction to DNA Sequencing

Many techniques and technologies have been developed towards fully categorizing the sequence variation found in genomes. In the early work of Hubby and Lewontin[19], gel electrophoresis was introduced as a method for examining the diversity in protein mobility between individuals as a surrogate for DNA sequence changes. However, as noted by Kreitman in 1983[20], there are many nucleotide differences that do not alter the amino acid sequence, protein mobility itself is a phenotypic measurement, and that it is therefore necessary to directly study the DNA sequence. Kreitman's study of the primary sequence of the alcohol dehydrogenase locus within eleven individual *Drosophila melanogaster* fruit flies examined SNPs and CNVs including those with protein-coding effects and those without[20]. The technology leveraged by Kreitman that was not available to Hubby and Lewontin has now been termed Sanger sequencing, named for the investigator who introduced it, and has become the gold standard for determining directly determining DNA sequence.

## *DNA sequencing reads and assembly*

The fundamental output unit from DNA sequencing technologies is the read. A read is a subsequence of DNA nucleotides, ranging in length from approximately 30 to 1000 bases. In whole genome shotgun (WGS) strategies, which most modern methods employ, the region sampled by each read is uniform across the genome. Reconstructing the original DNA sequence from which these reads were taken is the process known as

assembly; reads are required to overlap, covering the same regions of DNA sequence multiple times. Due to the randomness of the WGS approach, the number of reads overlapping in any single location is assumed to follow a Poisson distribution with a mean, termed the fold-coverage or sequence depth, equal to the average read length times the number of reads produced, divided by the length of the genome. Given a particular genome length and read length, the desired level of fold-coverage can be approximately set by sequencing the calculated number of requisite reads (~900 million 100-base pair reads will provide 30x coverage of a mammalian-sized genome[21]). Achieving a high coverage level is important for not only assembly purposes, but for capturing heterozygous variation and producing a true diploid genome.

Errors in the alignment or assembly process can reduce the actual level of coverage achieved, when some reads are not placed along the genome at the position where they were originally sampled. One strategy that assists in the assembly or alignment process is the creation of the read libraries in 'mate-pairs' or 'paired-ends'[22]. From an assembler's view, these two are virtually identical. They each produce reads in pairs that are a known distance apart from one another. The DNA sequence between the ends is called the insert. The methods differ in that 'mate-pair' reads have larger insert sizes so cover a larger area of the genome, while 'paired-end' reads are closer. The insert lengths between 'mate-pair' reads can vary between pairs, and 'paired-end' reads have a more precisely known insert size[22]. When one read aligns, information is immediately available concerning the expected position of the second read. If the second read does not align correctly, mapping

closer, further, or in reverse orientation than what is expected relative to its mate, it is possible to capture deletions, insertions, or inversions of the subject's DNA sequence. These types of variation fall under the category of structural variants and can play important roles in genetic diseases[22].

### *Inferring genotype from read alignment*

Capturing both alleles at heterozygous loci is crucial to obtain an accurate picture of the genetic variation present. Missing these sources of variation will not only influence population genetic conclusions observed levels of heterozygosity, but heterozygous positions can play an important role in both causing[23-25] and offering protection from disease[26]. A single read can sample only one of two chromosomes in a diploid genome, so capturing both alleles at a heterozygous site can be considered a binomial sampling problem: a sufficiently high fold-coverage is necessary to make sampling only one of the two chromosomes unlikely. The idea is equivalent to tossing a coin a sufficient number of times to guarantee seeing both sides of the coin land face-up at least once. There is often a bias between which chromosome is sampled in practice, so the binomial probability is not always equal to 0.5. Assuming no such bias between chromosomes exists, sampling a single site at a depth of 2x will sample both chromosomes 50% of the time, while using a 10x fold coverage of a site will produce at least one sequence from both chromosomes ~99.8% of the time. Taking into account the possibility of sequencing errors complicates

the task; when one out of ten reads contains a unique allele, it may more likely be due to sequencing error than represent heterozygous position.

## Human Genome Sequencing Review

The DNA sequencing method introduced by Sanger, et al in 1977[27] was used to produce the first draft sequences of the human genome[28,29], drawing genetic material from several individuals. Since the early 1990's large-scale sequencing has used 'Sanger sequencing' almost exclusively. Gradual improvements over the years have led to a high per-base accuracy (99.999%) and long sequencing reads (~1,000 bases) at a cost on the order of $0.50 per thousand bases produced in high-throughput contexts[30]. In many current applications, Sanger sequencing is used as the validation method for predicted results[31], or as a reliable strategy for direct re-sequencing of a small target region[32]. However, in high-throughput scenarios this method is costly in terms of reagents, machinery, and personnel, and is relatively slow compared with modern methods which are more scalable, automated, and parallelizable[33].

Sanger sequencing is considered the first-generation of sequencing technology, and those immediately following it are collectively referred to as 'next-generation' sequencing (NGS). The general 'NGS' moniker has remained a popular aggregate term despite the emergence of newer generations of technologies. All whole genome, shotgun DNA sequencing experiments have a small set of general steps in common: 1) sample preparation, 2) sequencing reaction, 3) assembly of the data (or forming a consensus

sequence), and 4) variation capturing and other analyses. Steps 1) and 2) are primarily based on the type of sequencing technology being used. These steps largely deal with detailed biochemistry and are often proprietary to the provider of the technology, so they will not be covered in extensive detail. There are many thorough reviews available describing and comparing each of the technologies mentioned here, and the interested reader is encouraged to examine any of these[18,30,34-38].

### *The genomes of Craig Venter and James Watson*

Beginning in 2007, the first fully resequenced diploid human genomes began to appear with the genome of two well-known scientists: Craig Venter [12], and James Watson in 2008 [39]. After these initial genome reports came a burst of resequencing activity, with three more studies appearing in 2008, five in 2009, and the most recent in 2010 including the whole genomes of three separate individuals. In this section these projects are reviewed, and a discussion of their contributions to the body of knowledge about human genetic variation is presented.

While Sanger sequencing was used in partial validation for many of these studies, only the first personal genome released (Venter) in September 2007 used it exclusively[12]. While this came at a cost exceeding $70 million, the effort was rewarded by achieving per-base accuracies as high as 99.999% and sequencing reads 800 bases in length[30]. Prior to this undertaking, there were two haploid versions of the human genome available: the Human Genome Sequencing Consortium (HGSC)[40] assembly is a composite haploid

sequence derived from numerous donors, and the Celera Genomics assembly is a consensus sequence derived from five individuals[29]. Virtually all of the genetic variation in these two assemblies was reported in the form of SNPs, although it became well understood that unreported structural variants (SVs) were equally significant: indels, copy number variants (CNVs), and segmental duplications[41-44]. An important feature of the Venter genome and the personal genomes following it is the analysis of all classes of variation. All variations located in the Venter sequence are reported in comparison with the HGSC resulting sequence: the National Center for Biotechnology Information version 36 of the human genome (NCBI36).

One of the major results from the Venter genome was only made possible by use of the Sanger sequencing method. Due to the length of the reads produced, as well as the utilization of mate-pairs, individual haplotypes could be constructed, specifying the chromosome on which each allele resides. This procedure is also referred to as determining the phase of the data, and is useful since haplotypes have more power to assist in phenotype-genotype association studies as well as predicting disease risk[45,46]. Prior to this study, haplotypes were constructed based upon genotyping studies of known variant locations, which can be separated by regions of high recombination rates, i.e. low LD. For this genome, haplotypes were identified covering 75% of the autosome, across regions of low LD and included many novel variant positions that were not available to previous studies. Each haplotype carried at least four variants.

In total this study identified 4.1 million variants in the Venter sequence, 30% of which had not been described previously. Of the 4.1 million variants, 3.2 million were classified as SNPs with ~45% being homozygous. This is a higher percentage than had been previously reported in population-based studies[12]. The remaining variants were SVs ranging in size from 1 base to 670,345 bases, including .82 million indels, 82 thousand "complex" and "multiple nucleotide polymorphisms", and 90 inversions. Notably, while 78% of all variants detected were SNPs, the vast majority (74%) of all variable bases are located within SVs. Focusing on the indels, again a higher than expected number homozygotic variant positions were identified (~68% of all indels). Several factors were suggested which contributed to the abundance of homozygous variants. One contributing factor is the low sequencing depth (7.5-fold) of the study. With this level of read redundancy, it's conceivable that many times only one chromosome at a heterozygous position is sampled. Indeed, given the filtering criteria used in variant discovery, the investigators statistically determined that between 44% and 54% of the time heterozygous variants would be missed due to insufficient read coverage at 7.5x. This was confirmed when they attempted to validate their SNPs with hybridization-based SNP microarrays[12]. Over 91% of the SNP genotypes predicted were concordant with the microarray validation, while 84% of those that were discordant were heterozygous positions incorrectly labeled as homozygous. While producing an unrivaled representation of the human genome in haplotype structure, this study perhaps more importantly illustrated the necessity of deep coverage levels when performing WGS on the human genome.

Six months after the publication of the Venter genome, the second personal genome was released[39], that of James D. Watson, the coauthor of the landmark study that originally identified the structure of the DNA molecule[47]. The Watson genome was the first constructed using NGS technology, and was sequenced in 2 months at a reagent cost of US $1 million[18], a small fraction of resources required for the Venter genome. The sequencing technology used was the 454 Sequencing platform by Roche[48]. They generated 106.5 million high-quality reads, each with a length of 250 bases, which were aligned to the NCBI36 genome to locate sequence variants. They achieved a 7.4x fold-coverage, similar to that of the Venter genome. Therefore one would expect this study to have the same difficulty in determining genotype due to failing to sample both chromosomes at heterozygous positions. Indeed, this was shown to be a problem as microarray genotyping indicated at least 32,770 heterozygous variants out of the 135,413 tested were missed (24.2%)[39]. In total they located 3.3 million SNPs, 2.7 million of which had been previously identified (dbSNP: www.ncbi.nlm.nih.gov/projects/SNP/), and estimate that the subject's genome actually contained a total of 3.7 million SNPs. While valuable as proof that a personal genome can be produced using NGS sequencing, this study lacked the overall coverage level to provide a comprehensive catalog of the variation within the Watson genome. Making use of mate-paired reads would have provided more accuracy in alignment and potentially higher coverage levels, allowing the study to capture more of the genomic variation.

*Three personal genomes released in November, 2008*

The November 6, 2008, issue of the journal Nature focused on personal genomics, and included three new reports of individual genomes, more than doubling the number previously published. Included were the diploid genomes of anonymous individuals: one male of African ancestry (Yoruba from Ibadan, Nigeria - NA18507)[49], one male of Asian descent (YH)[50], and the European female diagnosed with acute myeloid leukemia (AML)[51] (two genomes were produced for the AML patient, one from the leukemia and one from normal skin tissue). All carried out the DNA sequencing using the NGS technology platform that has been most widely used to date, the Illumina/Solexa Genome Analyzer (GA1) with reversible terminator chemistry[49]. They each achieved a significant reduction in per-genome cost compared with the Venter and Watson genomes, costing US $250,000, $500,000, and ~$800,000 respectively[18], and they all made large advances in the sequencing coverage achieved: 40.6x, 36x, and 32x (leukemia)/35x (skin) respectively. These advances were largely due to the massive parallelization of sequencing and processing of very short reads (32 to 35 bases). All of these studies assembled their genome(s) and performed variant discovery by aligning the reads to the NCBI36 reference.

The GA1 platform is introduced in the publication of the NA18507 genome[49], named for the HapMap sample acquired from the same anonymous individual[52,53]. Initially the investigators tested their sequencing, assembling, and variant calling strategy by focusing only on the X chromosome of the sample. This strategy allowed them to quickly test and validate various methods that were applied to the full genome. The genome was

sampled with ~4 billion paired-end 35-base pair reads which required 8 weeks to produce on six GA1 machines, and covered the genome at a level of 40.6x. This was the first personal genome that performed assembly using two competing algorithms: the short-read alignment programs ELAND [49] and MAQ [54]. MAQ slightly outperforms the ELAND aligner in terms of capturing more (99.29% vs. 99.24%) of 3.7 million validated SNPs from the HapMap project sample and having fewer genotype discrepancies (.88% vs. 1.20%) [49]. Both provided coverage for 99.9% of the reference genome. In total the number of SNPs reported using MAQ was 4.1 million while ELAND reported 3.8 million, both on the order of the number of SNPs reported in the Watson and Venter genomes.

The YH genome is the first diploid sequence of a member of an East Asian population that accounts for nearly 30% of all humans[50]. This study used a combined single-end, paired-end strategy for creating 35-base reads, producing two libraries totaling 3.9 billion reads. A third alignment program, Short Oligonucleotide Alignment Program (SOAP)[55], was used to place these reads on the NCBI36 reference, resulting in a 36-fold average coverage over 99.97% of the reference genome. There were 86.1% of these reads that aligned uniquely to the reference, and these were used to build the YH consensus sequence and detect genetic variations. That ~13% of the reads did not align to the reference raises the possibility of an incomplete human reference sequence, particularly one which was base on primarily individuals of European descent. The unmappable reads were assembled into contigs and compared with GenBank, as well as aligned to the chimpanzee genome. Roughly 50% of these contigs aligned well with unplaced human clones in

GenBank, and 5% aligned to the chimpanzee genome with greater than 90% identity. These regions may have been missed in the construction of the reference, or in fact represent a deletion in populations of European ancestry. This study employed a more stringent set of filtering criteria than the previous ones did which removed 8% of the reference sequence from their variant-discovery analysis, including among others a minimum of 4 overlapping reads and no evidence of flanking repetitive sequence. As a result, the number of identified variants was less than that of the other personal genomes: 3.07 million SNPs. While certainly reducing the number of false positive variant calls, the number of missed variants likely increased as well.

The final personal genome appearing in 2008 was the first such study specifically targeted at discovering disease causing mutations, illustrating the medical utility of whole genome resequencing. An acute myeloid leukemia genome was sequenced along with its normal counterpart obtained from the patient's skin was sequenced to 32.7x and 13.5x coverage respectively[51]. By creating and comparing two consensus sequences for this patient, the investigators were able to identify potential disease-related somatic mutations. The AML sample was covered by 2.7 billion reads, covering 91% of the genome when aligned to NCBI36 using MAQ, and the skin sample was covered by 1.1 billion reads, covering 80%. 3.8 million sequence variants were predicted in the leukemia genome and 2.36 million in the skin. Through several tiers of analysis and filtering, 8 nonsynonymous, high-quality somatic mutations were found to be unique to the leukemia genome, starting from an initial list of 31,623 leukemia-specific variants that had not been described before.

Two indels were also found in this study that were previously validated in other AML cases[51,56]. Through additional validation steps, the authors proceed to tie these mutations to genes supporting their claim that these may be 'driver' mutations in the pathenogenisis of AML[51].

## *Personal genomics in 2009*

By the spring of 2009, it was becoming more understood that the rapidly progressing high-throughput DNA sequencing technologies with their applications to human genomes were outpacing the current bioinformatics and annotation capabilities. While more studies like those from the previous year appeared, with billions of short reads produced cheaply and quickly, arguably much more effort was necessary for comparing, validating, and interpreting the results than in producing the reads themselves. In addition, a new generation of technologies began to immerge in the later half of the year including single-molecule sequencing (SMS), decreasing the sequencing costs by another order of magnitude, with the cheapest claiming a sequencing-consumables cost of $4,400 per genome. Two of the new personal genomes published in 2009 offer new views on two studies presented in 2008: a different technology was applied to the same NA18507 sample[57], and a second study of the AML genome was done by the same laboratory that produced the first. A second pair of personal genomes was reported based on a previously unrepresented socio-ethnic group: individuals of Korean descent.

The Applied Biosystems, Inc. (ABI) sequencing by oligo ligation detection (SOLiD) platform is in direct competition with the Illumina/Solexa GA, so it is fitting they were used to sequence the same individual, HapMap sample NA18507. The second whole-genome sequencing of NA18507 was released in September of 2009[57] with a reagent cost of up to US $60,000[18]. While only achieving producing ~18x of haploid coverage, an efficient use of mate-pairs allowed for an extensive interrogation of structural variation and the reconstruction of accurate haplotypes. The reads produced by SOLiD have a lower per-base error rate than other NGS technologies due to a two-base encoding scheme that self-corrects for errors[34]. This mechanism allowed the identification of variants with only two observed reads of each allele, rather than the three or four required in previous studies. Across the genome, 3.9 million SNPs were identified, 19% of which were novel. They presented a false-discovery rate analysis that indicated 99.88% of their SNP calls were accurate. Structural variants were identified ranging from single-base indels to deletions of up to ~97 kilobases, along with 91 inversions, 22 of which were also found in the Venter genome. Using mate-pair reads that covered at least two homozygous positions matching the reference allele, but were known to be variable in dbSNP, the genotypes at two-thirds of the loci were phased into haplotype "blocks" of sizes up to 215 kilobases. For haplotypes that were previously known for this individual from the HapMap data, the phases presented here were in accordance 98.95% of the time. Finally, the investigators studied the functional consequences of the cataloged variation. Their findings included five disease-relevant homozygous alleles, and 49 heterozygous sites using the Online

Mendelian Inheritance in Man database of gene-disease relationships[58,59]. In total, they located over 1500 putatively deleterious mutations and 2000 possible gene disruption events. Two interesting points were made using these results. First, an underrepresentation of SNPs was found in the coding regions for genes essential for cell survival; and second, an overrepresentation of SNPs occurred in genes undergoing rapid evolution in human populations. These genes include those involved with olfaction and immunity, whose variants have previously been described to differ in frequency between human populations, suggesting a relaxed mode of purifying selection, or even directional selection at these sites. These findings are also concordant with the population genetics prediction that the number of lethal mutations per individuals should be low[60]. Within this single genome, full genome resequencing was able to uncover signals of natural selection and human demography, indicating the utility of personal genomes in population genetics and personalized medicine.

In an effort to take full advantage of evolving NGS capabilities and bioinformatics methods, the same group that produced the original AML genome reported the full genome of a second AML case (AML-2) in September of 2009[61]. Using improved sequencing techniques the AML-2 genome was covered more completely (98% vs. 91%) than that of AML, along with a dramatically reduced cost in data production ($0.5 million vs. $1.6 million). As with the first case, individual genomes were created for samples derived from healthy skin tissue and from leukemia tissue. The focus of this study was on classifying potential cancer "driver" point mutations. An initial set of 3.8 million SNPs were identified

by MAQ in the leukemia genome relative to the NCBI36 reference, and 3.4 million of these passed the MAQ SNPfilter algorithm[61]. After comparing these to the SNPs found in the skin sample and previously described variable sites, 20,256 SNPs were identified to be most likely the result of leukemia-specific somatic mutation. Validation was performed on 395 SNPs that had the greatest potential for functional results, resulting in 62 confirmed somatic mutations. These were further analyzed in other 187 AML patients, and four were recurrent in at least one other AML sample. The authors argue that the likelihood of even 2 of 188 patients carrying the same mutation at the same mutation is small enough to warrant identification of these as contributing to pathogenesis. Therefore in this study, the authors have confirmed the potential of NGS platforms for uncovering the genetics of cancer at a dramatically reduced cost than previously reported.

The sequencing of the YH genome in 2008 produced a significant quantity of reads which could not be placed on previously reported genomes, including the NCBI36 reference, which was constructed primarily from European samples. This raises the question of what other human variation has been undiscovered in underrepresented socio-ethnic populations. The first full sequencing of a Korean genome, labeled SJK, was partially motivated by providing an answer to that question[62]. Koreans and Chinese are thought to have a common founder population and have been admixed for millennia, so an effort is being made to use elements from the YH genome and the SJK genome to construct a Korean reference genome (ftp://ftp.kobic.kr/pub/KOBIC-KoreanGenome/). The SJK consensus genome was built using MAQ[54] to align ~1.25 billion 36-base paired-end reads

and ~0.5 billion 75-base reads onto the NCBI36 reference, covering 99.9% of the reference with and average of 26x-fold coverage. Approximately 6% of the reads could not be aligned. While some of the unmappable reads are likely due to contamination or sequencing error, a portion of them could be used to construct 3,286 contigs that were homologous with unanchored scaffolds in the NCBI36 reference. This supported the finding of YH that the reference sequence was in fact not complete for all populations. The study identified 3.4 million SNPs, out of which 40% were novel. Paired-end reads were used to located structural variation and revealed 2,920 deletions and 415 inversion ranging from 0.1-100 kilobases in length, with 11.3% of structural variants identified as unique to SJK. The amount of novel variation located in SJK even when compared with the YH genome, suggest that the overall genetic differences among individuals from closely related populations may yet be significant, requiring further resequencing of minor groups to reveal the extent of human genetic variation.

To more fully address this question, a second whole-genome sequence of a Korean individual, labeled AK1, was also released in 2009. However, the notable distinction between this study and that of SJK is the level of redundancy that was used to fully characterize this individual. NGS technology in the form of Illumina/Solexa GA was used to cover the genome to an average depth of 27.8x-fold coverage using mate-pairs at various insert-sizes. To obtain haplotype sequencing, overlapping BAC clones were used to specifically target 390 genome-wide regions commonly affected by copy number variation to an impressive level of 151x-fold coverage. The entire chromosome 20 was also covered

in this manner, up to 155x, to estimate the overall sensitivity of their SNP and indel detection. Using the program GSNAP[63] 3.45 million SNPs were found by comparison with the NCBI36 reference; 17.1% of these were novel, a lesser number of novel SNPs than those found in SJK. In addition they identified 170,202 short indels (up to 29 bases), of which 62% were novel. Large structural variants were detected primarily using microarrays. For each class of variation the investigators trained quality filters and validated their results using a combination of microarray genotyping and Sanger resequencing. The validated variants were annotated to uncover their functional results primarily using an unpublished algorithm, Trait-o-matic, and 773 were found to be involved with a wide range of clinical phenotypes. Through the combined efforts of massively parallel sequencing, utilization of highly redundant BAC clones, multiple microarray genotyping, and Sanger validation sequencing, AK1 may be more substantially complete and accurate than any other personal genome to date. However this extensive treatment of a single genome is not routinely feasible and the finished product may still be incomplete[64].

### *The current generation of personal genomics*

Single-molecule sequencing (SMS) is a direct-DNA sequencing approach that does not rely on a PCR-amplification step. Various DNA molecules have variable success rates with PCR, which may introduce sequence changes as the DNA is replicated. By avoiding this step, SMS technologies are expected to lead to more accurate DNA sequencing,

especially when applied to DNA quantity-sensitive applications such as gene expression profiling[65]. The first commercially available SMS technology used to perform whole genome human sequencing is the Heliscope sequencer, sold by Helicos. The individual genome, sampled from a Caucasian male of European ancestry, was introduced in August 2009 [66]. This study achieved a 28x fold coverage of 90% of the NCBI36 reference using 2.7 billion 32-bp reads at a consumable-reagent cost of US $48,000[18]. The read alignment to the reference genome was performed with a new, unpublished software algorithm, IndexDP, which was developed to out-perform other short-read aligners on data generated by the Heliscope sequencer. A separate program, called UMKA, was developed to call the variant bases by using alignment quality scores, accounting for the Heliscope sequencing error profile, and incorporating a naïve prior probability distribution on the distribution of variation throughout the human genome. Using the alignments produced by IndexDP and the variant calling capability of UMKA, ~2.8 million SNPs were identified in the personal genome. The SNP calls were validated by comparing the genotyping results of an Illumina Human610-Quad SNP BeadArray assay, resulting in a100% coverage of the sites genotyped with 99.8% concordance. Seventy-six percent of the identified SNPs were listed as validated in dbSNP. Assuming that CNV dominates the structural variant class of human polymorphism, the authors specifically target CNVs by examining changing read densities over the genome. With this method 752 CNV regions were predicted, and validation was successful on 25 of 27 regions selected for testing. While this particular personal genome was not presented as fully described and validated as some produced by

NGS technologies, this study servers as an effective proof of principle that SMS technologies can currently be applied to the field of personal genomics.

The most recent technology-introducing publication of a personal genome, produced by Complete Genomics (http://www.completegenomics.com/), included the independent sequencing and assembly of three unrelated individuals[14]. This impressive report claims that each individual genome was produced at an average cost of US$4,400 for sequencing consumables, a representation of an order-of-magnitude reduction from the personal genomes published most recently[18]. The individuals sequenced included a Caucasian male of European descent (NA07022), a Yoruban female (NA19240), and a second Caucasian male (NA20431). The first two were originally characterized by the HapMap project[53], and the third is a member of the Personal Genome Project[67]. These individuals were chosen to facilitate validation of the results by comparison to previously reported findings. A custom alignment algorithm was used to align the generated mate-pair reads from NA20451, NA19240, and NA07022 to the NCBI36 reference, resulting in an overall genome coverage of 45x, 63x, and 87x, respectively. In summary, the respective genetic variation detected in each genome was 2.90 million, 4.04 million, and 3.07 million SNPs along with 270 thousand, 469 thousand, and 338 thousand indels. The NA19240 sample contained the highest amount of novel variation, with 19% of the SNPs and 42% of the indels being previously undescribed. Validation was carried out primarily on the NA07022 data set by comparing the called SNPs with the HapMap genotypes reported, revealing a 99.15% concordance with the 94% of the HapMap genotypes having coverage.

Targeted sequencing was performed on a random subset of novel nonsynonymous variants, and a false positive rate of 1 false variant per 100,000 bases was extrapolated from the analysis. This variant call accuracy compares well with the other reported human genome sequences, and has again been achieved at a fraction of the cost needed in the previous technology generation[14,68].

## Conclusion – Lingering Challenges in Human Genomics

Across the published personal human genomes several common themes emerge, highlighting the ways in which the practice of whole genome sequencing must improve. These advances are necessary before the extent of information contained within each human genome can be comprehensively understood or the maximal medical utility can be achieved through personalized medicine. Assuming the former is a necessary prerequisite to the latter, motivation in this regard should be high given the public interest and growing market for 'consumer genetics.' A key element to both goals will be the population-specific characterization of rare variants that may be the key to uncovering the genetic basis of common, complex human diseases. The 1,000 Genomes Project and the Personal Genome Project (PGP) are important efforts to this end. However, as progress is made, a priority shift may be necessary from cheaply and quickly producing billions of minimally informative reads to developing, perfecting, and standardizing methods for sequencing analysis and linking a myriad of genetic variations to their diverse phenotypic results.

As illustrated by the individual genomes presented, there has been a tremendous advancement in base sequencing technology since the first individual genomes were published: in two years there has been two orders of magnitude reduction in reagent cost for the production of a diploid human genome sequence[14,18]. It is frequently noted that the bioinformatics methods for comprehensively extracting useful biological information has not kept pace[69]. Indeed, it is possible that the manner in which sequencing technologies have advanced has made the bioinformatics problems more difficult. By initially focusing on the massively parallel production of short reads sized to a few tens of bases, the challenge of recreating the genome has been exacerbated. Sequencing the reads in pairs, with clearly defined distances separating their original genomic locations, has been an improvement, but continues to fall short of the lengths required for complex genomes[22]. Complete, accurate reconstruction of an original diploid sequence with extensive structural variations orders of magnitude larger in scale than the length of a read may be an unsolvable task. Although methods have been presented that offer solutions, utilizing a de Bruijn graph data structure[70-72], they are all limited by the unmet requirement of at least a proportion of read pairs covering more bases than the longest common near-identical repeat in the human genome[73]. Using a reference assembly as a guide has substituted for this full, *de novo* reconstruction, a practice that has added a level of uncertainty and error to the process. Ideally read lengths would be on the order of millions of bases, up to the length of a full chromosome, and technologies reaching toward this goal will likely contribute more

to the study of human genetics than ones which instead focus on reducing the cost of producing reads of a few hundred bases in length.

Until such a technology exists, investigators must continue making the most of the streams of reads flowing unabated from the latest generation of sequencers. As suggested above, the assembly task becomes of primary importance. Currently, a full *de novo* assembly is not a viable solution for mammalian genomes since the de Bruijn graph strategies alluded to also are hindered by the amount of real memory required, necessitating several terabytes of memory to process a complex genome[73]. Thus, with the exception of the Venter genome, every personal genome was produced by aligning the reads to the NCBI36 reference genome, relying on a variety of software packages. MAQ[54], which was shown in the original presentation of the NA18507 genome to perform slightly better than the Illumina software ELAND[49], emerged as a front-runner in alignment of short reads to the reference genome: four of the ten such studies presented used MAQ[51,61,62,74]. Studies that did not use MAQ typically make use of a short read aligner custom built to manage the specific error profile of the sequencing technology in use: the ABI presentation of the NA18507 genome used Corona-lite[57], the genome sequenced by the Heliscope SMS platform was formed with IndexDP[66], and all three of the genomes presented by Complete Genomics were constructed with an unpublished method highly dependant on their method of read production[14]. The remaining programs used included BLAT[75], SOAP[55], ELAND, and GSNAP[76]. MAQ, SOAP, ELAND, and IndexDP make use of the same type of alignment strategy known as hash-base alignment. Using the concept of 'spaced seeds'

popularized for sequence alignment by the PatternHunter[77] program, these programs compress either the set of reads or the reference sequence into a specialized data structure, known as a 'hash table', that is designed to index complex and non-sequential data, allowing for rapid searching[73]. Given that DNA sequencing reads are most likely to contain duplicated segments rather than every possible combination of nucleotides, the hash-indexing strategy is an appropriate one. Following the construction of the hash table, the exact placement of each read is determined by specialized alignment algorithms that make use of gapped and ungapped versions of the Smith-Waterman dynamic programming algorithm[78] while taking advantage of the quality values of the sequenced bases. A competing set of programs, including BWA[79], BOWTIE [80], and SOAP2 [81] for aligning short reads to a reference genome has appeared that make use of the Burrows-Wheeler transform (BWT) method of creating a suffix tree[82-84] to compress the sequence data in such a way that placing reads on the genome is completed quite rapidly[73]. These methods are several times faster than their hash-based counterparts while admitting very slight degradation in accuracy. It is possible these programs will become commonly utilized as WGS reads are produced with increasing swiftness. Further reviews and comparisons are available for software implementations of both hash-based and BWT-based alignment strategies[65,73,85].

The NCBI36 reference genome played a significant role in each personal genome produced. Most of the genetic variation reported is given in the context of this DNA sequence. This has led to two major sources of difficulty with regards to fully

characterizing the genetic variants found in each individual genome. The first is the incompleteness of the reference itself. Due to certain regions being refractory to Sanger sequencing, 341 gaps exist in the NCBI36 reference genome, 250 of which are located in euchromatic regions, likely to contain actively expressed genes[13]. Secondly, the reference sequence has been primarily derived from individuals of European origin, potentially leaving regions unique to other populations unrepresented. The YH[50], AK1[76], and SJK[62] studies each contained a substantial amount of sequence data that could not be mapped to the NCBI36 reference. The investigators in the SJK report nearly 6% of the reads could not be aligned indicating the need for a comprehensive reference unique to minor socio-ethnic groups in the future of massive individual genome sequencing[62].

Making meaningful comparisons between the reported personal genomes is complicated by the variety of methods of sample preparation, sequencing, assembly, and quality filtering of the data used in each. Where comparisons have been made, in making inferences of genetic ancestry between the samples for example, it is encouraging to find the results are in concordance with expectations. The NA18507 genome falls in line with other HapMap African samples, the Watson and Venter genomes cluster with samples of European origin, and the YH, AK1, and SJK samples are affiliated with population samples of East Asia [64]. Exemplary of the complications presented by making these comparisons across methodologies for variant calling, the Watson genome shows a significant correlation with African ancestry, while there has been no previously published indication that Watson's lineage contains African admixture[64]. Whether this is a novel finding or is

simply an artifact of the low sequencing coverage and poorer sequence quality of the Watson genome remains an unanswered question. It is notable that these conclusions could have been drawn from data produced by genotyping, much more quickly and cheaply than the cost of producing the complete genome. However what should not be overlooked is the contribution of cataloging rare variations, for example those which would distinguish the SJK from the YH genomes. These may be useful in the future for making ancestral inferences on a finer scale than what is currently possible, but many more personal genomes will have to be produced before these comparisons will be fully beneficial.

Two ongoing projects promise to contribute greatly to achieving a more complete catalog of human variation, common and rare: the 1000 Genomes Project (1KGP) (www.1000genomes.org) and the Personal Genome Project (PGP) (www.personalgenomes.org). Announced in 2008, and currently proceeding into its primary project phase after completing three pilot investigations, the 1KGP has the specific goal of characterizing all human genetic variation which occur with frequencies greater than 1% in non-protein coding regions, and greater than .1% in coding regions[13]. The individuals sequenced will include those from diverse populations which have so far been under represented in population-based whole-genome investigations including Gujarati Indians from Houston, Toscani from Italy, Mexican-Americans from Los Angeles, Chinese from Denver and Beijing, Africans from Nigeria and Kenya, and African-Americans from the southwestern USA[86]. Perhaps the greatest utility of the data sequenced in this project will be richer, more diverse human reference sequence containing population-specific

motifs and genetic landscapes, as compared to the relatively narrow scope of the current standard reference sequence when viewed in light of the enormous amounts of unrepresented human diversity. One facet of the 1KGP limiting its utility however is the lack of phenotype information captured for the samples. Without such information, making ties between newly discovered rare variation and phenotypic differences including susceptibility to disease will be difficult. However, the PGP aims to excel in this regard by linking extensive phenotypic information, including familial disease history as well as facial photographs, with the collected DNA. There are ethical, legal, and social issues (often termed ELSI) associated with either strategy that have been well discussed, and legislation has already begun to appear in the United States governing the use of such personal data[87,88]. First introduced by George Church in 2005, the PGP is a public, volunteer-base program that aims to fully sequence over 100,000 individuals over the course of 25 years[67]. With currently fewer than 100 enrollees, as of this writing no published results have been generated from the PGP, however the possibilities towards uncovering novel phenotype-genotype correlations are tremendous.

# References


1. Chakravarti, A. Population genetics--making sense out of sequence. in *Nat Genet* Vol. 21 56-60 (1999).

2. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R. & Amos, C.I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. in *Am J Hum Genet* Vol. 82 100-12 (2008).

3. Schork, N.J., Murray, S.S., Frazer, K.A. & Topol, E.J. Common vs. rare allele hypotheses for complex diseases. in *Current Opinion in Genetics & Development* Vol. 19 212-9 (2009).

4. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. in *Nat Rev Genet* Vol. 10 241-51 (2009).

5. Lander, E.S. The new genomics: global views of biology. in *Science* Vol. 274 536-9 (1996).

6. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. in *Trends Genet* Vol. 17 502-10 (2001).

7. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? in *Am J Hum Genet* Vol. 69 124-37 (2001).

8. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. in *Nat Genet* Vol. 40 695-701 (2008).

9. Consortium, W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. in *Nature* Vol. 447 661-78 (2007).

10. Manolio, T.A., Brooks, L.D. & Collins, F.S. A HapMap harvest of insights into the genetics of common disease. in *J Clin Invest* Vol. 118 1590-605 (2008).

11. Manolio, T.A. et al. Finding the missing heritability of complex diseases. in *Nature* Vol. 461 747-53 (2009).

12. Levy, S. et al. The diploid genome sequence of an individual human. in *Plos Biol* Vol. 5 e254 (2007).

13. Mir, K.U. Sequencing genomes: from individuals to populations. in *Brief Funct Genomic Proteomic* Vol. 8 367-78 (2009).

14. Drmanac, R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. in *Science* Vol. 327 78-81 (2010).

15. Tjio, J.H. & Levan, A. The Chromosome Number of Man. in *Hereditas* Vol. 42 1-8 (1956).

16. Painter, T.S. The Sex Chromosomes of Man. in *The American Naturalist* Vol. 58 506-524 (1924).

17. Kimura, M. The neutral theory of molecular evolution. in *Sci Am* Vol. 241 98-100, 102, 108 passim (1979).

18. Metzker, M.L. Sequencing technologies - the next generation. in *Nat Rev Genet* Vol. 11 31-46 (2010).

19. Hubby, J.L. & Lewontin, R.C. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. in *Genetics* Vol. 54 577-94 (1966).

20. Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. in *Nature* Vol. 304 412-7 (1983).

21. Metzker, M.L. Emerging technologies in DNA sequencing. in *Genome Research* Vol. 15 1767-76 (2005).

22. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. in *Nat Meth* Vol. 6 S13-20 (2009).

23. Macaya, D. et al. A synonymous mutation inTCOF1causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. *Am J Med Genet A* **149A**, 1624-1627 (2009).

24. Marshall, M., Solomon, S. & Lawrence Wickerham, D. Case report:de novo BRCA2gene mutation in a 35-year-old woman with breast cancer. *Clin Genet*, 1-4 (2009).

25. Gauthier, J. et al. Novel de novo SHANK3 mutation in autistic patients. *Am. J. Med. Genet.* **150B**, 421-424 (2009).

26. Khor, C.C. et al. A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. in *Nat Genet* Vol. 39 523-8 (2007).

27. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. in *Proc Natl Acad Sci USA* Vol. 74 5463-7 (1977).

28. Lander, E.S. et al. Initial sequencing and analysis of the human genome. in *Nature* Vol. 409 860-921 (2001).

29. Venter, J.C. et al. The sequence of the human genome. in *Science* Vol. 291 1304-51 (2001).

30. Shendure, J. & Ji, H. Next-generation DNA sequencing. in *Nature Biotechnology* Vol. 26 1135-45 (2008).

31. Ng, S.B. et al. Exome sequencing identifies the cause of a mendelian disorder. in *Nat Genet* Vol. 42 30-5 (2010).

32. Haaland, W.C. et al. A-beta-subtype of ketosis-prone diabetes is not predominantly a monogenic diabetic syndrome. in *Diabetes Care* Vol. 32 873-7 (2009).

33. Fox, S., Filichkin, S. & Mockler, T.C. Applications of ultra-high-throughput sequencing. in *Methods Mol Biol* Vol. 553 79-108 (2009).

34. Mardis, E.R. The impact of next-generation sequencing technology on genetics. in *Trends Genet* Vol. 24 133-41 (2008).

35. von Bubnoff, A. Next-generation sequencing: the race is on. in *Cell* Vol. 132 721-3 (2008).

36. Ansorge, W.J. Next-generation DNA sequencing techniques. in *N Biotechnol* Vol. 25 195-203 (2009).

37. Liu, G.E. Applications and Case Studies of the Next-Generation Sequencing Technologies in Food, Nutrition and Agriculture. in *Recent Patents on Food, Nutrition & Agriculture* Vol. 1 75-79 (2009).

38. Pettersson, E., Lundeberg, J. & Ahmadian, A. Generations of sequencing technologies. in *Genomics* Vol. 93 105-11 (2009).

39. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. in *Nature* Vol. 452 872-6 (2008).

40. Consortium, I.H.G.S. Finishing the euchromatic sequence of the human genome. in *Nature* Vol. 431 931-45 (2004).

41. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. in *Nat Rev Genet* Vol. 7 85-97 (2006).

42. She, X. et al. Shotgun sequence assembly and recent segmental duplications within the human genome. in *Nature* Vol. 431 927-30 (2004).

43. Freeman, J.L. et al. Copy number variation: new insights in genome diversity. in *Genome Research* Vol. 16 949-61 (2006).

44. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. in *Annual review of genomics and human genetics* Vol. 7 407-42 (2006).

45. Stephens, J.C. et al. Haplotype variation and linkage disequilibrium in 313 human genes. in *Science* Vol. 293 489-93 (2001).

46. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. in *Nat Genet* Vol. 29 229-32 (2001).

47. Watson, J.D. & Crick, F.H.C. A Structure for Deoxyribose Nucleic Acid. in *Nature* Vol. 3 737-738 (1953).

48. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. in *Nature* Vol. 437 376-80 (2005).

49. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. in *Nature* Vol. 456 53-9 (2008).

50. Wang, J. et al. The diploid genome sequence of an Asian individual. in *Nature* Vol. 456 60-65 (2008).

51. Ley, T.J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. in *Nature* Vol. 456 66-72 (2008).

52. Consortium, I.H. A haplotype map of the human genome. in *Nature* Vol. 437 1299-320 (2005).

53. Consortium, I.H. et al. A second generation human haplotype map of over 3.1 million SNPs. in *Nature* Vol. 449 851-61 (2007).

54. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. in *Genome Research* (2008).

55. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. in *Bioinformatics* Vol. 24 713-4 (2008).

56. Link, D.C. et al. Distinct patterns of mutations occurring in de novo AML versus AML arising in the setting of severe congenital neutropenia. in *Blood* Vol. 110 1648-55 (2007).

57. McKernan, K.J. et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. in *Genome Research* Vol. 19 1527-41 (2009).

58. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. in *Nucleic Acids Research* Vol. 33 D514-7 (2005).

59. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. in *Am J Hum Genet* Vol. 80 588-604 (2007).

60. Morton, N.E., Crow, J.F. & Muller, H.J. AN ESTIMATE OF THE MUTATIONAL DAMAGE IN MAN FROM DATA ON CONSANGUINEOUS MARRIAGES. in *Proc Natl Acad Sci USA* Vol. 42 855-63 (1956).

61. Mardis, E.R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. in *N Engl J Med* Vol. 361 1058-66 (2009).

62. Ahn, S.-M. et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. in *Genome Research* Vol. 19 1622-9 (2009).

63. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. in *Bioinformatics* Vol. 21 1859-75 (2005).

64. Yngvadottir, B., Macarthur, D., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. in *Genome Biol* Vol. 10 237 (2009).

65. Dalca, A.V. & Brudno, M. Genome variation discovery with high-throughput sequencing data. in *Briefings in Bioinformatics* (2010).

66. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. in *Nature Biotechnology* Vol. 27 847-52 (2009).

67. Church, G.M. The personal genome project. in *Mol Syst Biol* Vol. 1 2005.0030 (2005).

68. Porreca, G.J. Genome sequencing on nanoballs. in *Nature Biotechnology* Vol. 28 43-4 (2010).

69. Mcpherson, J.D. Next-generation gap. in *Nat Meth* Vol. 6 S2-S5 (2009).

70. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. in *Genome Research* Vol. 17 1697-706 (2007).

71. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. in *Genome Research* Vol. 18 821-9 (2008).

72. Simpson, J.T. et al. ABySS: a parallel assembler for short read sequence data. in *Genome Research* Vol. 19 1117-23 (2009).

73. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. in *Nat Meth* Vol. 6 S6-S12 (2009).

74. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).

75. Kent, W.J. BLAT--the BLAST-like alignment tool. in *Genome Research* Vol. 12 656-64 (2002).

76. Kim, J.-I. et al. A highly annotated whole-genome sequence of a Korean individual. in *Nature* Vol. 460 1011-5 (2009).

77. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. in *Bioinformatics* Vol. 18 440-5 (2002).

78. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. in *J Mol Biol* Vol. 147 195-7 (1981).

79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. in *Bioinformatics* Vol. 25 1754-60 (2009).

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

81. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. in *Bioinformatics* Vol. 25 1966-1967 (2009).

82. Burrows, M. & Wheeler, D.J. A Block-sorting Lossless Data Compression Algorithm. in *Digital Equipment Corporation* Vol. Technical Report 124 1-24 (1994).

83. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *ANNUAL SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE* Vol. 41 390-398 (2000).

84. Lam, T.W., Sung, W.K., Tam, S.L., Wong, C.K. & Yiu, S.M. Compressed indexing and local alignment of DNA. in *Bioinformatics* Vol. 24 791-7 (2008).

85. Voelkerding, K.V., Dames, S.A. & Durtschi, J.D. Next-Generation Sequencing: From Basic Research to Diagnostics. in *Clinical Chemistry* Vol. 55 641-658 (2009).

86. Need, A.C. & Goldstein, D.B. Next generation disparities in human genomics: concerns and remedies. in *Trends Genet* Vol. 25 489-94 (2009).

87. Lunshorf, E.R.M.J.E. A focus on personal genomics. in *Personalized Medicine* Vol. 6 603-606 (2009).

88. McGuire, A.L. 1000 Genomes: on the road to personalized medicine. in *Personalized Medicine* Vol. 5 195-197 (2008).

# Chapter II

# Probabilistic Inference of Pedigree Genotypes using High-Throughput Sequencing*

* with major contributions on model and method development by Reed Cartwright, Ph.D.

# Introduction

As the characterization of genetic variation is a primary goal of genetics, an important aspect of that pursuit is studying the mutational origin of DNA sequence variants and quantifying the rate at which they occur. Motivation for understanding the nature of spontaneous, *de novo*, mutation is especially high in human medical genetics where they can contribute to disease and disease susceptibility. There is increasing evidence that the interaction of *de novo* mutation with segregating background variation is a significant contributor to complex diseases such as Autism[1], Schizophrenia[2], Breast Cancer[3], Diabetes[4], Treacher Collins Syndrome[5], Darier Disease[6], and Beals syndrome[7]. These cases highlight a few novel mutations that have been linked to complex disease susceptibility, but likely represent only the crest of the mountainous supply of functional genetic variation beneath the surface of common, complex disease phenotypes. With high throughput sequencing, a new type of case-control study has become feasible; complete genome sequencing of a well-defined pedigree and somatic tissues provides an unbiased survey for screening affected individuals for spontaneous mutations. Multiple methods have been developed for detecting *de novo* mutations using whole-genome sequences of a family cohort, each based on the detection of Mendelian errors. Here a novel approach is presented involving a probabilistic framework that uses the relatedness between the individuals to produce a joint probability for the entire pedigree at each site, taking full advantage of the information contained within the pedigree. The utility of this methodology is two-fold: every genomic site investigated is ranked according to a posterior

probability of carrying a spontaneous mutation, and a direct estimate is produced of the spontaneous mutation rate, as well as other important parameters such as the error rate of the sequencing technology and the population mutation rate for the population from which the samples were drawn.

The estimation of spontaneous mutation rate has been an active area of inquiry since the inception of modern mathematical formulations of genetics. Haldane (1935) estimated the appearance of a 'human gene for hemophilia' to occur once in every 50,000 human generations based on existing medical records of disease incidence. Since then, the methods for making empirical estimates in humans have proceeded in one of two forms: direct and indirect. The direct approach is somewhat similar to Haldane's original method in that specific reporter loci at which mutations have drastic phenotypic changes are queried for mutations at the sequence level[8-13]. Sperm-typing these candidate loci can be performed using a variety of technologies to ensure the mutations in question occur within germline tissue. Calculations involving the observed per-locus rate of such a mutation and the number of nucleotide sites where a mutation would lead to the abnormal phenotype can be used to generate a direct estimate assuming many different such mutant alleles have been sequenced[8]. These studies suffer from making assumptions on the fraction of mutations that have observable phenotypic effects, and the estimates put forward may be biased towards the genomic context of the loci tested. It is well established that the mutation rate can vary greatly throughout the genome[14-16] and that failing to take the variation into account can have drastic effects on down stream analyses[17]. If the direct-

estimation methods approach the problem on a short time scale, then indirect methods consider much longer time scales; these studies compare the sequences data from closely related species[18-22]. The primary drawbacks with these methods are the reliance upon estimates for species divergence time, for generation times, and for the effective population size of the common ancestral species. In addition they can produce underestimates of the actual per-generation mutation rate since, in theory, many of the mutations that occurred over the time of divergence were deleterious, and were removed by natural selection. Only recently with high-throughput sequencing have direct methods been applied to whole-genome analysis. The first such study was applied to a classic model organism, the yeast *Saccharomyces cerevisiae* [23], and the first human application relied on a scan of the Y Chromosomes taken from males separated by 13 generations[24]. By applying high-throughput sequencing to members of a nuclear family with accurate pedigree information, the stage is set to make the most direct estimates of the human mutation rate to date.

Spontaneous mutations can occur any time a cell divides. Beginning from a single-cell zygote, a human will develop into a multicellular organism consisting of from ten to 100 trillion cells, after at least the same number of cell divisions, creating ample opportunity for mutation to occur within any particular tissue. Cells involved in gametogenesis, as well as the resulting ova and sperm sex cells, are referred to as germ cells while cells comprising the remainder of body tissues are somatic cells. Mutations occurring in the formation of these cell types are correspondingly labeled as germline mutations or somatic mutations. Both somatic and germline mutations carry

epidemiological significance, yet only mutations in the germline are passed to subsequent generations. Therefore the per-generation mutation rate of humans is specifically limited to germline mutations. Germline mutations arguably play a more significant role in human evolution by providing the raw genetic variation upon which natural selection can act from generation to generation. Because DNA samples are often obtained from somatic tissues, such as blood or skin, the possibility exists of mutations unique to the somatic tissue within individual sampled, and may be different than the zygotic genotype the individual inherited from his or her parents or the germline genotype he or she may pass to the following generation. The number of somatic mutations could be quite large due to the number of cell divisions that occur, and attempts must be made to distinguish germline mutations from somatic mosaicism[25-27]. Any empirical germline mutation rate estimation method based on resequenced somatic cells should take into account somatic sources of variation. It will be shown that a single sample from each individual in the simplest nuclear family (two parents, one child) is insufficient to disentangle somatic and germline mutations. However, experimental validation has become cost effective to the point of assaying many more candidate mutation sites than would be expected by current estimations of the expected mutation rate, allowing for locus-specific verification of predicted *de novo* events. The framework we present here can be used to disentangle the effects of these mutational processes in larger family designs, which will make possible accurate germline mutation rate estimates without the extra commitment of resources to experimentally validate putative *de novo* mutations. In addition to the rate of germline and somatic mutation, we

simultaneously estimate the effect of sequencing error and the initial genetic variation in the population from which the parents arise. By having good estimates of these rates, performing scans for actual spontaneous mutations will be made easier by providing some expectation for the number of mutations.

## Methods

Detecting Mendelian errors is the primary way by which we discover new mutations in resequenced nuclear families. A Mendel error is said to have occurred when an offspring has a genotype that is incompatible with any possible pattern of inheritance taken from the two parent genotypes. For example, if both parents have genotypes AA, and the child is heterozygous AC, then a Mendel error has occurred since neither parent carries the 'C' allele, with the implication that a new A to C mutation has occurred. In an effort to more closely model the true biology of DNA sequence, we do not assume an infinite number of mutable sites, so mutations are possible at polymorphic sites. Some mutations may occur where the mutant allele will be consistent with Mendelian inheritance, and so will be undetectable. Without making use of the inheritance pattern of linked sites, these mutations will go undetected. In these rare cases a mutation rate based on occurrence of *de novo* mutations has the potential of being slightly. Aside from these recurrent mutations, it is clear when a *de novo* mutation occurs given perfect knowledge of the family members' genotypes and inheritance pattern. However, several levels of uncertainty make comprehension of the pedigree challenging. The primary sources of uncertainty are: (1)

inadequate sampling of both chromosomes to detect a heterozygous genotype in all individuals, (2) somatic or cell line mutations which are distinguished from *de novo* events in that they do not occur within the germline passed from one generation to the next, (3) sequencing error, and (4) errors during read assembly as mentioned in Chapter 1.

Detecting heterozygous genotypes is imperative to this discovery process. Missing heterozygous sites in a child will contribute to false negative mutational calls while an undetected heterozygous parent can lead to false positives. In order to have any high degree of certainty in calling a site a mutation or a non-mutation, it is necessary to have a sufficient level of sequencing coverage for each individual in order to guarantee sampling of both chromosomes at a diploid site. At low coverage levels, an under-sampled chromosome bearing a true minor allele can resemble sequencing error or mapping error. In order to make the most accurate genotyping calls possible, the method presented leverages the relatedness between individuals in order to take full advantage of the information contained within the pedigree. At each site the probability of the pedigree is calculated rather than for individual genotypes. This is the primary way in which this method for spontaneous mutation discovery differs from other procedures: typically the genotypes for individuals are determined independently before searching for Mendelian errors.

As an illustration, consider the high-throughput sequencing data presented in Table 2-1. In high-throughput sequencing multiple reads are expected to sample any given

genomic position of interest, so the number of observed nucleotides of each type among the set of reads covering the site for each individual can be summarized as shown.

**Table 2-1.** The data structure used to represent the observed nucleotides at a site for all three members of a trio nuclear family. Aligned high-throughput sequencing reads are transformed into this format as input into the *de novo* prediction algorithm.

| Observed Nucleotide at a Single Site | Adenine (A) | Cytosine (C) | Guanine (G) | Thymine (T) |
|---|---|---|---|---|
| Frequency in the Mother | 0 | 0 | 0 | 9 |
| Frequency in the Father | 0 | 0 | 0 | 15 |
| Frequency in the Child | 1 | 12 | 0 | 18 |

Considered individually, the most parsimonious genotype call for the mother is homozygous-T, for the father homozygous-T, and heterozygous CT for the child. However, when considering the possibility that the mother is also heterozygous CT, the probability of observing 9 reads bearing a Thymine nucleotide and none which sampled the chromosome bearing the Cytosine must be compared to the probability of mutation. Assuming any read can independently sample either chromosome at a site with equal chance, this question is analogous to tossing a coin 9 times and the coin never landing on 'tails', according to binomial probability. This event is calculated to occur with the approximate frequency of 2 in one thousand, $2 \times 10^{-3}$. A spontaneous Adenine-to-Cytosine mutation is the best explanation for the heterozygosity of the child if the mother is truly homozygous. However, previous estimates of the per-site frequency of spontaneous

mutation are on the order of one to three in ten million, $1-3\times10^{-8}$ [8,18,28,29]. By several orders of magnitude, it is much more likely that unequal sampling of a truly heterozygous mother produced the observed data rather than a mutation event. A lack of coverage in the mother for this site can lead to a false positive mutation call when independently determining the genotypes. The method presented here, by considering the data observed in the child when determining the most likely genotype of the mother, avoids this source of error in an automated calculation and infers the presence of an unobserved allele in low coverage situations to prevent false positive mutation calls.

## Model

A probabilistic model was constructed to account for the uncertainty and error in the process of *de novo* mutation discovery. As mentioned above, the method makes use of the relatedness between the individuals and produces posterior probabilities of *pedigrees* at each site rather than posterior probabilities on individual *genotypes* to facilitate correctly determining the family members' genotypes. While the data for a single site is considered jointly among the family members, within a single individual the data for each site is treated independently. This simplifies calculations without sacrificing much accuracy in terms of modeling mutation patterns. Polymorphic sites will be approximately independent from one another since closely linked double heterozygotes will be rare when the per-site diversity level is low. In fact, the vast majority of sites are expected to be non-polymorphic, in which case linkage between sites has little to no effect.

Conceptually the data in our model belongs to one of two categories. The *observed data*, $R$, consists of aligned sequence reads from each individual. The *hidden data*, $H$, from which the observed data is derived, is comprised of the actual parental and offspring genotypes, the pattern of inheritance, somatic and germ-line mutation events, how the chromosomes are sampled by the sequencing reads, and any sequencing error events. Thus perfect knowledge of the *hidden data* makes discovering *de novo* mutations a simple task, and the rate at which they occur is then trivially estimated. I now turn to the mathematical formulation of the model. Given the independence of sites, the total probability of all of the data is the product of the probabilities at each site:

$$P_T(R,H \mid \Theta) = \prod_{s=1}^{N_s} P_S(R_s,H_s \mid \Theta) \quad (1)$$

where $R_s$ and $H_s$ are the observed and hidden data associated with site $s$, and $\Theta$ contains the parameters of our model: the population per-site diversity parameter $\theta$, the per-site per-generation germline mutation rate $\mu$, the per-site somatic mutation rate $\mu_S$, and the per-site sequencing error rate $\varepsilon$. From here the explanation will focus on a single site, so the subscript $s$ will be suppressed for clarity. For a single family including two parents and one offspring (hereafter referred to as a trio), the relationship between these parameters and the individual genotypes is given in Figure 2-1. Although the model can be extended to more complex pedigrees, a simple trio pedigree will be used initially for illustration. Starting from the root of the pedigree, the parental alleles $m_a$, $m_b$, $f_a$, and $f_b$ are sampled from a population at equilibrium allowing up to three segregating alleles. The distribution

of these alleles is calculated in a coalescent framework utilizing $\theta$ and allowing for at most two mutations on the coalescent genealogy. These four alleles are the founders of the pedigree and form the original zygotic or germline genotypes of the two parents. From this sample of alleles, one is transferred from each parent to the offspring, *m** and *f**, with the possibility of germline mutation at rate $\mu$, to form the zygotic genotype $o_a o_b$. The allele from chromosome *a* in the offspring is arbitrarily labeled as the allele inherited from the mother (*m**). Throughout development and stem cell replenishment, the original zygotic genotypes for each individual are independently passed through a somatic cell lineage to the tissue sampled for sequencing. Over the course of these cell lineages somatic mutations accrue at the per-nucleotide (not per cell-division) rate of $\mu_S$ to form the three somatic genotypes *m'*, *f'*, and *o'*. The collections of reads that sample the somatic genotypes at this site are contained within $R_M$, $R_F$, and $R_O$ with a sequencing error rate of $\varepsilon$ per base sequenced, constituting the observed data such that $R = \{R_M, R_F, R_O\}$. The final piece of hidden data which is not shown in this figure is the set of indicator variables, $S = \{S_M, S_F, S_O\}$, corresponding to the chromosome each particular read samples. There are $N_{RM}$ reads in total sampling the mother at this site, $N_{RF}$ reads from the father, and $N_{RO}$ offspring reads, so *R* and *S* are fully partitioned into:

$$\begin{aligned} R &= \left\{ R_I = \left\{ R_{I1}, R_{I2}, \ldots, R_{Ik}, \ldots, R_{IN_{RI}} \right\}, R_{Ik} \in \{A,C,G,T\}, I = M,F,O \right\} \\ S &= \left\{ S_I = \left\{ S_{I1}, S_{I2}, \ldots, S_{Ik}, \ldots, S_{IN_{RI}} \right\}, S_{Ik} \in \{a,b\}, I = M,F,O \right\} \end{aligned} \tag{2}$$

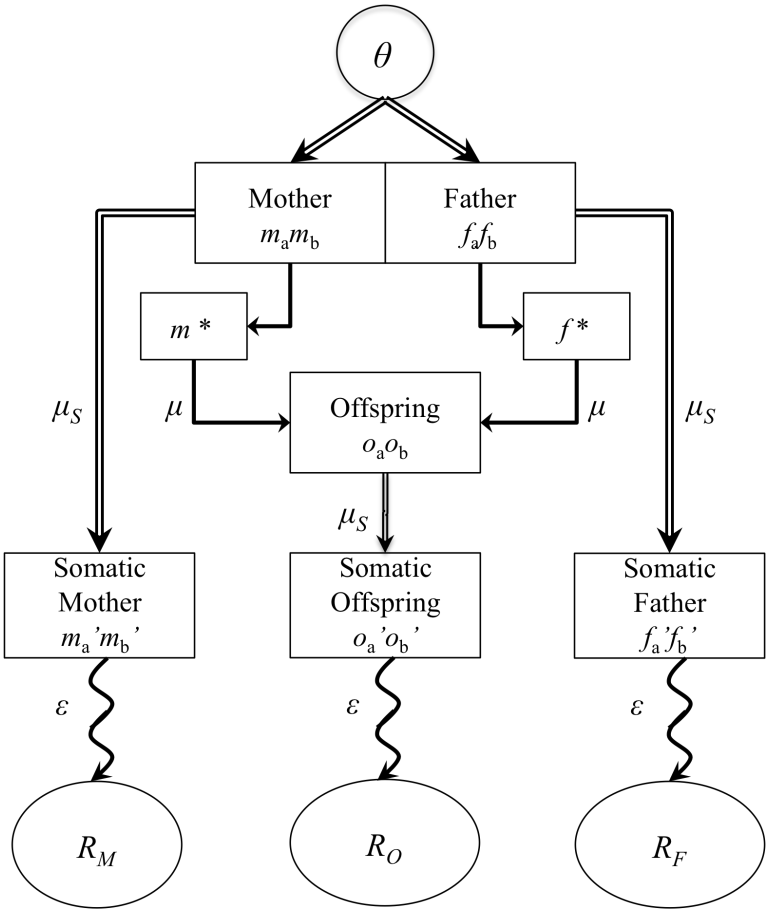



**Figure 2-1.** Trio model for a single site. Each straight line in the figure represents the transmission of a single allele. The boxes represent genotypes, and the circles represent a collection of nucleotide bases observed among a set of sequencing reads for the site in question. Only the nucleotide bases from the reads aligning to this site are observed. The true germ and somatic genotypes, the read-sampling pattern of the somatic chromosomes, and the specific parental alleles inherited by the offspring comprise the hidden data.

## *Probability of the observed data*

The full probability of the data given the parameters at a single site is given below. Applying probability laws and making progressive assumptions on the independence of

individual probability statements, this probability is reduced into several kernel functions. First the full components of $H$ are listed.

$$P(R,H \mid \Theta) = P(R,S,o',m',f',m*,f*,m,f \mid \Theta) \tag{3}$$

Taking advantage of Bayes' Rule allows breaking the statement into three initial terms.

$$\begin{aligned} P(R,H \mid \Theta) = {} & P(R_O,S_O,o',o,m*,f* \mid R_M,S_M,R_F,S_F,m',f',m,f,\Theta) \\ & \times P(R_M,S_M,R_F,S_F,m',f' \mid m,f,\Theta) \times P(m,f \mid \Theta) \end{aligned} \tag{4}$$

Since the probability of observing reads for any individual are not directly dependent upon the reads observed for other individuals, the data terms for individuals may be broken apart, conditional on the genotypes. The probability of observing the offspring zygotic genotype is also separated from the offspring read data.

$$\begin{aligned} P(R,H \mid \Theta) = {} & P(R_O,S_O,o' \mid o,m*,f*,\Theta) \times P(o,m*,f* \mid m,f,\Theta) \\ & \times P(R_M,S_M,m' \mid m,f,\Theta) \times P(R_F,S_F,f' \mid m,f,\Theta) \times P(m,f \mid \Theta) \end{aligned} \tag{5}$$

Next we can separate the probability of each individual's somatic genotype from the probability of observing the reads, which depends only on the somatic genotype.

$$\begin{aligned} P(R,H \mid \Theta) = {} & P(R_O,S_O \mid o',\Theta) \times P(o' \mid o,m*,f*,\Theta) \\ & \times P(o_a \mid m*,\Theta) \times P(o_b \mid f*,\Theta) \times P(m*,f* \mid m,f,\Theta) \\ & \times P(R_M,S_M \mid m',\Theta) \times P(m' \mid m,f,\Theta) \\ & \times P(R_F,S_M \mid f',\Theta) \times P(f' \mid m,f,\Theta) \times P(m,f \mid \Theta) \end{aligned} \tag{6}$$

Finally the sampling indicators for each read are separated, and the probability for each individual read is specified. Only the parameter that each probability function depends upon is specified. Furthermore, each chromosome is assumed to mutate independently of the other chromosome allele for all individuals.

$$P(R,H \mid \Theta) = \prod_{k=1}^{N_{RO}} P(R_{Ok} \mid S_{Ok}, o', \varepsilon) P(S_{Ok}) \tag{7}$$

$$\times P(o_a' \mid o_a, \mu_S) \times P(o_b' \mid o_b, \mu_S) \tag{8}$$

$$\times P(o_a \mid m^*, \mu) \times P(o_b \mid f^*, \mu) \times P(m^* \mid m) \times P(f^* \mid f) \tag{9}$$

$$\times \prod_{k=1}^{N_{RM}} P(R_{Mk} \mid S_{Mk}, o', \varepsilon) P(S_{Mk}) \tag{10}$$

$$\times P(m_a' \mid m_a, \mu_S) \times P(m_b' \mid m_b, \mu_S) \tag{11}$$

$$\times \prod_{k=1}^{N_{RF}} P(R_{Fk} \mid S_{Fk}, o', \varepsilon) P(S_{Fk}) \tag{12}$$

$$\times P(f_a' \mid f_a, \mu_S) \times P(f_b' \mid f_b, \mu_S) \tag{13}$$

$$\times P(m, f \mid \theta) \tag{14}$$

The equation line numbers shown above now define the kernel functions: the Error kernel is seen in lines (7), (10), and (12), the Somatic Mutation kernel is seen in lines (8), (11), and (13), the Germline Mutation kernel is defined as line (9), and the Population kernel is line (14).

$$\begin{aligned} P(R,H \mid \Theta) = {} & K_E(R_O, S_O, o', \varepsilon) \times K_S(o', o, \mu_S) \times K_G(o, m^*, f^*, m, f, \mu) \\ & \times K_E(R_M, S_M, m', \varepsilon) \times K_S(m', m, \mu_S) \\ & \times K_E(R_F, S_F, f', \varepsilon) \times K_S(f', f, \mu_S) \times K_P(m, f, \theta) \end{aligned} \tag{15}$$

Before proceeding to the definitions of each of these four kernels, recall the nature of *R* and *H*, the data within our model. The hidden data, *H*, is not observable so we are interested in specifically calculating the marginal probability of the observed data, *R*, given

the parameters. The law of total probability allows this calculation using the joint probability of $R$ and $H$ derived above.

$$P(R \mid \Theta) = \sum_{H} P(R, H \mid \Theta) \tag{16}$$

Applying the kernels of equation (15) to this result and partitioning $R$ and $H$ into their constituent parts gives:

$$P(R \mid \Theta) = \sum_{m}\sum_{f}\sum_{m^*}\sum_{f^*}\sum_{o}\sum_{o'}\sum_{S_O}\sum_{m'}\sum_{S_M}\sum_{f'}\sum_{S_F} \begin{bmatrix} K_P(m,f,\theta) \\ \times K_S(o',o,\mu_S) \\ \times K_E(R_O,S_O,o',\varepsilon) \\ \times K_G(o,m^*,f^*,m,f,\mu) \\ \times K_S(m',m,\mu_S) \\ \times K_E(R_M,S_M,m',\varepsilon) \\ \times K_S(f',f,\mu_S) \\ \times K_E(R_F,S_F,f',\varepsilon) \end{bmatrix} \tag{17}$$

Factoring the Population kernel to only the sum over the possible zygotic genotypes of the parents is the first step of factoring kernels outside of sums where they vary, leading to a simpler, more intuitive formulation.

$$P(R \mid \Theta) = \sum_{m}\sum_{f} K_P(m,f,\theta) \sum_{m^*}\sum_{f^*}\sum_{o}\sum_{o'}\sum_{S_O}\sum_{m'}\sum_{S_M}\sum_{f'}\sum_{S_F} \begin{bmatrix} K_S(o',o,\mu_S) \\ \times K_E(R_O,S_O,o',\varepsilon) \\ \times K_G(o,m^*,f^*,m,f,\mu) \\ \times K_S(m',m,\mu_S) \\ \times K_E(R_M,S_M,m',\varepsilon) \\ \times K_S(f',f,\mu_S) \\ \times K_E(R_F,S_F,f',\varepsilon) \end{bmatrix} \tag{18}$$

Continuing in this manner produces the algorithm used to calculate the probability at each site.

$$P(R \mid \Theta) = \sum_{m,f} K_P(m,f,\theta) \times \left[ \begin{array}{l} \sum_{f'} \left( \begin{array}{l} K_S(f',f,\mu_S) \\ \times \sum_{S_F} K_E(R_F,S_F,f',\varepsilon) \end{array} \right) \\ \times \sum_{m'} \left( \begin{array}{l} K_S(m',m,\mu_S) \\ \times \sum_{S_M} K_E(R_M,S_M,m',\varepsilon) \end{array} \right) \\ \times \sum_{m^*,f^*} \sum_{o} \left( \begin{array}{l} K_G(o,m^*,f^*,m,f,\mu) \\ \times \sum_{o'} \left[ \begin{array}{l} K_S(o',o,\mu_S) \\ \times \sum_{S_O} K_E(R_O,S_O,o',\varepsilon) \end{array} \right] \end{array} \right) \end{array} \right] \tag{19}$$

The sums over $S_I$ are incorporated within the three instances of $K_E$ and the sum over *m** and *f** within $K_G$, forming $\tilde{K}_E$ and $\tilde{K}_G$, to improve algorithm performance.

$$P(R \mid \Theta) = \sum_{m,f} K_P(m,f,\theta) \times \left[ \begin{array}{l} \sum_{f'} \left( \begin{array}{l} K_S(f',f,\mu_S) \\ \times \tilde{K}_E(R_F,f',\varepsilon) \end{array} \right) \\ \times \sum_{m'} \left( \begin{array}{l} K_S(m',m,\mu_S) \\ \times \tilde{K}_E(R_M,m',\varepsilon) \end{array} \right) \\ \times \sum_{o} \left( \begin{array}{l} \tilde{K}_G(o,m,f,\mu) \\ \times \sum_{o'} \left[ \begin{array}{l} K_S(o',o,\mu_S) \\ \times \tilde{K}_E(R_O,o',\varepsilon) \end{array} \right] \end{array} \right) \end{array} \right] \tag{20}$$

With the algorithm in place for determining the total probability of the observed data, I continue by defining the specific kernel functions used in the calculation.

*Kernel functions*

There is a fundamental assumption governing each kernel which contains a base substitution, whether through sequencing error, somatic mutation, or germline mutation. The simplest base substitution model, the Jukes-Cantor (1969) model, is assumed which dictates that each of the four nucleotides, or bases, is equally likely and any base may mutate to any of the others with equal probability. In addition, each of the rate parameters are assumed to take place in continuous time and are in fact randomization rates, which randomly substitute the observed nucleotide with any of the four, including the one previously observed. These randomization rates determine the mutation and error rates that are ultimately reported to the user. Allowing for virtual bases in general allows the capturing of multiple mutation hits on the same branch of a tree, which may in fact arise as a possibility on long somatic branches covering many millions of generations of cell divisions. A single kernel function, $K_{\Delta}$, encompassing these ideas is used to build all of the substitution-based kernels, including the Error kernel, the Somatic mutation kernel, and the Germline mutation kernel

$$K_{\Delta}(x_i, y_i, \delta) = \tfrac{1}{4}(1 - e^{-\delta}) + I(x_i = y_i)e^{-\delta} \tag{21}$$

where $x_i$ and $y_i$ are single alleles and $\delta$ is the substitution rate. Each kernel presented is reduced to a form that is easily pre-calculated and accessed by a simple lookup within the software implementation.

*Somatic mutation kernel*

The Somatic mutation kernel is the most basic of the kernel functions, including the product of two basic substitution kernels, one for each allele:

$$\begin{aligned} K_S(x',x,\mu_S) &= P(x' \mid x,\mu_S) \\ &= P(x_a' \mid x_a,\mu_S) \times P(x_b' \mid x_b,\mu_S) = K_\Delta(x_a',x_a,\mu_S) \times K_\Delta(x_b',x_b,\mu_S) \end{aligned} \tag{22}$$

$$K_S(x',x,\mu_S) = \begin{cases} (\frac{1}{4} + \frac{3}{4}e^{-\mu_S})^2, & \text{if } x_a' = x_a \text{ and } x_b' = x_b \\ (\frac{1}{4} + \frac{3}{4}e^{-\mu_S}) \times (\frac{1}{4} - \frac{1}{4}e^{-\mu_S}), & \text{if } x_a' = x_a \text{ and } x_b' \neq x_b, \\ & \text{or } x_a' \neq x_a \text{ and } x_b' = x_b \\ (\frac{1}{4} - \frac{1}{4}e^{-\mu_S})^2, & \text{if } x_a' \neq x_a \text{ and } x_b' \neq x_b \end{cases} \tag{23}$$

*Germline mutation kernel*

The Germline mutation kernel is similar in form to the Somatic mutation kernel, but adds the complicating factor of selecting the alleles inherited from the parents to the offspring zygote. The possible inheritance patterns are summed over within this kernel.

$$\begin{aligned} \tilde{K}_G(o,m,f,\mu) &= \sum_{m^*f^*} K_G(o,m^*,f^*,m,f,\mu) \\ &= \sum_{m^*f^*} P(o \mid m^*,f^*,m,f,\mu) \times P(m^*,f^* \mid m,f) \\ &= \sum_{m^*f^*} P(o_a \mid m^*,m,\mu) \times P(o_b \mid f^*,f,\mu) \times P(m^* \mid m) \times P(f^* \mid f) \end{aligned} \tag{24}$$

$$\tilde{K}_G(o,m,f,\mu)=\frac{1}{4}\begin{bmatrix} P(o_a \mid m^*=m_a,\mu)\times P(o_b \mid f^*=f_a,\mu) \\ +P(o_a \mid m^*=m_a,\mu)\times P(o_b \mid f^*=f_b,\mu) \\ +P(o_a \mid m^*=m_b,\mu)\times P(o_b \mid f^*=f_a,\mu) \\ +P(o_a \mid m^*=m_b,\mu)\times P(o_b \mid f^*=f_b,\mu) \end{bmatrix} \tag{25}$$

$$\tilde{K}_G(o,m,f,\mu)=\frac{1}{4}\begin{bmatrix} K_\Delta(o_a,m_a,\mu)\times K_\Delta(o_b,f_a,\mu) \\ +K_\Delta(o_a,m_a,\mu)\times K_\Delta(o_b,f_b,\mu) \\ +K_\Delta(o_a,m_b,\mu)\times K_\Delta(o_b,f_a,\mu) \\ +K_\Delta(o_a,m_b,\mu)\times K_\Delta(o_b,f_b,\mu) \end{bmatrix} \tag{26}$$

$$\tilde{K}_G(o,m,f,\mu)=\left[\frac{K_\Delta(o_a,m_a,\mu)+K_\Delta(o_a,m_b,\mu)}{2}\right]\left[\frac{K_\Delta(o_b,f_a,\mu)+K_\Delta(o_b,f_b,\mu)}{2}\right] \tag{27}$$

$$\tilde{K}_G(o,m,f,\mu)=\begin{bmatrix} \begin{Bmatrix} \frac{1}{4}+\frac{3}{4}e^{-\mu}, & o_a=m_a=m_b \\ \frac{1}{4}+\frac{1}{4}e^{-\mu}, & o_a=m_a\neq m_b \text{ or } o_a=m_b\neq m_a \\ \frac{1}{4}-\frac{1}{4}e^{-\mu}, & o_a\neq m_a \text{ and } o_a\neq m_b \end{Bmatrix} \\ \times \begin{Bmatrix} \frac{1}{4}+\frac{3}{4}e^{-\mu}, & o_b=f_a=f_b \\ \frac{1}{4}+\frac{1}{4}e^{-\mu}, & o_b=f_a\neq f_b \text{ or } o_b=f_b\neq f_a \\ \frac{1}{4}-\frac{1}{4}e^{-\mu}, & o_b\neq f_a \text{ and } o_b\neq f_b \end{Bmatrix} \end{bmatrix} \tag{28}$$

*Error kernel*

As with the germline mutation kernel, the full error kernel includes the basic substitution kernel at its core, but adds on the complexity of summing over one of the hidden data states, in this case the $S_I$ indicators for the chromosome sampling by the reads.

In addition the Error kernel calculates the joint probability of all of the reads for one individual at a site, so there are $N_{RI}$ opportunities for base substitution in the kernel. An additional element to this kernel is the probability, *p*, that chromosome *a* is sampled by a read. With some high-throughput sequencing applications *p* has been shown to be greater than 0.5 for one or the other allele (Craig A.W., personal communication). Possible sources of this bias are heterogeneity in the sequenced DNA sample leading to unequal amplification of alleles, biases in sequencing chemistry, or biases towards the reference allele when using reference-guided assembly.

$$\begin{aligned}\tilde{K}_E(R_I,x',\varepsilon) &= \sum_{S_I} K_E(R_I,S_I,x',\varepsilon) \\ &= \sum_{S_{I1}=a}^{b}\sum_{S_{I2}=a}^{b}\cdots\sum_{S_{IN_{RI}}=a}^{b}\prod_{k=1}^{N_{RI}} P(R_{Ik}\mid S_{Ik},x',\varepsilon)P(S_{Ik}\mid p) \\ &= \prod_{k=1}^{N_{RI}}\sum_{S_{Ik}=a}^{b} P(R_{Ik}\mid S_{Ik},x',\varepsilon)P(S_{Ik}\mid p) \\ &= \prod_{k=1}^{N_{RI}} K_\Delta(R_{Ik},x_a',\varepsilon)p + K_\Delta(R_{Ik},x_b',\varepsilon)(1-p)\end{aligned} \tag{29}$$

When *x'* is homozygous, this simplifies to:

$$K_\Delta(R_I,x',\varepsilon) = \left(\frac{1}{4}+\frac{3}{4}e^{-\varepsilon}\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}=x_a')} \times \left(\frac{1}{4}-\frac{1}{4}e^{-\varepsilon}\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}\neq x_a')}, \quad x_a' = x_b' \tag{30}$$

With heterozygous *x'* the reduced form is:

$$\tilde{K}_E(R_I,x',\varepsilon)=\begin{cases}\left(\frac{1}{4}+e^{-\varepsilon}(p-\frac{1}{4})\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}=x'_a)} & \\ \quad\times\left(\frac{1}{4}+e^{-\varepsilon}(\frac{3}{4}-p)\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}=x'_a)}\times\left(\frac{1}{4}-\frac{1}{4}e^{-\varepsilon}\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}\neq x'_a)} & , x'_a\neq x'_b \end{cases} \quad (31)$$

When $p$ = .5, the heterozygous term of $\tilde{K}_E$ reduces to:

$$\tilde{K}_E(R_I,x',\varepsilon)=\begin{cases}\left(\frac{1}{4}+\frac{1}{4}e^{-\varepsilon}\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}=x'_a)+I(R_{Ik}=x'_b)}\times\left(\frac{1}{4}-\frac{1}{4}e^{-\varepsilon}\right)^{\sum_{k=1}^{N_{RI}} I(R_{Ik}\neq x'_a)\times I(R_{Ik}\neq x'_b)} & , \; x'_a\neq x'_b \end{cases} \quad (32)$$

## *Population kernel*

In this kernel the probability of a sample of four alleles constituting the two zygotic genotypes of the parents, or the parent-pair allele spectrum, is defined using coalescent theory[30] under the finite sites model[31]. The finite sites model is used since *de novo* mutations are considered a possibility even at sites with two alleles already segregating in the population. Biologically such sites could be quite informative as hyper-mutable sites. In addition through investigations of real data, many sites are found to contain three parental alleles, a phenomena that has also been described in other human data sets[32]. Properly dealing with rare cases is essential when working with whole-genome human sequence data because even rare phenomena will be present in non-trivial quantities when three billion bases are considered. The resulting kernel will take the form:

$$K_P(m,f,\theta)=P(m,f\mid\theta)=P(m_a,m_b,f_a,f_b\mid\theta) \quad (33)$$

The nucleotides $m_a$, $m_b$, $f_a$, and $f_b$ constitute a sample of four alleles from a finite, randomly-mating population, with effective size of $N_e$. Allowing up to three different allele states amongst this sample of four, the following allele spectra are possible:

$$\Phi(m_a, m_b, f_a, f_b) \in \begin{Bmatrix} 4-0-0, & \textit{all alleles are the same} \\ 3-1-0, & \textit{two allele states with one occuring once} \\ 2-2-0, & \textit{two allele states with both occuring twice} \\ 2-1-1, & \textit{three allele states} \end{Bmatrix} \quad (34)$$

When calculating the probability of each of these spectra, the possible specific nucleotide used (assuming all are equal in frequency) and the possible order in which they occur must be considered.

$$P_O(\Phi(m_a.m_b, f_a, f_b)) = \begin{cases} \frac{1}{4}, & \Phi(m_a.m_b, f_a, f_b) = 4-0-0 \\ \frac{1}{12} \times \frac{1}{4}, & \Phi(m_a.m_b, f_a, f_b) = 3-1-0 \\ \frac{1}{12} \times \frac{1}{3}, & \Phi(m_a.m_b, f_a, f_b) = 2-2-0 \\ \frac{1}{24} \times \frac{1}{6}, & \Phi(m_a.m_b, f_a, f_b) = 2-1-1 \end{cases} \quad (35)$$

Applying these concepts to the formulation of the kernel function gives

$$K_P(m, f, \theta) = P_\Phi(\Phi(m_a, m_b, f_a, f_b) \mid \theta) \times P_O(\Phi(m_a, m_b, f_a, f_b)) \quad (36)$$

All that must be determined at this point is the calculation of the first term of equation (36). Mutations are allowed to occur continuously along the genealogical tree connecting these four sampled chromosomes, back to their most recent common ancestor. The allele spectrum observed in the sample is a function of the number of mutations

occurring in this tree and the branches on which they occur. Consider that there are two possible tree structures, shown in Figure 2-2.

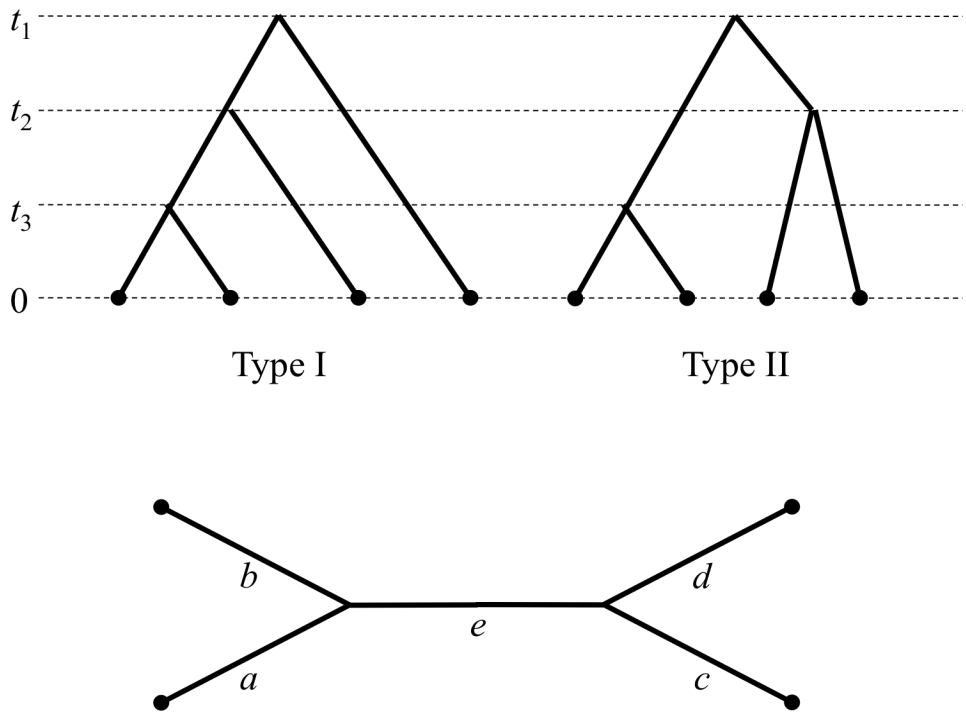



**Figure 2-2.** Possible genealogical trees for a sample of four alleles from a single population.

In this illustration, the $t_i$ are the times before coalescent events of the chromosomes, in units of $2N_e$. Note that in coalescent theory all times are in terms of the present looking backwards into the past. Under the finite sites model [31], the joint distribution of $t_1$, $t_2$, and $t_3$ is:

$$f(t_1,t_2,t_3) = 18e^{-t_1-2t_2-3t_3} \tag{37}$$

and the distribution of $T$, the total length of the actual genealogy is given by

$$f(T) = \tfrac{3}{2} e^{-T/2} (1 - e^{-T/2})^2 \tag{38}$$

The total length of $T$ is also easily put in terms of the $t_i$:

$$T = 2(t_1 - t_2) + 3(t_2 - t_3) + 4t_3 \tag{39}$$

The probability no mutation will occur in a tree of length $T$ is $e^{-T\theta/2}$. The probability no mutation occurs across in the genealogy of the four samples is constructed by integrating over the length of the tree using equation (38). Let $k$ be the number of mutations that occur.

$$P(k=0) = \int_0^\infty e^{-T\theta/2} \times \tfrac{3}{2} e^{-T/2} \left(1 - e^{-T/2}\right)^2 dT \approx \frac{6}{6 + 11\theta} \tag{40}$$

In general the probability of number of mutations occurring is given by (Cartwright R., personal communication):

$$P(k) = 3\theta^k \left( \frac{1}{(1+\theta)^{k+1}} - \frac{2}{(2+\theta)^{k+1}} + \frac{1}{(3+\theta)^{k+1}} \right) \tag{41}$$

Substituting $k$=1 into equation (41) gives the probability of a single mutation as

$$P(k=1) = \frac{6\theta\left(11 + 12\theta + 3\theta^2\right)}{(1+\theta)^2 (2+\theta)^2 (3+\theta)^2} \approx \frac{66\theta}{36 + 132\theta} \tag{42}$$

Again assuming only two mutations can occur at this site along the genealogical tree, the probability of two mutations is

$$P(k=2) \approx 1 - P(k=1) - P(k=0) = \frac{121\theta^2}{2(3+11\theta)(6+11\theta)} \tag{43}$$

Returning to the probability of each allele spectrum possibility, consider first $P_{\Phi}(\Phi(m_a, m_b, f_a, f_b) = 4-0-0)$. This spectrum of four copies of the same allele can occur either when $k = 0$ or when $k = 2$ and the two mutations occur on the same branch with the second resetting the first. To simplify the calculation the probabilities are given in terms of the branches of the unrooted tree in Figure 2-2. With genealogy Type I, $a = t_3, b = t_3, c = t_2, d = 2t_1 - t_2, and\ e = t_2 - t_3$. For genealogy Type II, which occurs with half the frequency of Type I, $a = t_3, b = t_3, c = t_2, d = t_2, and\ e = 2t_1 - t_2 - t_3$. The probability of the possible locations of the two-mutation case are then given by the following, incorporating the joint distribution of $t_i$ from equation (37):

$$P_{\Phi}(4-0-0 \mid k=2) = \frac{1}{3}\int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{a^2+b^2+c^2+d^2+e^2}{TT} f(t_1,t_2,t_3)\,dt_3\,dt_2\,dt_1 \qquad (44)$$

Using the probability of zero mutation from equation (40) and substituting in the appropriate branch lengths into the integral, the probability of the $4-0-0$ allele spectrum is

$$\begin{aligned} P_{\Phi}(\Phi(m_a,m_b,f_a,f_b) = 4-0-0) &= P(4-0-0 \mid k=0) + P(4-0-0 \mid k=2) \\ &= P(4-0-0 \mid k=0) + \frac{2P(4-0-0 \mid k=2, Type\ I) + P(4-0-0 \mid k=2, Type\ II)}{3} \\ &= \frac{6}{6+11\theta} + P(k=2)\times\left(\frac{2\times 0.166982031953 + 0.153953001646}{3}\right) \\ &= \frac{6}{6+11\theta} + \frac{121\theta^2}{2(3+11\theta)(6+11\theta)} \times 0.162639021851 \end{aligned} \qquad (45)$$

Moving on to the probability of the $2-2-0$ spectrum, which can occur when a single mutation occurs on the internal branch *e*, or when two mutations occur within the

internal branch as long as the second does not resent the first. In general the integrals for this spectrum is

$$\begin{aligned} P_{\Phi}(2-2-0 \mid k=1) &= \int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{e}{T} f\left(t_1,t_2,t_3\right) dt_3\, dt_2\, dt_1 \\ P_{\Phi}(2-2-0 \mid k=2) &= \frac{2}{3}\int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{e^2}{TT} f\left(t_1,t_2,t_3\right) dt_3\, dt_2\, dt_1 \end{aligned} \tag{46}$$

and the full probability is therefore

$$\begin{aligned} P_{\Phi}\left(\Phi(m_a,m_b,f_a,f_b)=2-2-0\right) &= P\left(2-2-0 \mid k=1\right)+P(2-2-0 \mid k=2) \\ &= \frac{2P(2-2-0 \mid k=1, Type\ I)+P(2-2-0 \mid k=1, Type\ II)}{3} \\ &\quad + \frac{2P(2-2-0 \mid k=2, Type\ I)+P(2-2-0 \mid k=2, Type\ II)}{3} \\ &= \frac{17.2671887442\times\theta}{36+132\theta} + \frac{121\theta^2}{2\left(3+11\theta\right)\left(6+11\theta\right)} \times 0.224534553633 \end{aligned} \tag{47}$$

The $3-1-1$ spectrum will occur with a single mutation on any terminal branch (*a* to *d*) and with two mutations where both occur on a terminal branch without the second resetting the first or where one occurs on the internal branch and the second on an external branch while resetting the first.

$$\begin{aligned} P_{\Phi}(3-1-0 \mid k=1) &= \int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{a+b+c+d}{T} f\left(t_1,t_2,t_3\right) dt_3\, dt_2\, dt_1 \\ P_{\Phi}(3-1-0 \mid k=2) &= \frac{2}{3}\int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{a^2+b^2+c^2+d^2}{TT} f\left(t_1,t_2,t_3\right) dt_3\, dt_2\, dt_1 \\ &\quad + \frac{1}{3}\int_0^{\infty}\int_0^{t_1}\int_0^{t_2} \frac{e(T-e)}{TT} f\left(t_1,t_2,t_3\right) dt_3\, dt_2\, dt_1 \end{aligned} \tag{48}$$

$$
\begin{aligned}
P_{\Phi}\left(\Phi(m_a,m_b,f_a,f_b)=3-1-0\right) &= P\left(3-1-0 \mid k=1\right)+P(3-1-0 \mid k=2) \\
&= \frac{2P(3-1-0 \mid k=1, Type\ I)+P(3-1-0 \mid k=1, Type\ II)}{3} \\
&\quad + \frac{2P(3-1-0 \mid k=2, Type\ I)+P(3-1-0 \mid k=2, Type\ II)}{3} \\
&= \frac{48.7328112558\times\theta}{36+132\theta}+\frac{121\theta^2}{2\left(3+11\theta\right)\left(6+11\theta\right)}\times 0.271437801551
\end{aligned}
\tag{49}
$$

Finally the $2-1-1$ spectrum will occur only when two mutations occur on different branches and the second mutation does not reset the first.

$$
P_{\Phi}(2-1-1 \mid k=2)=\frac{2}{3}\int_0^{\infty}\int_0^{t_1}\int_0^{t_2} 1-\frac{a^2+b^2+c^2+d^2}{TT} f\left(t_1,t_2,t_3\right)dt_3\,dt_2\,dt_1 \tag{50}
$$

$$
\begin{aligned}
P_{\Phi}\left(\Phi(m_a,m_b,f_a,f_b)=2-1-1\right) &= P(2-1-1 \mid k=2) \\
&= \frac{2P(2-1-1 \mid k=2, Type\ I)+P(2-1-1 \mid k=2, Type\ II)}{3} \\
&= \frac{121\theta^2}{2\left(3+11\theta\right)\left(6+11\theta\right)}\times 0.341388622965
\end{aligned}
\tag{51}
$$

With all of the $P_{\Phi}\left(\Phi(m_a,m_b,f_a,f_b)\right)$ thus defined, the Population kernel is calculated by substituting in the appropriate results of equations (45), (47), (49), and (51) along with the ordering and "tagging" probability from equation (35) into equation (36).

## *Tree peeling algorithm*

In this section a method of recursively evaluating the probability equation (20) is introduced. This algorithm, referred to as tree peeling, in brief calculates the model

probability by peeling the family pedigree from the sequencing reads down to the zygotic genotype of the parents. This formulation of the model probability will be used to extend the model to multiple family types and as part of the algorithm for estimating the parameters from the observed data. For the pedigree in Figure 2-1, consider the individual genotypes as nodes, with a particular genotype at a node represented by $X$. The inheritance branches define a 'parent-child' relationship between each of the nodes. For example, the somatic genotype nodes are 'children' of their respective zygotic genotype nodes. Let $Y_j$ represent the genotype of the $j$-th child node of $X$. In this particular pedigree $j$ is equal to 1 at each internal node excepting at the parental zygotic node, where it is equal to 3, but this will not be the case with more complex pedigrees such as those including twins or separate somatic samplings of a single individual. This allows rewriting of equation (20) as the following

$$P(R \mid \Theta) = \sum_{X} P(X \mid \Theta) P(R \mid X, \Theta) \tag{52}$$

$$P(R \mid X, \Theta) = \prod_{j} \sum_{Y_j} P(Y_j \mid X, \Theta)\, P(R_j \mid Y_j, X, \Theta) \tag{53}$$

In equation (52) $X$ takes all of the possible states of the four-allele sample making up the parental zygotic genotypes. In equation (53) $X$ takes the possible genotypes of the parental zygotes and the child zygote, while $Y$ takes the genotypic states at the child zygotic node and the somatic tissue nodes respectively.

*Summary statistics*

In the course of executing the tree-peeling algorithm, the values of a set of summary statistics are also calculated that fully describe the hidden data. The expected values of these summaries are incorporated into our parameter estimation procedure. Specifically, the summaries are defined as:

$S_{400}$ - estimates the probability of a $4-0-0$ parental germline allele spectrum;

$S_{310}$ - estimates the probability of a $3-1-0$ parental germline allele spectrum;

$S_{220}$ - estimates the probability of a $2-2-0$ parental germline allele spectrum;

$S_{211}$ - estimates the probability of a $2-1-1$ parental germline allele spectrum.

$S_M$ - the number of nucleotide mismatches between *m** and $o_a$, representing germline mutations inherited from the mother;

$S_F$ - the number of nucleotide mismatches between *f** and $o_b$, representing germline mutations inherited from the father;

$S_{Som}$ - the number of nucleotide mismatches between all *x* and *x*', representing the total number of somatic mutations in the pedigree;

$S_{Hom}$ - the number of nucleotide matches between a somatic homozygous genotype and its sequencing reads, representing the read coverage of all somatic homozygotes in the pedigree;

$S_{Het}$ - the number of nucleotide matches between a somatic heterozygous genotype and its sequencing reads, representing the read coverage of all somatic heterozygotes in the pedigree;

$S_E$ - the number of nucleotide mismatches between a somatic genotype and its sequencing reads, representing the total number of sequencing errors in the pedigree.

Initially the values of these summary statistics are determined by the genotype comparisons between a parent node *X* and child node *Y*, as described in the tree-peeling algorithm. $S_{Sum}(R, X \to Y)$ denotes the contribution to a particular summary statistic due to the comparison between *X* and *Y*. Beginning at the tips of the pedigree tree at the branches between the somatic genotypes and the sequencing reads, the summary statistics are calculated as

$$S_{Hom}(R_I, Y \to R_I) = I(Y_a = Y_b) \sum_{k=1}^{N_{RI}} I(R_{Ik} = Y_a) \tag{54}$$

$$S_{Het}(R_I, Y \to R_I) = I(Y_a \neq Y_b) \sum_{k=1}^{N_{RI}} I(R_{Ik} = Y_a) + I(R_{Ik} = Y_b) \tag{55}$$

$$S_E(R_I, Y \to R_I) = \sum_{k=1}^{N_{RI}} I(R_{Ik} \neq Y_a) \times I(R_{Ik} \neq Y_b) \tag{56}$$

Moving to the somatic nodes, where $X \in \{m, f, o\}$ and $Y \in \{m', f', o'\}$, the summary statistic describing somatic mutation events is calculated.

$$S_{Som}(R_I, X \to Y) = I(X_a \neq Y_a) + I(X_b \neq Y_b) \tag{57}$$

On branches containing the transmission of parental alleles to the offspring zygote, $X \in \{m^*, f^*\}$ and $Y \in \{o\}$, the summary statistic for germline mutation is calculated.

$$S_m(R_O, X \to Y) = I(X \neq Y_a) \tag{58}$$

$$S_f(R_O, X \to Y) = I(X \neq Y_a) \tag{59}$$

The four parent-allele spectra summaries are the same as those from the Population kernel results of equation (36).

As the algorithm given in equations (52) and (53) for calculating the model probability (20) is performed, the cumulative summary statistics describing the hidden data at the site are built. Specifically, the value of a summary statistic at node $X$ is the weighted average of all of the summaries of the same type coming from the child nodes $Y$ plus the summary calculated along the branch from $X$ to $Y$.

$$S_{Sum}(R, X) = \sum_j \left[ \frac{\sum_{Y_j} P(Y_j \mid X)\, P(R_j \mid Y_j) \times \left[ S_{Sum}(R_j, Y_j) + S_{Sum}(R_j, X \to Y_j) \right]}{\sum_{Y_j} P(Y_j \mid X)\, P(R_j \mid Y_j)} \right] \tag{60}$$

By iteratively calculating the summary statistics in this manner, their expected value is found by summing the weighted average of equation (60) as $X$ takes all possible genotypic states at each internal node.

$$E(S_{Sum}) = \sum_H P(H \mid R, \Theta)\, S_{Sum}(R, H) = \frac{\sum_H P(H, R \mid \Theta) \times S_{Sum}(R, H)}{\sum_H P(H, R \mid \Theta)} \tag{61}$$

*Calculating the probability of de novo mutation, $\delta$*

Perhaps the most useful summary of the hidden data is calculated as a single probability that a mutation has occurred on a single branch or on a subset of branches. In particular, a *de novo* mutation will have occurred within a trio pedigree if a mutation occurs along either the maternal or paternal germline branches or along the somatic branch in the offspring. This probability, which I label with the letter $\delta$, is used to rank sites by order of those most likely to have a mutation event. In general, the probability of at least one mutation over the entire pedigree is calculated as one minus the probability of no mutations. Let $\varnothing$ denote the event of no mutation and $\varnothing_{X \to R}$ the event of no mutation on along the branches between node *X* and observed data $R_X$. Using the probability of the observed data calculated in equation (20) or recursively in equation (52), I calculate

$$\delta = P(\textit{child de novo}) = 1 - P\left(\varnothing \mid R, \Theta\right) = 1 - \frac{P\left(\varnothing, R \mid \Theta\right)}{P\left(R \mid \Theta\right)} \tag{62}$$

$$P\left(\varnothing, R \mid \Theta\right) = \sum_X P\left(X \mid \Theta\right) P\left(R, \varnothing_{X \to R} \mid X, \Theta\right) \tag{63}$$

$$P\left(R, \varnothing_{X \to R} \mid X\right) = \prod_j \sum_{Y_j} P\left(Y_j, \varnothing_{X \to Y_j} \mid X, \Theta\right) \times P\left(R_j, \varnothing_{Y_j \to R_j} \mid Y_j, \Theta\right) \tag{64}$$

$$P\left(\varnothing_{X \to Y_j}, Y_j \mid \Theta\right) = I\left(Y_j = X\right) \times P\left(Y_j \mid X, \Theta\right) \tag{65}$$

The full, unpacked formula used to calculate the probability of a child (germline or somatic) mutation in the trio pedigree is given by

$$\delta = 1 - \frac{\sum_{m,f}\left(\left[\begin{array}{l} P(m,f \mid \theta)\times P(R_M \mid m,\Theta) \\ \times P(R_F \mid \Theta)\times\left[\frac{1/4 + 3/4\, e^{-\mu_S}}{2}\right]^2 \end{array}\right]\times\sum_o\left[\begin{array}{l} \left(I(o_a = m_a)+I(o_a = m_b)\right) \\ \times\left(I(o_b = f_a)+I(o_b = f_b)\right) \\ \times P(R_O \mid o' = o)\times\left[1/4 + 3/4\, e^{-\mu_S}\right]^2 \end{array}\right]\right)}{P(R \mid \Theta)} \quad (66)$$

*Maximum-likelihood estimations of model parameters*

The ideal use of this model would be to estimate the parameters and apply the estimated rates to create the posterior probability of *de novo* mutation at each site examined. It is trivial to estimate the parameters from the full data by maximizing the likelihood function

$$L(\Theta \mid R,H) = P(R,H \mid \Theta) \quad (67)$$

However the hidden data is not observable so the marginal likelihood must be maximized instead which is not trivial.

$$L(\Theta \mid R) = \sum_H P(R,H \mid \Theta) \quad (68)$$

An expectation-maximization algorithm (EM) [33] is employed to iteratively maximize the function

$$Q(\Theta \mid R,\Theta_n) = \sum_H P(H \mid R,\Theta_n)\times \log P(R,H \mid \Theta) \quad (69)$$

until it converges, allowing the use of the solution for maximizing equation (67) to eventually maximize equation (68). After selecting an initial $\Theta_0$ we alternate between an expectation (E-) step, in which we calculate the expected values of the summary statistics

across all sites using the method given by equation (61), and a maximization (M-) step, where the MLEs of the model parameters are calculated. The MLEs resulting from each M-step are used in the following E-step. A feature of the EM algorithm is the overall likelihood of the model, $L(\Theta_n \mid R)$, is guaranteed to increase from iteration to iteration [33]. Therefore the MLEs will converge to their locally optimal values.

Following all iterations, the expected value of each summary statistic is summed across all sites and the result is used to set the relevant rate parameter using the equations below.

$$\tilde{S}_{Sum} = \sum_{k=1}^{N_s} E\left(S_{k,Sum}\right) \tag{70}$$

$$\hat{\mu} = -\log\left[1 - \frac{4}{3}\frac{\tilde{S}_m + \tilde{S}_f}{N_S}\right] \tag{71}$$

$$\hat{\mu}_S = -\log\left[1 - \frac{4}{3}\frac{\tilde{S}_{Som}}{N_S}\right] \tag{72}$$

$$\hat{\varepsilon} = -\log\left[1 - \frac{1}{3}\frac{3\tilde{S}_{Hom} + 2\tilde{S}_{Het} + 5\tilde{S}_E - \sqrt{9\tilde{S}_{Hom}^2 + \left(2\tilde{S}_{Het} - \tilde{S}_E\right)^2 + 6\tilde{S}_{Hom}\left(2\tilde{S}_{Het} + \tilde{S}_E\right)}}{\tilde{S}_{Hom} + \tilde{S}_{Het} + \tilde{S}_E}\right] \tag{73}$$

The MLE of the population mutation parameter, $\theta$, is found as the positive root of a fifth order polynomial with coefficients defined by a 6x1 vector, $MS$, where $S = \{S_{400}, S_{310}, S_{220}, S_{211}\}$ is the 4x1 vector of summary statistics calculated through

equations (36) and (61). The required pre-calculated 6x4 matrix $M$ is given below (Cartwright R., personal communication).

$$M = \begin{bmatrix} 0 & 0.03185903 & 0.03185903 & 0.06371801 \\ -0.05840823 & 0.38511249 & 0.41376867 & 0.78568976 \\ -0.73892515 & 1.77060787 & 2.03328953 & 3.63959210 \\ -3.51554359 & 3.73506542 & 4.52126313 & 7.59386481 \\ -7.43876468 & 3.23339216 & 4.02587058 & 6.35199953 \\ -5.88936646 & 0.41934538 & 0.52464792 & 0.81818182 \end{bmatrix} \quad (74)$$

## *Model extension to larger pedigrees*

A feature of this method is the ease with which it can be extended to apply to multiple kinds of pedigrees. Figure 2-3 depicts the current theoretical limits to the size of pedigree which the model can currently accommodate; note that the specific parental alleles which are transferred to the offspring are not shown for graph clarity.

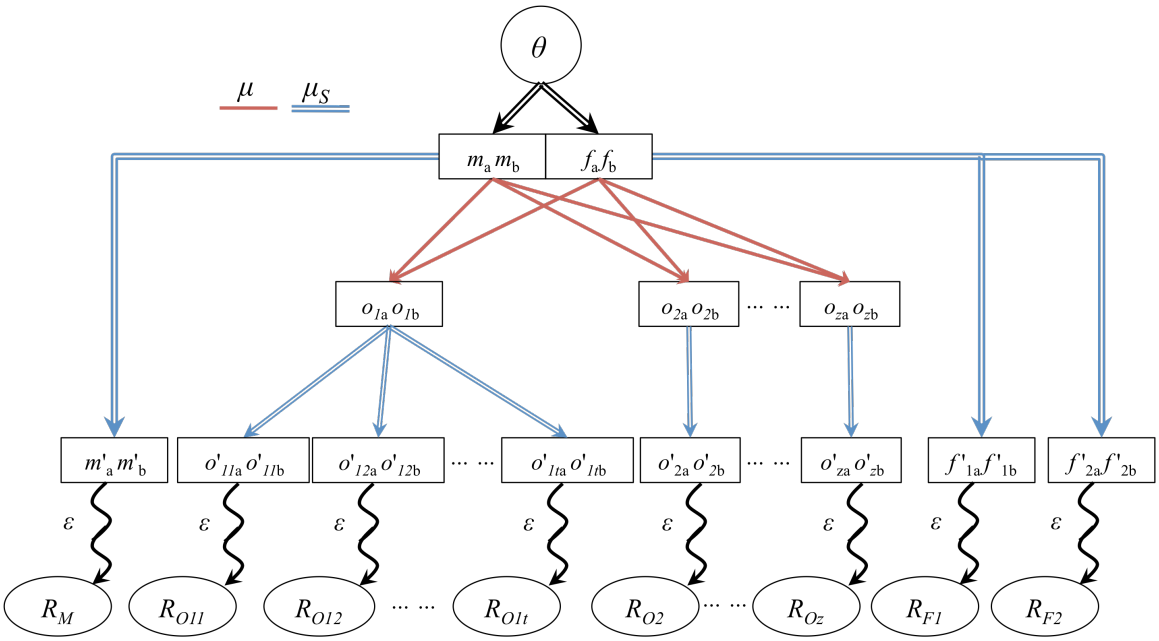



**Figure 2-3.** A pedigree model including two somatic samplings of a parent, *z* multiple zygote offspring (full-siblings), and *t* multiple offspring from a single zygote (monozygotic t-tuplets).

Compared with the trio pedigree shown in Figure 2-1, this pedigree has three notable differences. First it allows for a second somatic sampling of the father. This is a simple modification to the study design that is often used to validate predicted variants. Purposefully the double lines leading to the two somatic father samples overlap, to indicate the possibility of a most recent common ancestor cell or cells within the father somatic tissue that was not part of the germline, allowing the possibility for a shared somatic mutation between the two samples. Such will be the case if two skin samples are used for example. Mathematically, this adjustment creates a second instance of the somatic and error kernels along with an extra sum over the possible somatic genotypes within equation (20); the result shown in equation (75). Equivalently, within tree peeling algorithm when $X$

is taking the values of the parental genotypes, there is an extra $Y_j$ to include in the product for the probability and in the sum for the accumulating summary statistics.

$$P(R \mid \Theta) = \sum_{m,f} K_P(m,f,\theta) \times \left[ \begin{array}{l} \sum_{f_1'} \left( \begin{array}{l} K_S(f_1',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_1',\varepsilon) \end{array} \right) \\ \times \sum_{f_2'} \left( \begin{array}{l} K_S(f_2',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_2',\varepsilon) \end{array} \right) \\ \times \sum_{m'} \left( \begin{array}{l} K_S(m',m,\mu_S) \\ \times \tilde{K}_E(R_M,m',\varepsilon) \end{array} \right) \\ \times \sum_{o} \left( \begin{array}{l} \tilde{K}_G(o,m,f,\mu) \\ \times \sum_{o'} \left[ \begin{array}{l} K_S(o',o,\mu_S) \\ \times \tilde{K}_E(R_O,o',\varepsilon) \end{array} \right] \end{array} \right) \end{array} \right] \tag{75}$$

The second extension shown in Figure 2-3 is the inclusion of multiple offspring zygotes, from 1 to $z$. This represents multiple full-siblings within the nuclear family. In accordance with Mendelian inheritance laws, each zygote is formed independently of the others has a 50% probability of sharing the same allele from either parent with another full sibling before mutation. Incorporating multiple zygotes in the model amounts to multiplying the results of multiple sums over the possible offspring zygote genotypes, each with an included somatic component and germline mutation kernel, as shown in equation (76).

$$P(R \mid \Theta) = \sum_{m,f} K_P(m,f,\theta) \times \left[ \begin{array}{l} \sum_{f_1'} \left( \begin{array}{l} K_S(f_1',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_1',\varepsilon) \end{array} \right) \\ \times \sum_{f_2'} \left( \begin{array}{l} K_S(f_2',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_2',\varepsilon) \end{array} \right) \\ \times \sum_{m'} \left( \begin{array}{l} K_S(m',m,\mu_S) \\ \times \tilde{K}_E(R_M,m',\varepsilon) \end{array} \right) \\ \times \prod_{j=1}^{z} \sum_{oj} \left( \begin{array}{l} \tilde{K}_G(o_j,m,f,\mu) \\ \times \sum_{o'} \left[ \begin{array}{l} K_S(o_j',o_j,\mu_S) \\ \times \tilde{K}_E(R_{Oj},o_j',\varepsilon) \end{array} \right] \end{array} \right) \end{array} \right] \quad (76)$$

The third extension is the inclusion of monozygotic twins (or *t*-tuplets). Multiple offspring from a single zygote mathematically are represented identically to multiple somatic samplings from a single individual. However the important distinction to be made between the two cases is the most recent common ancestor cell or cells to the monozygotic offspring samples is by definition within the germline, and therefore they are assumed to share no somatic mutations. Equation (77) completes the model extension to include all elements of the pedigree shown in Figure 2-3.

$$P(R \mid \Theta) = \sum_{m,f} K_P(m,f,\theta) \times \left[ \begin{array}{l} \sum_{f_1'} \left( \begin{array}{l} K_S(f_1',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_1',\varepsilon) \end{array} \right) \\ \times \sum_{f_2'} \left( \begin{array}{l} K_S(f_2',f,\mu_S) \\ \times \tilde{K}_E(R_F,f_2',\varepsilon) \end{array} \right) \\ \times \sum_{m'} \left( \begin{array}{l} K_S(m',m,\mu_S) \\ \times \tilde{K}_E(R_M,m',\varepsilon) \end{array} \right) \\ \times \sum_{o_1} \left( \begin{array}{l} \tilde{K}_G(o_1,m,f,\mu) \\ \times \prod_{j=1}^{t} \sum_{o_{1j}'} \left[ \begin{array}{l} K_S(o_{1j}',o_{1j},\mu_S) \\ \times \tilde{K}_E(R_{O1j},o_{1j}',\varepsilon) \end{array} \right] \end{array} \right) \\ \\ \times \prod_{j=2}^{z} \sum_{oj} \left( \begin{array}{l} \tilde{K}_G(o_j,m,f,\mu) \\ \times \sum_{o'} \left[ \begin{array}{l} K_S(o_j',o_j,\mu_S) \\ \times \tilde{K}_E(R_{Oj},o_j',\varepsilon) \end{array} \right] \end{array} \right) \end{array} \right] \quad (77)$$

*Distinguishing sources of mutation using various pedigree structures*

When searching for *de novo* mutations through resequencing parents and offspring, it is difficult to discern mutations that have taken place in the soma of any sequenced family member from those that have occurred in the germline. This is an important distinction since only germline mutations are passed from generation to generation and can therefore respond to the forces of natural selection as the species evolves. Similarly, an estimate of the per-generation mutation rate in humans must depend on counts of germline mutations only.

A fundamental rule can be adopted to determine whether the data is sufficient for disentangling somatic and germline mutations: amongst the samples taken from the family

members there must be at least three alleles that are copies of the same ancestral allele across at least one generational boundary. Such alleles are called identical by descent (IBD). Figure 2-4 shows examples of alleles that are IBD.

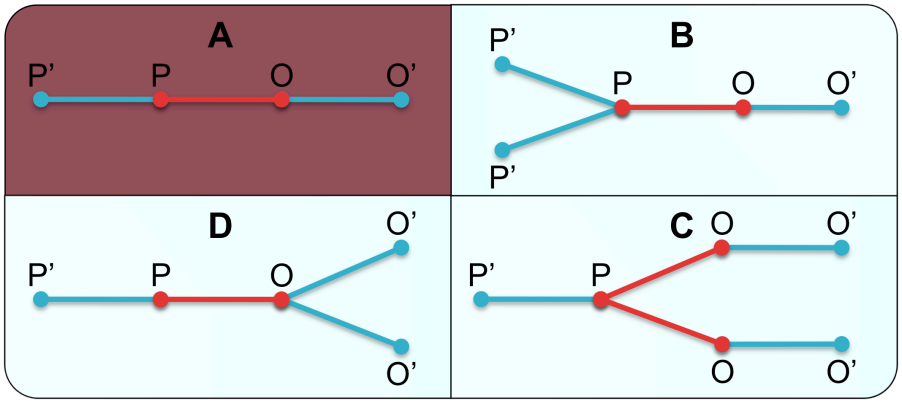



**Figure 2-4.** Non-identifiable and Identifiable Un-rooted Pedigrees. The pedigree of a single allele is shown in each quadrant, with nodes representing the state of the allele at a genotype and edges representing transitions that may bear a mutation. P' and O' are the sampled somatic allele states for the parent and offspring; P and O are the germline allele states. The red edges represent germline transitions, while the blue edges are somatic transitions. In order to be identifiable, the pedigree must terminate with three observable alleles that are identical by descent, assuming all somatic branches are independent given the germline node genotype. Working from the upper left quadrant clockwise, the darkly colored quadrant (A) contains a pedigree in which mutations are non-identifiable as specifically germ or somatic; the next (B) contains two somatic samplings of the parental allele so there are three somatic and one germline mutation branch that are identifiable; next (C) there are two offspring so there are three somatic and two germline mutations identifiable; finally (D) there is either two somatic samplings of a single offspring or samplings of monozygotic twins, with three somatic branches and one germline mutation branch identifiable.

Pedigree A is non-identifiable since a mutations occurring on different segments have identical effects: causing different allele states at *P*' and *O*'. However, Pedigree B is identifiable since any mutation occurring on either *P*-*P*' branch will be seen in only the

affected *P*' allele state, and not observed in the *O*' or alternate *P*' nodes. If many such sites are examined, the number of events occurring on a single of these somatic *P*-*P*' branches can be used to infer a somatic mutation rate; subtracting this value from the rate at which mutations occur on the combined *P*-*O*-*O*' branch will yield an estimate of the germline mutation rate. Similarly, in Pedigree C when both *O*' nodes have the same allele state which differs from the allele state at *P*', a somatic mutation on the *P*-*P*' branch is the most likely explanation. Subtracting the rate at which this occurs from the rate at which mutations arrive on either *O*' branch will again yield a germline mutation rate estimate.

When striving to fully distinguish between somatic and germline mutations some study designs are more efficient than others. Efficiency in this context refers to the proportion of identifiable single-allele pedigrees included in the study relative to the total number of samples sequenced. Samples could either be taken from a single individual or from related family members. Consider the case of a trio study design as depicted in Figure 2-5A. Here there are no identifiable mutations since the fundamental allele IBD pedigree is the same as the pedigree in Figure M 2-4A. Three somatic samplings are needed and no information is provided in distinguishing somatic from germline mutations. While technically there may be an IBD signal between the two parents since they are drawn from the same population, this effect will be slight and provides essentially no power for discerning mutation identity.

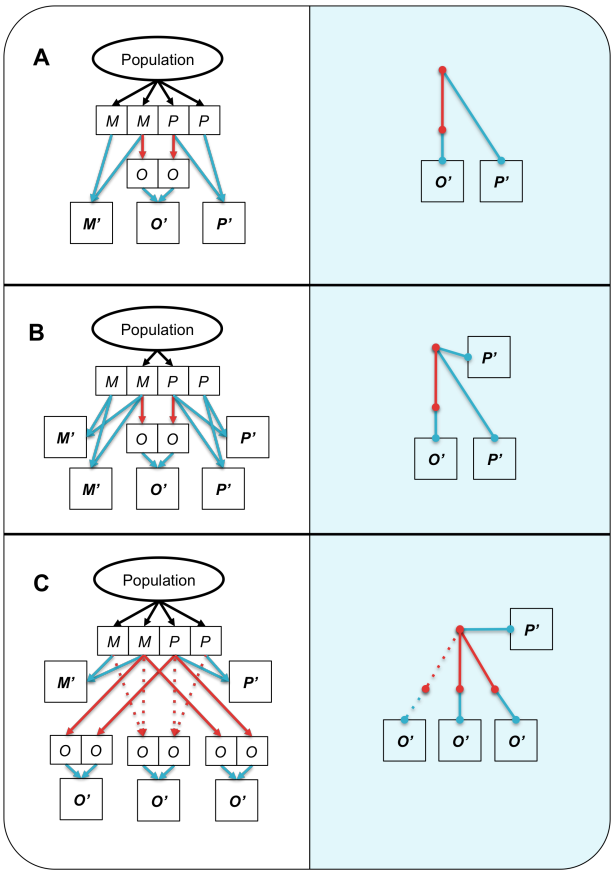



**Figure 2-5.** Family Resequencing study designs with corresponding IBD allele pedigrees. On the left is a simplified tree model of the resequencing study design. In each, four parental alleles are drawn from a population, and arrows point to transmission of these alleles either to somatic samplings or through the germline to offspring. Red lines indicate a germline transmission. The bolded, larger boxes are somatic samplings and are the only observable allele states. A. Trio design: two somatic and one germline mutational branches for three somatic samplings, none identifiable. B. Trio design with two somatic samplings: 3 of 5 somatic and 1 of 1 germline mutational branches identifiable for five somatic samplings. C. Three Sib design: 3.25 of 5 somatic and 2.25 of 3 germline identifiable mutations for five somatic samplings. Since two offspring are guaranteed to inherit the same allele from a single parent, a more precise measure of the germline rate is achievable for the same sampling cost. For one quarter of the sites (dotted line) all of the offspring will inherit the same allele, giving three sources for estimating the germline mutation rate.

By obtaining a second somatic sampling of each parent (Figure 2-5B), the allele pedigree becomes identifiable using five somatic samplings. Computationally, there are 3

of 5 somatic and 1 of 1 germline mutational branches identifiable per parent, giving an overall proportion identifiably of 6:10 (0.60) somatic and 2:2 (1.0) germline. However it is important to notice that obtaining a second somatic sample from a single individual is only specifically identifiable for somatic mutations that occurred after the cell lines diverged. In other words, if a somatic mutation occurred early in development, both cell lines leading to the independent somatic samplings will contain the variant. So, the mutation event will be included with the combined germ-plus-somatic branch, and could lead to an overestimation of the germline mutation rate. Ideally all somatic samples drawn will be separated from the germline genotype by the same number of cell divisions. While this is unlikely in any study given the age difference of the individuals sampled, it is grossly violated when two samples are taken from one individual, while the third is taken from another. However, this effect can be assuaged to some degree by ensuring the samples are taken from different tissues (such as blood vs. cheek swab) whose founder cell lines likely diverged early in development. Assuming this early divergence of somatic cell lines, monozygotic twin studies are identical in structure to obtaining two somatic samples from the child.

In Figure 2-6, a simplified version of the two-sib, three-sib, and grandparent design pedigrees are shown to illustrate the calculation of the power to distinguish germline and somatic mutations for different pedigree types. In terms of strict efficiency, the poorest performing designs are the two-sib and grandparent design. The two-sib design only requires four somatic samples, but is only identifiable in half of the sites sampled. Due to Mendelian inheritance laws, a parent will transfer the same allele to two offspring 50% of

the time. Therefore only half of the sites sampled will contain three IBD alleles. The grandparent design is identifiable for every site, but suffers from requiring seven somatic samples. In this design it is necessary to capture all four grandparental alleles since the grandchild could inherit any one of these four from the parent. The pedigree is identifiable at a site only for that transmitted allele. The two best performers are the three-sib and monozygotic-twin designs. In the three-sib design, every site is identifiable with the added bonus of having 25% of sites identifiable across the entire pedigree when the same allele is inherited by all three offspring, according to Mendelian inheritance laws. This is a large improvement over the two-sib design at the cost of only one extra somatic sample. Monozygotic-twin designs are fully identifiable at every site since they have the same pedigree structure as the trio with second Somatic Sample design. However, they do not suffer from the non-independence of somatic samples, and so will provide a more accurate estimation of germline and somatic mutation rates. Due to the higher number of included, identifiable germline transmissions in the three-sib design which provide more power for capturing germline mutations, it may still be the better choice in design selection provided a third sibling is available and the cost of sampling a fifth individual is not prohibitive.

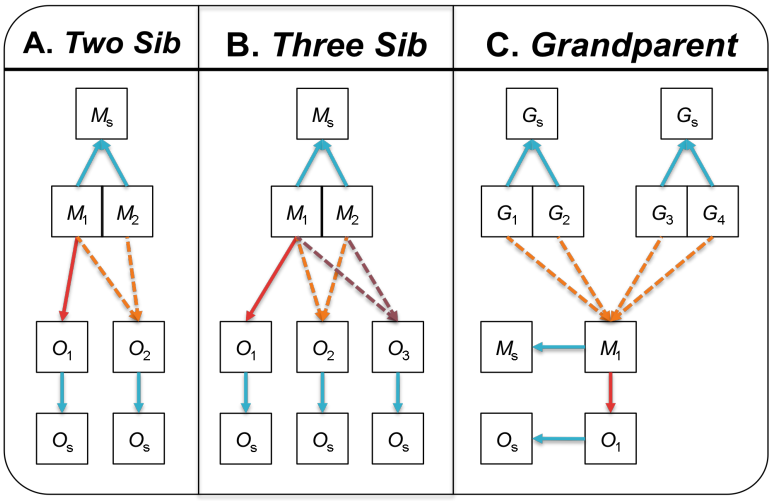



**Figure 2-6.** Mendelian Inheritance of the Two Sib, Three Sib, and Grandparent designs. This figure shows basic possible inheritance pattern of alleles in Two Sib, Three Sib, and Grandparent study designs in panels A, B, and C, respectively. Only a single parent (M) is shown, with the alleles transmitted by that parent to the set of offspring. Dashed lines represent transmissions that are uncertain (only one of each color can occur). In the Two Sib design (A), the allele transferred (red arrow) from the mother to the first offspring is arbitrarily labeled as M1. There is a 50% chance that this same allele is transferred to the second offspring (orange arrows). This pedigree becomes identifiable only when M1 is also transferred to O2, which occurs 50% of the time assuming independence of sites. Therefore a Two Sib design has (.5*2)+(.5*0) = 1 identifiable germline branch per parent, or two in total; there are (.5*3)+(.5*0) = 1.5 identifiable somatic branches per parent, or three in total. When a third offspring is included (B), there is again a 50% chance it will inherit (brown arrows) the same IBD allele as offspring one. Note that regardless of the inheritance pattern, a somatic sampling of three IBD alleles is guaranteed, whether they come from M1, O1, and O2 (red and orange lines), M1, O1, and O3 (red and brown lines), or M2, O2, and O3 (orange and brown lines). Furthermore, 25% of the time all three offspring samplings will be of the same allele, providing more opportunities to capture an identifiable germline mutation. This gives (.75*2)+(.25*3) per parent, or 4.5 total identifiable germline branches and (.75*3)+(.25*4) per parent, or 6.5 total identifiable somatic branches. In panel C, the Grandparent design is shown. Only the single parental allele that is transferred (red line) to the offspring is shown. The parental allele M1 has an equal probability of being inherited (orange lines) from any one of the four grand-parental alleles. Therefore every site has two identifiable germline branches and three identifiable somatic branches per parent, giving a total of four and six identifiable branches respectively. Since the Three Sib design contains the most power for identification while requiring only five somatic samples, it is the most efficient design of the three shown here.

## Method Implementation

The *de novo* discovery method, including the simulation functionality presented below, has been implemented as a single Java 6 application with the capability of being executed singularly on a workstation or distributed on a HPC cluster using additional Perl scripts. The observed data takes the form of discrete counts of observed nucleotides from sequencing reads sampled from all individuals in the study, as shown in Table 2-1 for each site in the genome. A maximum of 255 reads bearing a single nucleotide type are allowed at any site and space is reserved for an additional offspring. The maximum of 255 was chosen as a high-end threshold for depth, as well as being conveniently stored in a single byte ($2^8 = 256$). Therefore the memory required for the observed data at a single site is

$$\frac{1\ byte}{nucleotide} \times \frac{4\ nucleotides}{individual} \times 4\ individuals = 16\ bytes \tag{78}$$

For a three-gigabase genome, this is an inefficient solution since a single study would require 48 gigabytes of RAM, given a maximum depth for a single nucleotide of 255. This requirement is drastically reduced by applying a compression algorithm that, by treating the nucleotide types as columns and individuals as rows, sorts the data by the column sums. This increases the frequency at which a single structure is seen, and the frequency is stored as overhead with each structure type increasing the total size to 48 bytes. Before sorting the structure, this overhead space is used to store genomic position information for the site. However, the compression is quite effective on genomic data, reducing the number of structures to be stored by three orders of magnitude. Therefore the structures for an entire

human genome can effectively be in roughly two to five gigabytes of RAM. The overall effectiveness of the compression is dictated by the variability of structures observed. Cleaner, better aligned, data tends to have fewer structures, although the number of structures will increase with sequencing coverage.

The data compression employed also has the effect of reducing the execution time required, but it requires that the supporting evidence for each observed nucleotide be equal. This summarization of the alignment data does not incorporate data quality, which is a potential source of error when data of poor quality is treated in the same manner as high-quality observations. However, since the model does not treat the nucleotide bases differently, symmetrical calculations arise within the model which allow optimization of the implementation. Specifically, calculations done for a site which has ten A's observed for each individual at a site are identical to the calculations done for ten C's, G's, or T's. Therefore we were able to calculate the summary statistics mentioned above once per *structure* and multiply the results by the number of *sites* characterized by the structure. This reduces the number of computations needed by roughly 3 orders of magnitude. In general a single iteration of the EM algorithm using data from a complete resequenced human genome takes roughly 2 hours to complete on an 8-core 3GHz Intel Zeon server with 32 GB of memory, so the time of execution for convergence depends on the number of iterations required which can range from around five to over thirty. The criteria for convergence is met once the proportional difference of all parameter estimates between consecutive iterations was less than $1x10^{-6}$.

In addition to the compression algorithm benefits, the implementation takes advantage of symmetrical calculations resulting from the assumed Jukes-Cantor substitution model. Specifically, because nucleotides are assumed to have equal frequencies and may equally mutate to any other nucleotide, the values of the kernels are pre-calculated and a simple lookup is performed when comparing genotype pairs, such as the somatic and germline genotype comparisons done within the Somatic mutation kernel defined in equation (23). The kernel can be calculated a mere three times per parameter value regardless of the specific alleles being compared: for genotype comparisons that have zero, one, or two allelic matches. Furthermore by making the assumption that reads are equally likely to sample either chromosome in the Error kernel and setting $p$ to .5 in equation (31), the number of possible genotypes summed over and the number of calculations in total is cut geometrically, from $\sim 16^i$ to $\sim 10^i$ where $i$ is the number of biallelic genotype nodes in the pedigree, since the probability of heterozygous genotype is identical, regardless of the order of alleles (i.e. AC versus CA). While this allowance may expose the method to a small possibility of false inferences when severe reference bias exists, the tradeoff for execution speed, over an order-of-magnitude improvement with a trio study design, is arguably warranted.

The procedure developed to predict *de novo* mutations and estimate the model parameters using aligned read data from resequenced nuclear families includes four primary steps. The first step is to transform the multiple sequence alignment of the aligned reads from all family members into a compressed file which has a minimum of one data

structure shown in Table 2-1 per site which has aligned data from all individuals. At this point various quality filters can be applied or entire regions of the genome can be ignored depending on the scope of the search for *de novo* mutations. Two files are created at this point: a binary file containing the compressed data, and a text file which lists every site incorporated into the compressed data as including the chromosome, the chromosome position, the family type, a unique numerical identification number which ties the particular site to a specific data structure within the compressed data file. The second step is the execution of the EM algorithm to create maximum likelihood estimates of the model parameters from the compressed data. This step is optional since a researcher may already have access to expert parameter values to apply to the data set at hand. Third, the EM-generated or the expert parameter set are used to calculate $\delta$, using equation (62), and this value as well as the expected values of the summary statistics are reported for each data structure examined. Finally, a minimum threshold value for $\delta$ can be applied, and data structures with corresponding $\delta$ values greater than this threshold are reported along with the specific genomic coordinates of the original site examined.

## Simulation Studies

A series of three simulation studies were carried out to investigate the performance of the method when applied to high throughput sequencing data. The first examines the effect various read structures at a single site has upon the inferences of the model; whether the model predicts a mutation or sequencing error given a certain pattern of observed read

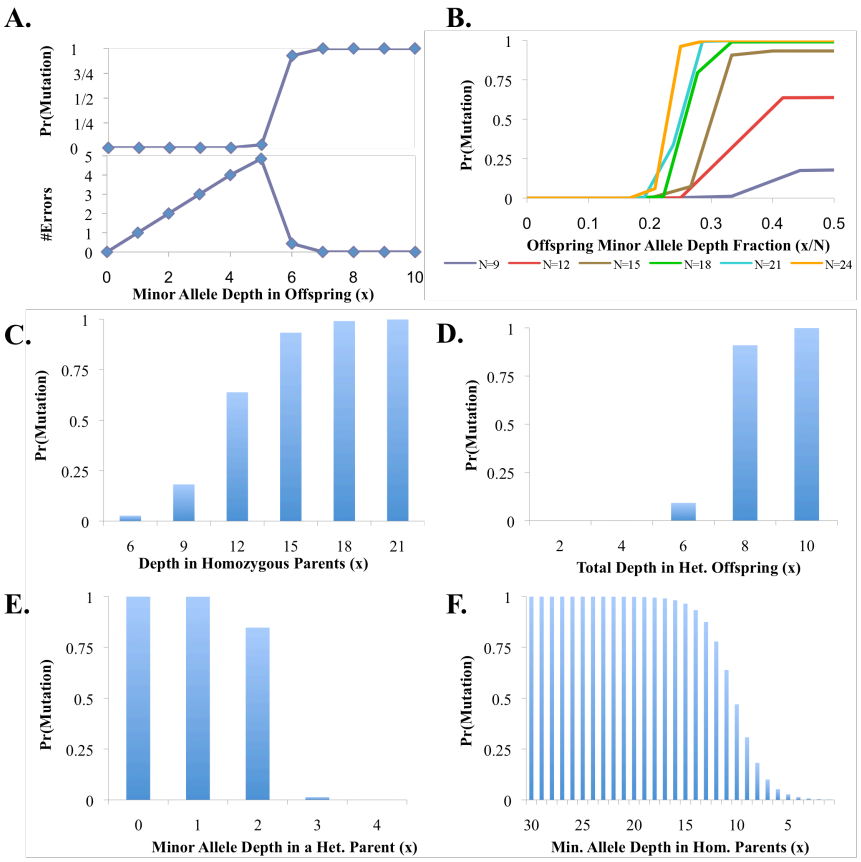
nucleotides often depends on merely a few reads. The second study considers the model performance using a trio pedigree. As mentioned in previous sections the distinction between somatic and germline status of a spontaneous mutation can not be made using data from a trio, so this simulation set ignores the possibility of somatic mutation. The third study includes somatic mutation by simulating a pedigree containing monozygotic twins. The ability of the method to accurately estimate the simulated model parameters and to reliably predict the mutations placed within the simulated data set are evaluated at varying coverage levels and with different magnitudes of alignment error.

### *Single-site simulations*

To illustrate the performance of the model, it is useful to examine the data points at which the inferences made about a site transitions from predicting Mendelian inheritance or sequencing errors to a predicting a mutation event. These points are demonstrated in Figure 2-7. For each panel A. through F., a range of read structures was tested with the model, producing the posterior probability of mutation for each. The range of structures tested is given in the table below the graphs, where each row represents the read data taken for a family member from a site, with the offspring data as the bottom row, and the columns represents the number of reads observed bearing an 'A', 'C', 'G', or 'T' allele. The individual whose observed data varies in the graph is bolded.

Multiple observations of more than one allele suggest a heterozygous genotype. When this pattern is caused by multiple sequencing errors stacking on a single allele in an

offspring when both parents are homozygous, one can be mislead to call a heterozygous genotype, and a mutation may be inferred (Figure 2-7A). Misalignment of short sequencing reads to the reference is another cause of this "minor-allele stacking." Given equal weightings to allele observations originating from correctly and incorrectly aligned reads, it is difficult to distinguish between a site bearing a mutation and one covered by many misaligned reads. One potential method is to make an assumption of the total depth expected at a site. If two alleles are observed among approximately this expected number of reads, the argument could be made the site bears a mutation, while two alleles observed among twice this number of reads would indicate mapping errors have occurred. However, with variable depth throughout a genome, it is difficult to accurately make any such expectation. Alternatively, mapping qualities such as those created by the MAQ alignment tool can be used to weight read-allele observations to further assist in making this distinction.




| | | |
|---|---|---|
| **A.** 25 00 00 00<br>25 00 00 00<br>**25-x x 00 00** | **B.** N 00 00 00<br>N 00 00 00<br>**N-x x 00 00** | **C.** **x 00 00 00**<br>**x 00 00 00**<br>15 15 00 00 |
| **D.** 30 00 00 00<br>30 00 00 00<br>**x/2 x/2 00 00** | **E.** **30-x x 00 00**<br>30 00 00 00<br>05 05 00 00 | **F.** **x 00 00 00**<br>30 00 00 00<br>05 05 00 00 |

**Figure 2-7.** Transition points in model inferences. For each graph, the corresponding data structures are given in the table. In every structure tested the following parameter values were used: $\theta = 0.001$, $\varepsilon = 0.01$, $\mu = 2.0\text{e-}7$, and $\mu_S = 0.0$: (A) transition from error to *de novo* mutation, (B) effect of depth on error-mutation transition, (C) effect of low parent depth, (D) effect of low offspring depth, (E) transition from *de novo* mutation to inherited minor allele, (F) transition from *de novo* mutation to inferred inheritance of minor allele.

The transition point from error inference to mutation inference shifts depending on the overall depth observed at the site. As shown in Figure 2-7B, the required ratio of minor-to-total allele depth for calling a mutation decreases as the depth increases. We examined depths up to 50, where the mutation probability becomes 1.0 with 9 of 50 (0.18) reads in the offspring bearing the mutant allele. As with all transition points, this value depends on the model parameters used, including mutation and error rates. The model at this point is largely affected by the binomial probability of observing a single chromosome a limited number of times relative to the number of samples taken. With low depth levels, such as 15x and below as shown in the figure, the model is unable to conclusively call a mutation due to the low parental coverage.

The effect of low parental read depths is shown in Figure 2-7C. In this panel each structure tested has a clear signal of heterozygosis in the child with 15 reads observed of both alleles. The parental depths vary around the transition point at which the model infers heterozygosity in either parent. In order to be confident in the call of a mutation, both parents appear to require a read depth of at least 15 in the non-mutant allele. As Figure 2-7D shows, much less coverage (~8x) is needed in the offspring provided both parents are clearly homozygotic and the two offspring alleles are equally sampled by the sequencing reads. This is an important fact to consider when designing mutation discovery experiments; it is more advantageous to have high sequencing depth in parents to be certain of their genotypes than to have deep sequencing in the offspring at the expense of parental coverage. One should also note that this is not simply an effect of sequencing a second

individual at a high depth. The same effect would not be observed if a single parent plus the child are sequenced at depth, since the genotypes of the low-coverage parent will remain uncertain.

Figures 2-7E and 2-7F demonstrate the transition from predicting a mutation to inferring the inheritance of a minor allele. In both panels, heterozygosity has been established in the offspring with five reads of two alleles each. A single parent is fixed as a homozygote with thirty reads bearing a single allele. In Figure 2-7E this transition happens quickly, but not immediately, as mutant alleles stack in the parent. In Figure 2-7F this transition more slowly occurs as the depth of the parental major allele decreases. This metric, the minimum depth of a parental allele, proves to be one of the key factors in locating mutations in our simulation studies.

### *Trio simulation study*

Resequencing of a nuclear family trio was simulated to test the ability of our algorithm to estimate the parameters of our model and discern the location of actual *de novo* events. Using the human chromosome 10 primary reference assembly (NCBI Reference Sequence NC_000010), two parental genomes were created, allowing up to three segregating alleles per site, with configurations in proportions expected from coalescent theory given the per-site population mutation parameter, $\theta = .001$ (see the population kernel section above). Next, each diploid parental genome was used to generate a single

haploid sequences which were combined to form a diploid offspring genome, allowing for germline mutations at per-site, per-generation rate $\mu = 1 \times 10^{-6}$. Sequencing reads of an average of 35 bases each were generated from the diploid genomes of all three individuals, inserting sequencing error events at the per-base rate of $\varepsilon = .01$. Because the family structure simulated was a trio, somatic mutations were not included because they would be indistinguishable from germline mutations as mentioned earlier. To complete the simulated data set, the reads were aligned back to the reference using the BWA program [34] and pileup files were created. BWA was chosen due to its speed for aligning short reads with minimal loss in alignment accuracy (Heng Li, personal correspondence). For each of the simulation replicates, consistently 99.45% of the reads simulated were aligned back to the reference; 91.10% of the reads were aligned correctly. Additionally the average and variance of the coverage levels for each individual in each replicate were approximately equal to 20. Finally, the EM algorithm was used to estimate the parameters from the aligned data, and, using the resulting parameter MLE's, a list was generated of candidate *de novo* mutation sites. Ten replicates were done at the 20x coverage level to illustrate the spread of our parameter estimates around the true value. Three different sequencing coverage levels, 10x, 20x, and 30x, were used to examine the effect average coverage has on the method's ability to recover the simulated *de novo* mutations. In addition a single replicate was done at each coverage level that included no alignment step in order to investigate the role of mapping error in the simulations including BWA alignments. Every read in these replicates was placed exactly at the genomic position from which it was created.

The results for the EM parameter estimation within the ten 20x replicate simulations are given in Figure 2-8. Across these replicates, the estimated mutation rate is consistently higher than the simulated rate, but also consistently varies with the simulated rate around the True Rate indicated by the green line. The variance around the True Rate was much less in the simulated sequence error rate and population mutation rate.

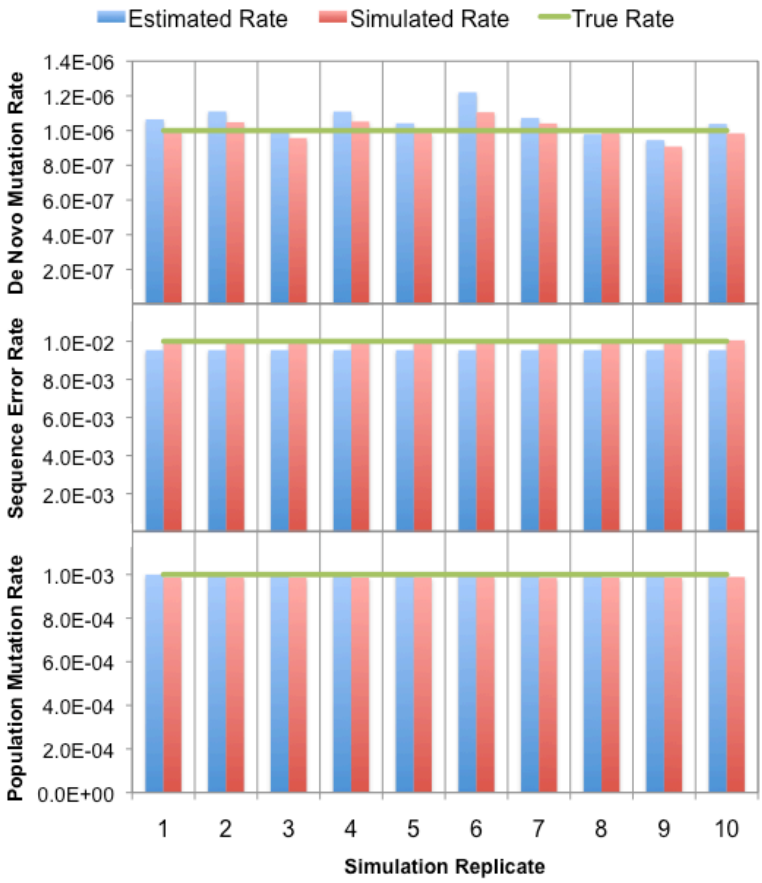



**Figure 2-8.** Parameter estimate results of ten replicate *de novo* discovery simulations with a mean 20x coverage in each family member of a trio pedigree.

The sequencing error rate is consistently under-estimated while the simulated population mutation rate is accurately recovered. Both the under-estimate of the sequencing error rate and the over-estimate of the *de novo* mutation rate is potentially the result of alignment error. Sequencing errors may prevent a read from being aligned to any region of the reference, so the data supporting the correct, higher error rate becomes lost. Alternatively several reads bearing the same mismatching allele for a single site can be misaligned in tandem, creating a signal that is indistinguishable from a mutation by supporting a genotype call that is consistent with Mendelian error at the site, inflating the *de novo* mutation rate. However, these simulations show that with a coverage level of 20x the EM algorithm produces accurate rate estimates in the presence of a small amount of alignment error.

I next sought to investigate the effect different levels of sequencing coverage had on the EM ability to estimate model rate parameters. A single family was simulated and reads from the family were produced in three independent sets: for coverage levels of 10x, 20x, and 30x. The EM-produced parameter estimates for these simulations are presented in Figure 2-9. Unsurprisingly, the estimates get progressively closer to the simulated rate as the coverage increases. Consistent with the replicates above, the estimate results at the 20x coverage level shows an overestimate of the *de novo* mutation rate and an under-estimate of the sequencing error rate.

To test the role which alignment error is playing in the simulations, the reads were aligned without error to the original reference. The resulting alignments were treated as if

they were produced by BWA and the *de novo* discovery method was applied. The results of these "control" EM parameter estimates are given in Figure 2-10. For the *de novo* mutation rate, the result is a consistent underestimate of the true rate. This is expected for two reasons. First, as stated in the introduction, some mutations will not cause a Mendelian error so will be undetectable by any method relying on the identification of Mendel errors to locate mutations. Second, a source of uncertainty remains even with perfectly aligned data in the sampling of the haplotypes by randomly generated reads.

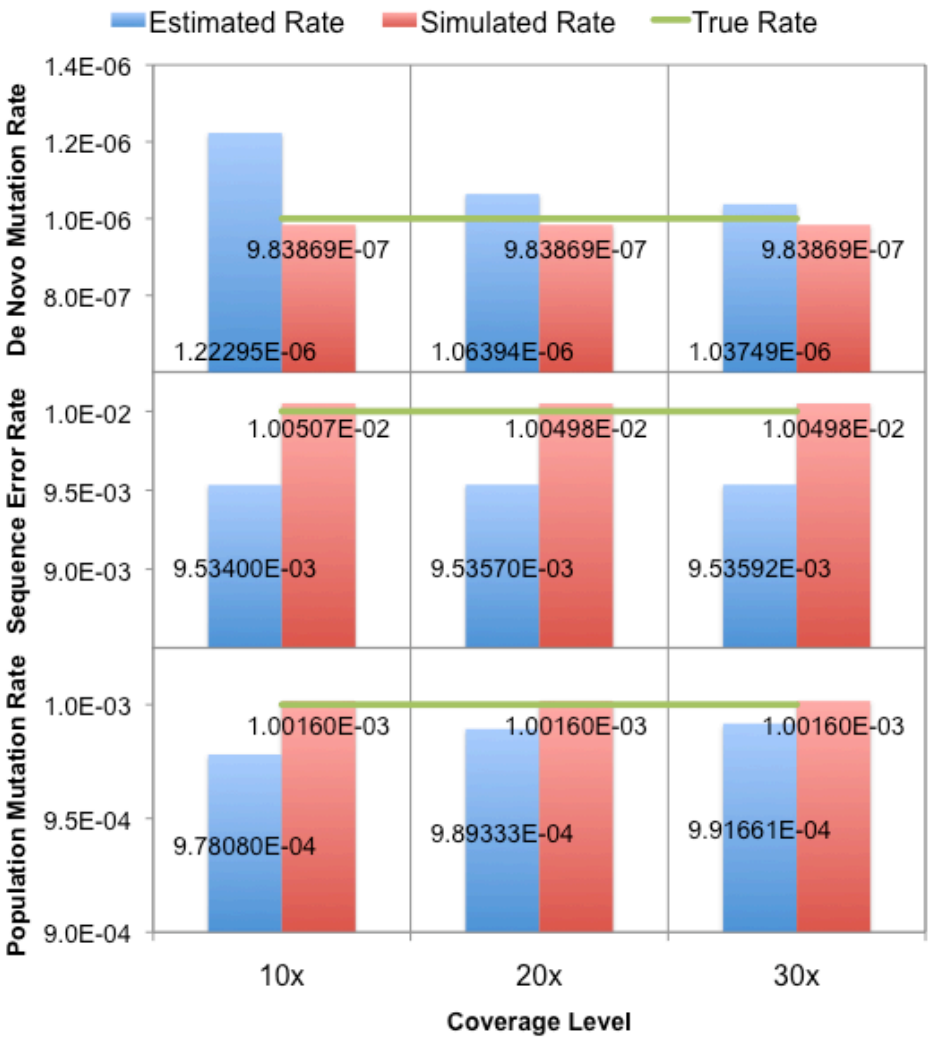



**Figure 2-9.** Variable coverage trio simulation parameter estimation results. A single family was simulated, and three independent sets of reads were created, one for each coverage level 10x, 20x, and 30x.

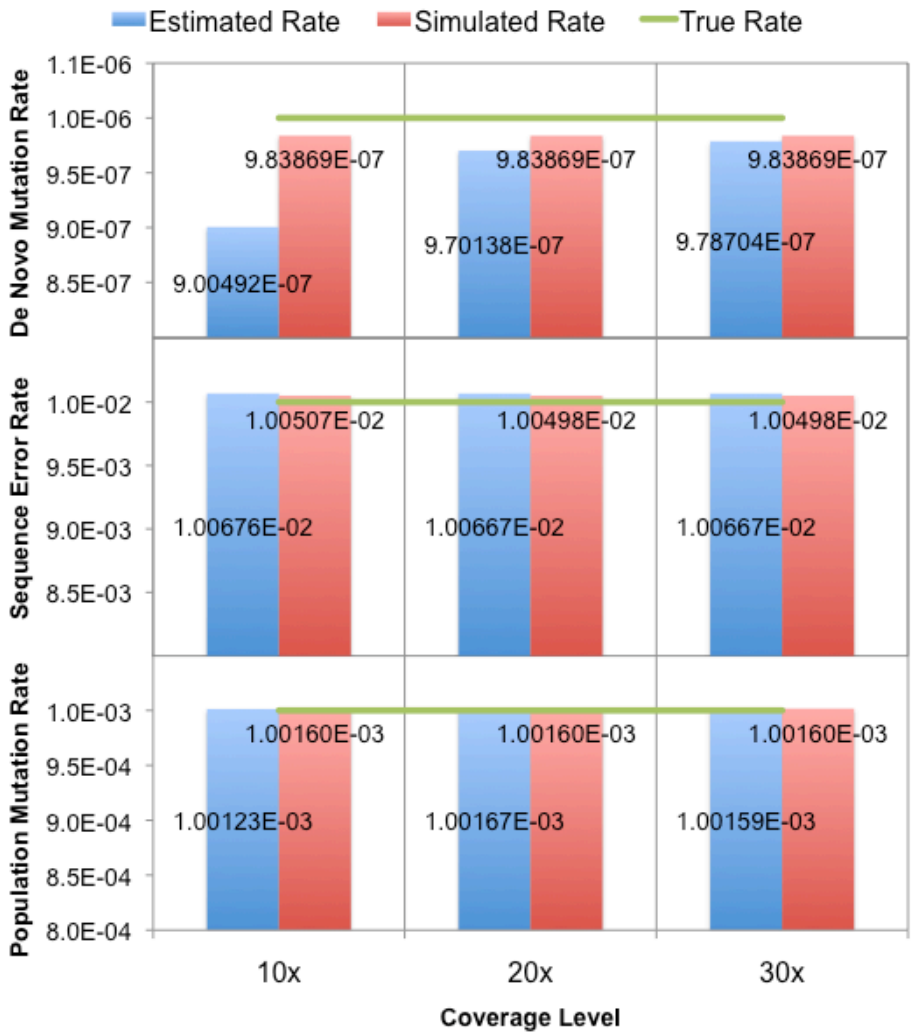



**Figure 2-10.** Variable-coverage, no alignment error trio simulations parameter estimation results. The reads simulated from each set were aligned without error on the reference. The only remaining source of uncertainty is the variable sampling of the simulated genomes by the reads. At the 30x coverage level even this effect is largely lost.

Some heterozygous mutation sites will not have complete coverage of both alleles; indeed some mutation sites may not be covered at all in one or another individual in the family. Of course this effect will be mitigated by increased coverage, as evidenced by these simulations; at the 30x level the under-estimate bias is less than .5% of the simulated rate. The biases in the other rate-parameter estimates are virtually eliminated by removal of alignment error, even at the 10x coverage level. Even with poor coverage the method is able to extract enough information in the perfectly aligned data to reconstruct the parental genotypes, leading to accurate estimation of the population mutation rate.

Another way to evaluate the method's performance is by examining the number of actual mutations relative to the number of false mutations that were predicted. Traditionally these are referred to as "true positives" and "false positives," respectively. A "false negative" is a mutation that was not predicted to be one by our method. The method predicts a mutation by ranking every site examined by the probability $\delta$ of a mutation event as calculated in equation (62) given the read data at the site and the estimated parameter rates. One can set a cutoff for $\delta$ and consider every site above that threshold a candidate site for further investigation. In this application our model acts as a binary classifier, assigning the data at a site to one of two groups: mutant site or no mutant site. To test the utility of $\delta$ as a classifier, we produced Receiver Operating Characteristic (ROC) curves based on every simulation performed[35]. ROC curves are plots of the increasing rate of false positive discovery against the true positive fraction. The perfect classifier would be represented as horizontal line at the top of the graph, having a high

specificity level without sacrificing sensitivity. A random, and therefore useless, classifier would show as a bisecting diagonal line, where the true positive rate and false positive rates are increasing at equal rates. At points where the ROC curve is beneath this diagonal it indicates a worse-than-random performance, while points where the ROC curve indicated improved performance. Therefore one useful metric of a ROC curve is the area under the curve (AUC); the ideal classifier will have an AUC of 1.0. Figure 2-11 shows the ROC curve of each simulation, with the ten replicates of 20x coverage in black above the colored variable coverage and control simulations, as well as the AUC for each. A dashed dark-green line is shown along the 'random' diagonal. Some important conclusions can be drawn from examining the ROC curves and their respective AUC values. First it is clear that at the 20x coverage level the method performs admirably, with a mean AUC of 0.92. Secondly, at the 10x coverage level removing alignment error doesn't seem to substantially improve performance. As shown earlier, with only 10x coverage it is difficult to accurately call heterozygote genotypes; therefore in this scenario it is better to increase coverage than increase alignment accuracy. However at high coverage levels, having perfect alignments creates a near perfect predictor in our method, given the AUC for the 30x control simulation of 0.996.

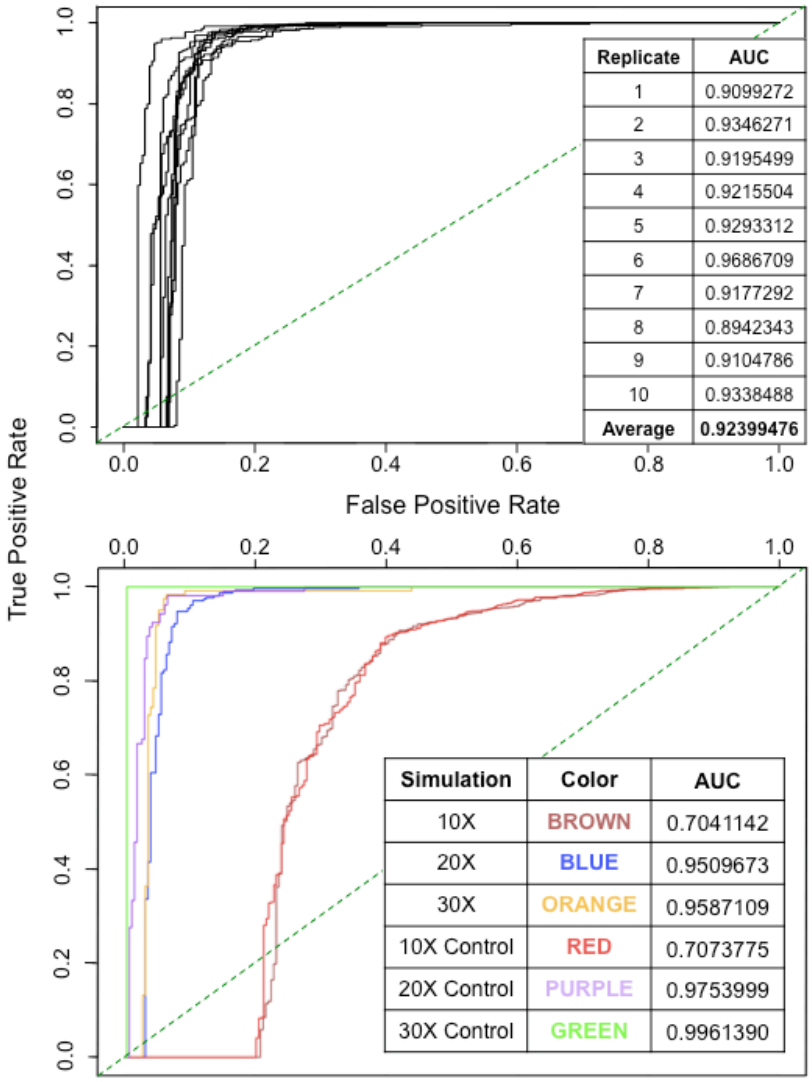



**Figure 2-11.** Trio Simulation ROC curves.

Figure 2-12 presents the number of actual mutations predicted (true positives) at three different threshold values of $\delta$, relative to the total number of predicted sites and total number of true mutant sites. At 20x coverage, a threshold of 0.9 has very few false positive calls (indicated by the small distances between the blue dashes and triangles), but also a

higher false negative rate (seen in the distance between the blue triangles and the top of the tan columns). Conversely, a threshold of 0.1 grants many more false positives and very few false negatives. By selecting an intermediate threshold, these values can both be equalized. When the validation cost of any given candidate site is low, it may be more desirable to capture all of the true mutations at the expense of capturing non-mutant sites, so selecting a lower threshold would be the most appropriate choice. This is a feature of the method that will be applied to real data sets in the upcoming chapters in order to capture as many potential *de novo* mutation sites as possible. The bottom section of Figure 2-12 clearly shows both the disadvantages of having low coverage (10x) and the advantages of improving alignment quality at higher coverage levels. At the 30x coverage level, there is little distinction between the results for the three threshold levels, suggesting a polarization of the $\delta$ values for true and false mutations. Again, eliminating alignment error at the 30x level creates a situation in which the method performs very well.

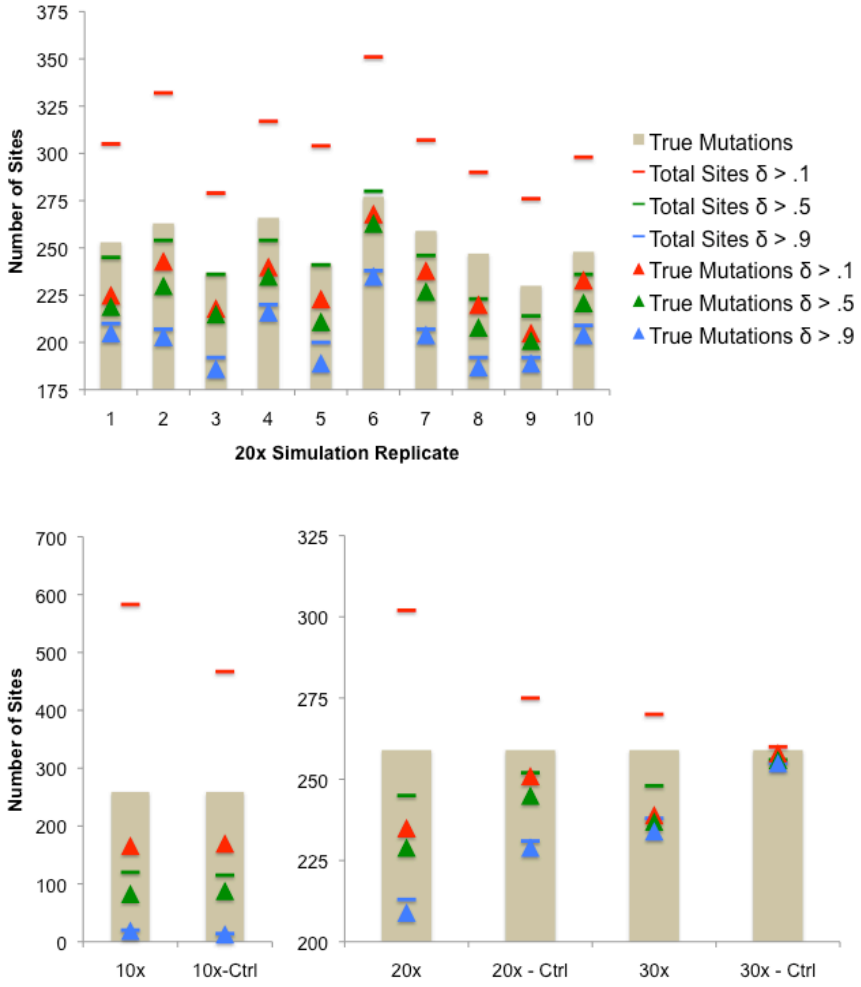



**Figure 2-12.** True positive, false positive, and false negative *de novo* mutations predictions from trio simulations. Each tan column shows the number of true de novo mutations simulated in each simulation run, identified by the label on the horizontal axis. For the columns labeled as 'Ctrl', the entire set of simulated reads were aligned back to the reference without error. The total number of positive mutation calls at three different levels of $\delta$ (.1 shown in red, .5 in green, .9 in blue) are given by the horizontal dashes for each simulation. The triangles indicate the number of true mutation sites with $\delta$ greater than the specified level, representing the amount of true positive mutations at that level. The distance between the dash and triangle represents the amount of false positives, and the difference between the triangle and the top of the tan column gives the number of false negatives. The panel at the top of the figure presents the ten replicate simulations at the 20x coverage level, and the bottom panel displays the variable coverage simulations.

*Monozygotic twin simulation study*

Following a procedure similar to that used in the trio simulation study, the performance of the *de novo* mutation discovery and rate estimation method was evaluated using a slightly more complicated, but also more informative pedigree. A trio pedigree structure is uninformative with regards to disentangling somatic and germline mutations because it lacks a third allele that is IBD with two others. A family with two offspring, or a quartet of samples, provides an opportunity for simultaneous estimation of the germline mutation rate without the confounding effects of somatic mutations. In general, when a genotype is differentially called between monozygotic twins, a somatic mutation must have occurred. Alternatively, when both offspring carry the same mutant allele that is unobserved in either parent, then it is likely due to a germline mutation in one of the parents. Note that due to the allowance of three segregating alleles in the parental zygotic genotypes, any somatic mutation in the parents will likely be missed, particularly if the presence of the mutant allele does not indicate a Mendelian error has occurred.

Two strategies for simulating a monozygotic-twin pedigree were taken: in one the true somatic mutation rate was set to be one-half the germline mutation, and in the other the somatic rate is several times higher than the germline rate. With the events occurring at different rates, the ability of the estimation method to specifically call one event or the other will be tested. This effect is increased in the simulation with the higher somatic rate when germline events occur at one fifth the rate of somatic mutations. This situation is the

more biologically relevant because within the span of one generation, depending on the ages of the donor individuals, many more somatic cell divisions have taken place than germline divisions, increasing the chances for mutation. In addition, in medical studies of cancer, it is often the high number of somatic mutations which are of interest within the leukemia sample, which must be distinguished from inherited *de novo* variants.

In the first set of mutations the somatic mutation rate was set to $0.50 \times 10^{-6}$, half the germline rate of $1.0 \times 10^{-6}$. These elevated rates were chosen to guarantee several mutations would occur across a simulated genome that is only the size of a single chromosome. Five full replicate simulations were done at this level following the same procedure as the trio simulation studies. The germline and somatic mutation rate estimates for these five replicates are shown in Figure 2-13. In the second set, shown in Figure 2-14, the somatic mutation rate was five times higher than the germline rate: $2.5 \times 10^{-6}$ and $5.0 \times 10^{-7}$. For both sets of mutations the somatic rate was neither consistently under-estimated or over-estimated, but varied with the simulated rate as expected. The germline rate estimate, however, was once again consistently overestimated as with the trio simulations. Both the error rate estimates and population mutation rate estimates were recovered with the same accuracy as seen in the trio simulations, suggesting that extension to a larger pedigree had no effect on the ability to estimate these rates.

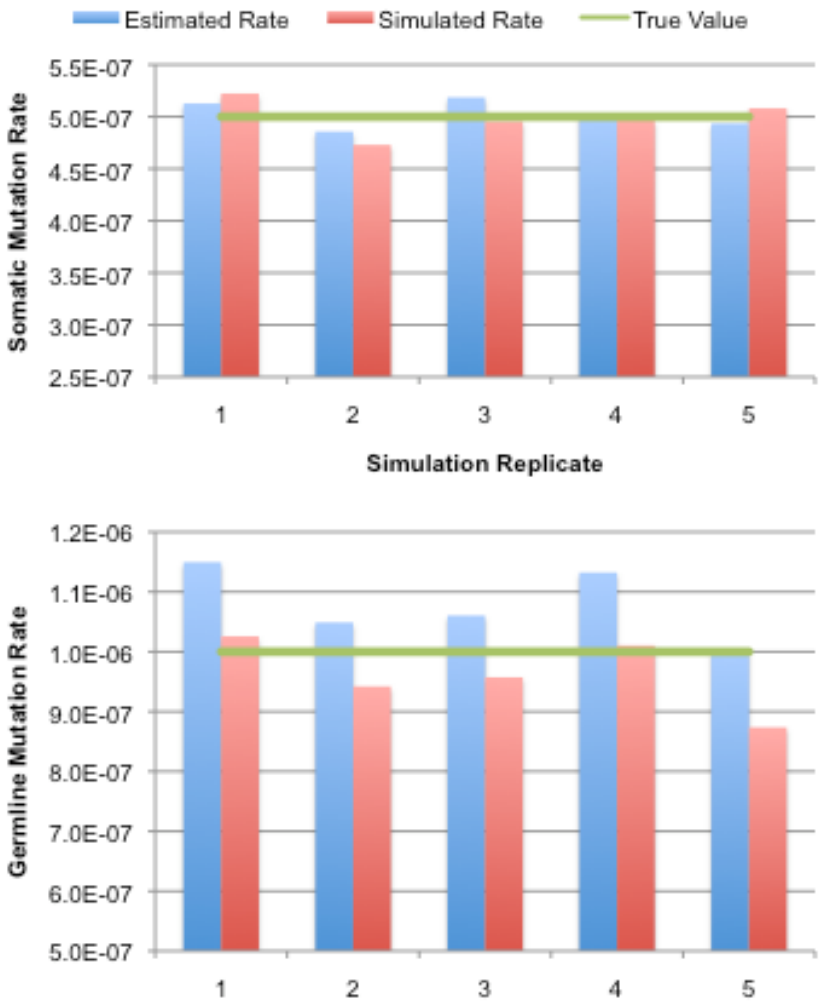



**Figure 2-13.** Rate estimates results for the first set of quartet simulation replicates.

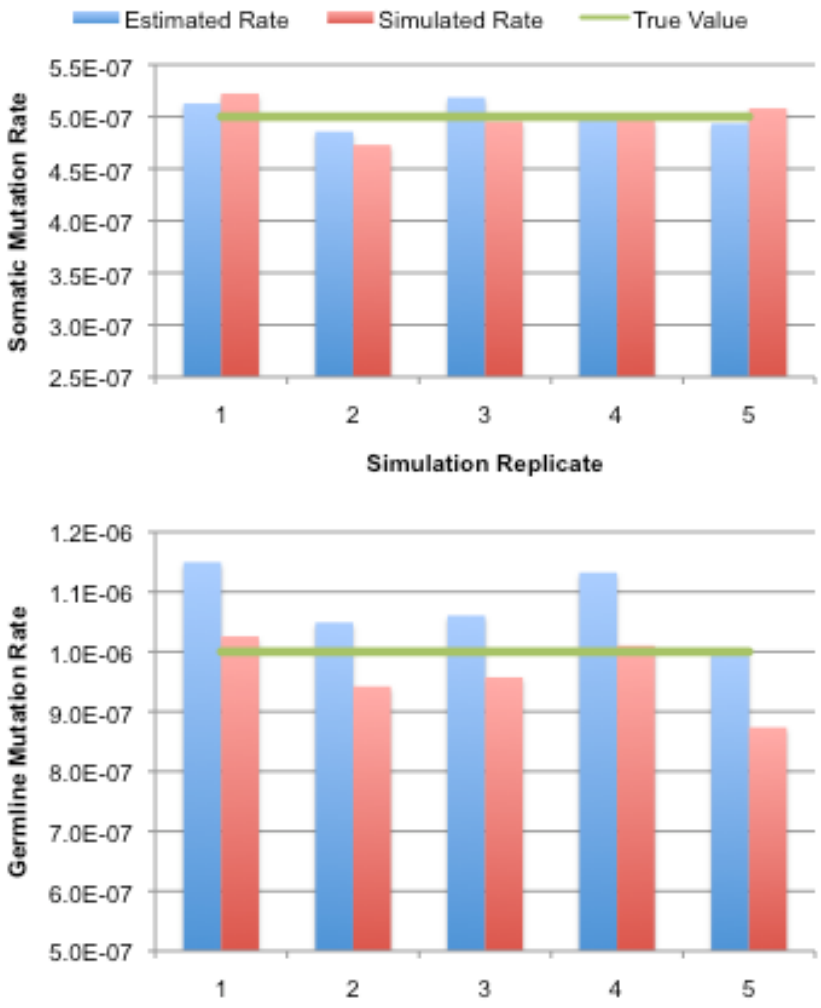



**Figure 2-14.** Rate estimates for the second set of quartet simulation replicates.

## Conclusion and Future Work

The simulation results presented indicate that both the parameter estimation and mutation prediction procedures within this method produce accurate results particularly at high coverage levels and with minimal alignment error. As alignment error increases the expectation is the sequencing error rate estimate will decrease as the mutation rates increase. Also as predicted the method is able to distinguish germline from somatic mutations in an automated manner when applied to data from an appropriate pedigree. At low coverage levels (~10x) the method has particular trouble in distinguishing true heterozygous sites from those affected by sequencing or alignment error. With the current state of high throughput sequencing where high error rates and reference-based alignments are the norm, making incorrect inferences about *de novo* mutations are unavoidable with low coverage levels.

This method can be improved through future work, both directed toward the mathematical model and to the software implementation. Within the model, incorporating mapping qualities and base qualities may improve the ability to distinguish errors from true sampling of novel alleles. Transitioning the fundamental data structure from a discrete count of observed alleles to continuous weighted average based upon these qualities is a possible solution however it would affect the data compression efficiency and therefore the performance time of the method as a whole. Extending the model to other forms of genetic variation would make the method more biologically and medically relevant as knowledge of the role structural variants plays in human disease phenotypes gains increasing interest.

Doing so would require including data from multiple sites compressed into a single data structure from which detection of Mendelian error can be done in an automated fashion. Finally it would be helpful to incorporate a mechanism for detecting and correcting for the role alignment error plays in the overestimation of mutation rates.

In terms of the Java software implementation, the method would gain significant speed improvements by being ported to C++. More direct access to raw data structures, optimized compiling, and the removal of a currently required virtual machine would all be ways in which performance could improve by moving away from the Java platform. If the speed gains were large enough, it may be possible to remove some of the efficiency-based assumptions that are currently made such as the equal sampling of heterozygous alleles and the equal substitution rates between different nucleotide types (transitions versus transversions). It is well understood that different areas of the human genome are evolving at different rates, so at some point it may be desirable to make estimates based only on certain regions of the genome. There is nothing within the software or the model itself that prevents this application, but increasing the automation of the software to this end would expedite this application. As the sampled number of sites decreases for making inference of any particular parameter, the variance of the estimate will increase. One method of maintaining a high sample size could be to pool data across families. Once again this is currently possible within the current implementation, but not directly so and not without extra work to make the multiple family data *resemble* data from a single family. Finally

the software may can be made available for use by the research community by being packaged into a single executable with documentation.

# References


1. Gauthier, J. et al. Novel de novo SHANK3 mutation in autistic patients. *Am. J. Med. Genet.* **150B**, 421-424 (2009).

2. Stefansson, H. et al. Large recurrent microdeletions associated with schizophrenia. in *Nature* Vol. 455 232-6 (2008).

3. Marshall, M., Solomon, S. & Lawrence Wickerham, D. Case report:de novo BRCA2gene mutation in a 35-year-old woman with breast cancer. *Clin Genet*, 1-4 (2009).

4. Sandal, T. et al. The spectrum of ABCC8 mutations in Norwegian patients with congenital hyperinsulinism of infancy. in *Clin Genet* Vol. 75 440-8 (2009).

5. Macaya, D. et al. A synonymous mutation inTCOF1causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. *Am J Med Genet A* **149A**, 1624-1627 (2009).

6. Castori, M. et al. Darier disease, multiple bone cysts, and aniridia due to double de novo heterozygous mutations inATP2A2andPAX6. *Am J Med Genet A* **149A**, 1768-1772 (2009).

7. Chen, Y. et al. A novel mutation (C1425Y) in the FBN2 gene in a father and son with congenital contractural arachnodactyly. *Genet Test Mol Biomarkers* **13**, 295-300 (2009).

8. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. in *Hum Mutat* Vol. 21 12-27 (2003).

9. Glaser, R.L. et al. The paternal-age effect in Apert syndrome is due, in part, to the increased frequency of mutations in sperm. in *Am J Hum Genet* Vol. 73 939-47 (2003).

10. Tiemann-Boege, I. et al. The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. in *Proc Natl Acad Sci USA* Vol. 99 14952-7 (2002).

11. Cole, D.N., Carlson, J.A. & Wilson, V.L. Human germline and somatic cells have similar TP53 and Kirsten-RAS gene single base mutation frequencies. in *Environmental and Molecular Mutagenesis* Vol. 49 417-25 (2008).

12. Choi, S.-K., Yoon, S.-R., Calabrese, P. & Arnheim, N. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. in *Proc Natl Acad Sci USA* Vol. 105 10143-8 (2008).

13. Qin, J. et al. The molecular anatomy of spontaneous germline mutations in human testes. in *Plos Biol* Vol. 5 e224 (2007).

14. Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. Cryptic variation in the human mutation rate. in *Plos Biol* Vol. 7 e1000027 (2009).

15. Duret, L. Mutation patterns in the human genome: more variable than expected. in *Plos Biol* Vol. 7 e1000028 (2009).

16. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. in *Am J Hum Genet* Vol. 63 474-88 (1998).

17. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. in *Trends in Ecology & Evolution* Vol. 11 367-372 (1996).

18. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. in *Genetics* Vol. 156 297-304 (2000).

19. Chen, F.C. & Li, W.H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. in *Am J Hum Genet* Vol. 68 444-56 (2001).

20. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. in *Proc Natl Acad Sci USA* Vol. 99 803-8 (2002).

21. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. in *Am J Hum Genet* Vol. 70 1490-7 (2002).

22. Consortium, C.S.a.A. Initial sequence of the chimpanzee genome and comparison with the human genome. in *Nature* Vol. 437 69-87 (2005).

23. Lynch, M. et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. in *Proc Natl Acad Sci USA* Vol. 105 9272-7 (2008).

24. Xue, Y. et al. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. *Current Biology*, 1-5 (2009).

25. Frank, S. Evolution in Health and Medicine Sackler Colloquium: Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. in *Proc Natl Acad Sci USA* (2009).

26. Piotrowski, A. et al. Somatic mosaicism for copy number variation in differentiated human tissues. in *Hum Mutat* Vol. 29 1118-24 (2008).

27. Iourov, I.Y., Vorsanova, S.G. & Yurov, Y.B. Chromosomal mosaicism goes global. in *Molecular Cytogenetics* Vol. 1 26 (2008).

28. Xue, Y. et al. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. in *Current Biology* 1-5 (2009).

29. Roach, J.C. et al. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* (2010).

30. Hudson, R. Gene genealogies and the coalescent process. in *Oxford surveys in evolutionary biology* Vol. 7 44 (1990).

31. Yang, Z. Statistical properties of a DNA sample under the finite-sites model. *Genetics* **144**, 1941-50 (1996).

32. Hodgkinson, A. & Eyre-Walker, A. Human triallelic sites: evidence for a new mutational mechanism? in *Genetics* Vol. 184 233-41 (2010).

33. Dempster, A., Laird, N. & Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. in *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 39 1-38 (1977).

34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. in *Bioinformatics* Vol. 25 1754-60 (2009).

35. Metz, C.E. Basic principles of ROC analysis. in *Semin Nucl Med* Vol. 8 283-98 (1978).

# Chapter III

# Discovering *de novo* mutations in collaboration with The 1,000 Genomes Project Analysis Group

# Introduction

## *The 1,000 Genomes Project*

The 1,000 Genomes Project (http://www.1000genomes.org), formed in the fall of 2007, is a multi-national consortium directed towards cataloging an unprecedented amount of human genetic variation. The primary purpose of the 1,000 Genomes Project (1kGP) is to support the discovery and understanding of genetic variants that are involved in human disease and disease susceptibility. While genome-wide association studies provided proof-of concept that this novel variation can be found by adding hundreds of clinically important genomic loci to the nascent field of human medical genetics, much of the heritable risk for diseases with complex phenotypes continues to go undescribed. For example, 18 loci have been identified as linked to Type 2 Diabetes while explaining only 6% of the heritable risk[1]. Currently, obtaining genotype information at these loci is of little medical utility because routinely obtained clinical information such as body-mass index and family history is both more predictive for this particular disease and much easier to query[2]. Improving the ability to explain the heritable component of complex diseases will be a necessary task in the future development of medical genetics.

The 'common disease, common variant' hypothesis[3] states that most of the common, complex disease in humans is attributable to genetic variation that is commonly found, with a minor allele frequency (MAF) above 5%. This hypothesis is the underlying assumption behind the genome-wide association studies. A competing hypothesis is the

'common disease, rare variant'[4] idea that the causal genetic variation is less frequent and more multi-faceted; very rare (0.5% to 1.0% MAF) variants at many different loci may be acting with medium to large effects to influence the overall population frequency of these diseases. By relying on linkage disequilibrium between causal loci and commonly queried 'tag' single nucleotide polymorphisms, genome-wide association studies are unable to locate rare variants. Therefore, by fully sequencing the genomes of roughly 1,200 individuals from diverse populations, the 1kGP attempts to locate genetic variants down to the 0.1% MAF level within protein-coding regions and 1.0% across the human genome. Obtaining a single, deep catalog of variation is more cost-effective than targeted resequencing on a case-by case basis and produces a data set formed without discovery bias towards any particular genomic region.

The initial steps of the 1kGP consist of a series of three pilot projects aimed at addressing issues needed to design the full project in order to meet the quantitative goals. The first and third pilot projects respectively focused on the ability of high throughput technologies applied at low and high coverage levels to uncover common and rare genetic variation from genome-wide and targeted exome resequencing. The second pilot project was constructed to assess the ability of high-coverage data from multiple sequencing technologies to detect variation in deeply sequenced nuclear families. Using immortalized cell cultures, two trios (consisting of two parents and a single child), were sequenced to an average coverage of 20x. While multiple technologies were used for many individuals, Illumina Solexa technology alone was used for all members of both families. Both families

included individual samples initially included in the HapMap project, with one family being of European descent from Utah, U.S.A. (CEU), and the second of Yoruban ancestry from western Africa (YRI). The single-generation pedigree for each family is shown in Figure 3-1. Due to their participation in the HapMap project, each of the family members are labeled by their HapMap identifiers. Throughout the rest of this text the individuals will be identified only by the last two digits of their HapMap numbers.

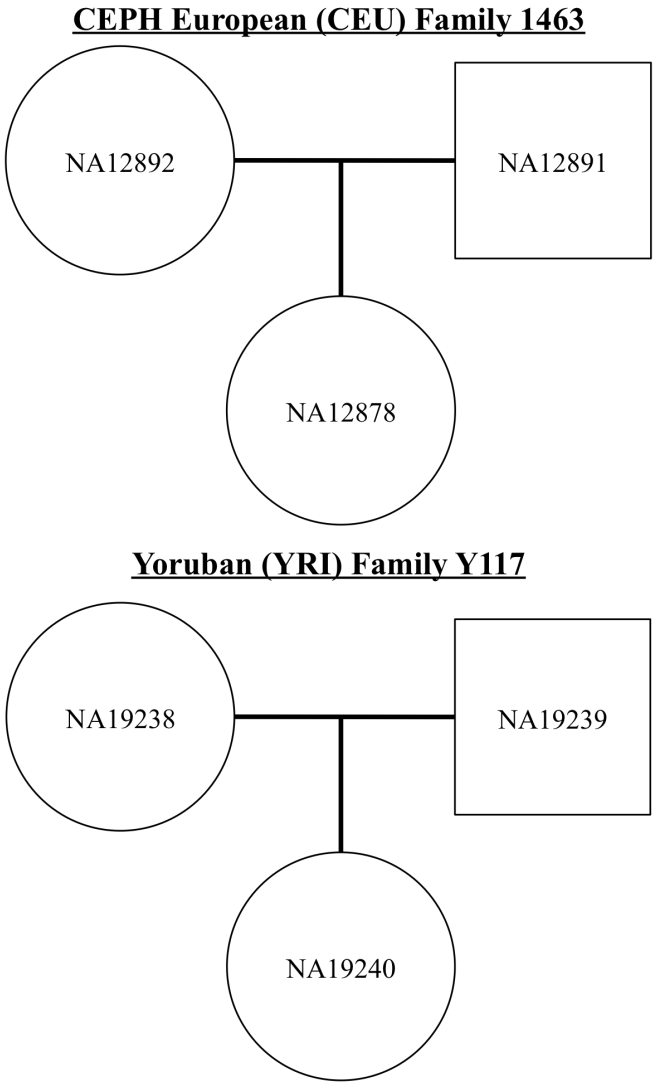



**Figure 3-1.** 1,000 Genomes Project Pilot 2 Deep Sequencing Trios. The family members in each pedigree, all sample donors to the HapMap Project, are labeled by their HapMap identifiers. From this point on, these individuals will be referred to only by the last two digits of their HapMap identifiers, i.e. 91, 92, and 78 for the CEU trio and 38, 39, and 40 for the YRI trio.

## *De novo Mutation Discovery in the 1,000 Genomes Project Pilot 2 Trios*

In collaboration with the analysis group of the 1kGP, I participated in a team effort to query these two families for spontaneous mutations in the form of Mendelian errors using the method presented in Chapter II. Two other groups, one from the Wellcome Trust Sanger Institute (WTSI) and one from the Broad Institute of MIT and Harvard (BI), were involved in spontaneous mutation discovery using independent methods. All groups used the same original data sets. Although several sequencing technologies have been used to produce sequence data for the families, only read data produced by Illumina Solexa technology was used as this type of data alone is available for all members of both families. Eliminating the between-technology variability due to different base-calling strategies and error profiles was preferential to including all of the sequencing data available. The specific data used by all groups was the July, 2009 release of the individual BAM files[5] produced by MAQ[6] alignments of Solexa reads downloaded from the 1000 genomes ftp repository (ftp://ftp-trace.ncbi.nih.gov/1000genomes).

The data pre-processing steps and the *de novo* mutation prediction methods varied between the three groups. My approach was to use the method described in Chapter II for discovering spontaneous mutations by jointly inferring genotypes for the family members, or Probabilistic Inference of the Pedigree using High-Throughput Sequencing (Piphits). In brief, the overall strategy was to (1) pre-processing of the data including application of quality filters, (2) execute Piphits, (3) extract loci with a slightly elevated probability of

mutation (> 0.001), (4) apply locus-annotation filters to this list, (5) re-execute Piphits on the resulting sites using unfiltered data, and (6) report a candidate list of loci with a higher probability of mutation (> 0.10) based on the second execution of Piphits. The final resulting list of mutation candidate loci for each family was merged with the lists provided by the WTSI and BI groups and experimental validation was carried out to test the predictions. In the sections that follow I will provide detail on each of these steps and present the merged validation results, as well as some conclusions drawn relating to the original quality of the data and the demonstrable capability of Piphits to handle real data.

# Methods

## *Data pre-processing and compression*

To begin, the downloaded alignment files were converted into the SAMtools-pileup (pileup) format using the default options[5]. One pileup file contains the full read alignment information for one sequenced sample or individual. After investigating the particular alignment qualities of each pileup, custom quality filters were applied to the pileup files of each individual at a stringent level. Each base call in high-throughput sequencing (HTS) is accompanied with a confidence measure (BaseQ) of the call that is reported by the sequencing machine. A separate quality metric (MapQ) is assigned to every sequencing read mapped to the genome by the alignment software, in this case by MAQ. Both values have log-based *phred*-scaling[7], whereby a quality score of 30 indicates the specified base

call or read alignment has a 0.001 probability of being incorrect. Taken together these values provide a measure of the degree of confidence one can have when determining a genotype based upon a set of reads. As introduced in Chapter II, the Piphits compression data structure uses discrete read-level base observations where one observation carries as much weight as any other. To approximate this uniformity in the real data, minimum quality filters were applied using levels set by a detailed investigation of the MapQ and BaseQ value distribution in each individual pileup. Note that if the data is filtered too stringently, coverage can be reduced to the point of being minimally informative at each site, leading to erroneous genotype inferences based upon the model used. However, too little filtering can lead to many high-confidence mutation calls based upon poor-quality data, over-taking any true signal mutation, particularly on a genome-wide scale. Therefore a two-step filtering strategy was used in which the first filters creating the input to the Piphits method were quite restrictive. After executing Piphits, a less stringent filter was applied in order to recover potentially informative reads lost initially in a manner that is discussed below. In addition to the quality filters, due to the ability of many reads from repetitive regions to align incorrectly in tandem, sites with a read depth greater than a maximum threshold defined by specific pileup investigation were ignored in the analysis. The quality and coverage filters used are given in Table 3-1. Following compression and filtering the resulting input of sites to Piphits consisted of 2.49 billion loci covered at an average depth of 19.7x for the CEU trio and 2.46 billion loci covered at 17.8x in the YRI trio.

**Table 3-1**. Custom Data Quality Filters. These were chosen to create a more uniform quality level within each pileup file. This was necessary due to the uniform treatment of read observations in the Piphits data compression. Without these filters, Piphits can make many high-confidence, false positive mutation predictions based only upon poor-quality data that overwhelm signals of true mutation.

| Individual | Minimum MapQ | Minimum BaseQ | Maximum Raw Sequence Depth |
|---|---|---|---|
| 91 | 63 | 21 | 74 |
| 92 | 67 | 21 | 63 |
| 78 | 67 | 21 | 75 |
| 39 | 65 | 21 | 63 |
| 38 | 63 | 21 | 54 |
| 40 | 65 | 21 | 75 |

## *Application of Piphits to the 1,000 Genomes Pilot 2 Trios*

The Piphits Expectation Maximization (EM) algorithm was attempted to produce maximum likelihood estimates of the model parameters for the filtered data. Based on the simulation studies performed in Chapter II, the EM is expected to over-estimate the *de novo* mutation rate and under-estimate the sequencing error rate as well as the population mutation rate, all primarily owing to alignment error. Several sources provide expectations for the results of the EM algorithm: high-throughput sequencing on the Illumina Genome Analyzer is reported to have a raw per-base sequencing error rate on the order of 0.01[8], the traditionally held value for the human population mutation parameter is 0.001 per base per generation[9], and the spontaneous mutation rate in humans has

consistently been reported to be on the order of $2.0 \times 10^{-8}$ per base per generation[10,11]. The EM-produced maximum likelihood estimates for these parameters are shown in Table 3-3. As expected, the sequencing error rate is lower than expected and the mutation rate is much higher. Because the source of over-estimation of the mutation rate was primarily alignment error in the simulation studies, observing an over-estimation by three orders of magnitude in this case is a cause for concern either with the post-filtering alignment quality of the data or with the robustness of Piphits to adequately account for mapping error. One fact to take into account is that the sequencing was performed on immortalized cell cultures, which likely have accumulated many mutations that are indistinguishable at this point from true Mendelian errors. It is likely that the frequency of discovered mutations based upon this data set will indeed be several times higher than traditional estimates. With regard to the population mutation rate, it is interesting to note that the estimate is higher (~26%) for the YRI family, given the larger expected effective size of the Yoruban population relative to that of western European populations[12].

**Table 3-3**. Maximum likelihood estimates of model parameters.

| EM-produced Maximum-Likelihood Parameter Estimates | CEU | YRI |
|---|---|---|
| Spontaneous mutation rate $\mu$ | $1.74 \times 10^{-5}$ | $3.28 \times 10^{-5}$ |
| Population mutation rate $\theta$ | $9.51 \times 10^{-4}$ | $1.19 \times 10^{-3}$ |
| Sequencing error rate $\varepsilon$ | $1.20 \times 10^{-3}$ | $1.14 \times 10^{-3}$ |

Due to the expected departures of the estimated values from the ranges, the actual calculation of the probability of mutation, $\delta$, for each site was performed using the previously reported rates mentioned above. The spontaneous mutation rate was set to $2.0 \times 10^{-7}$ to reflect the likely inclusion of cell-line mutations with the sampled data. No attempt was made in this discovery phase to distinguish cell-line or somatic mutations from germline events due to the lack of a third identical-by-descent allele within a trio pedigree, noted as being necessary in earlier discussions. Therefore the somatic mutation rate parameter within the model was set to 0.

## *Candidate mutation list post-processing*

Following execution of the expectation algorithm for every available site, those showing a moderately elevated level of $\delta > 0.001$ were segregated from the rest of the genome. These sites were then subjected to a progressive set of annotation-based filters to remove previously described genetic variants and sources of alignment error such as known

copy number variations and insertions / deletions. These filters were applied in tandem with the WTSI candidate list through collaboration on the definition of appropriate filters. In general, they removed loci that were in close proximity to previously called variants such as copy number variations and polymorphisms described in dbSNP, and sites that overlapped with genomic regions annotated variable number tandem repeats and segmental duplications. Application of these locus filters removed approximately 30% of the genome from primary inquiry. Furthermore any site observed to have a single read in either parent matching the putative mutant allele in the offspring was discarded. Given the stringent level of quality filtering, the presence of a few high-quality reads was treated as a signal of true heterozygosity in the parent that would be revealed with more confidence considering all the data available. The overall progressive culling of the candidate site list for each family is shown in Figure 3-2 and Figure 3-3. At the last step of the post-processing, the Piphits method was applied specifically to the remaining candidate sites, accepting base calls at this time with a minimum MapQ of 20 and BaseQ of 10. After application of a final cutoff of $\delta > 0.10$ the data was placed in descending order of $\delta$, with the most confident predictions at the top of the list. With the idea of over-calling in order to maximize sensitivity, the finalized candidate list included 2,089 putative mutation sites from the CEU family and 1,682 sites from the YRI family. The resulting distribution of $\delta$ values for the two sets of sites is shown in Figure 3-4.

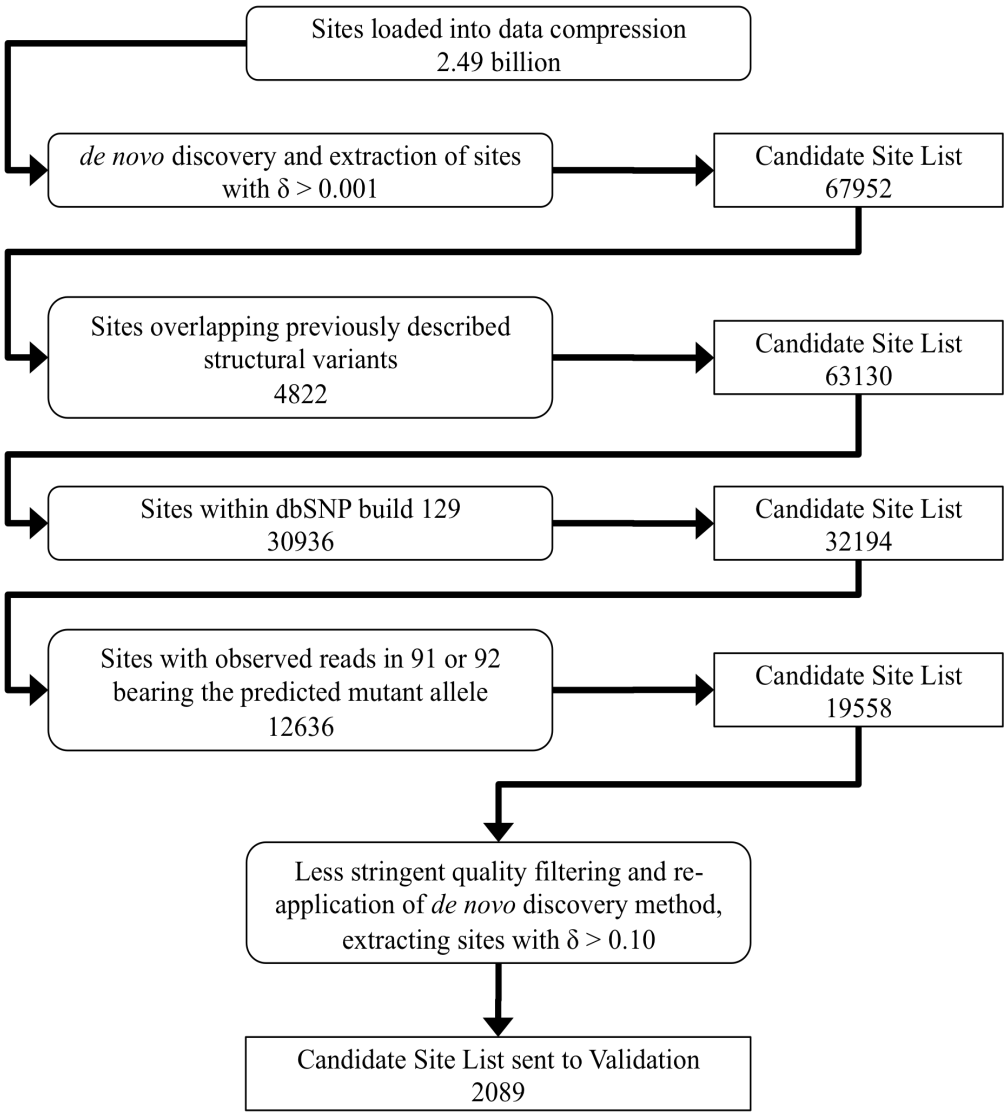



**Figure 3-2.** Generation of Validation Set for CEU Family

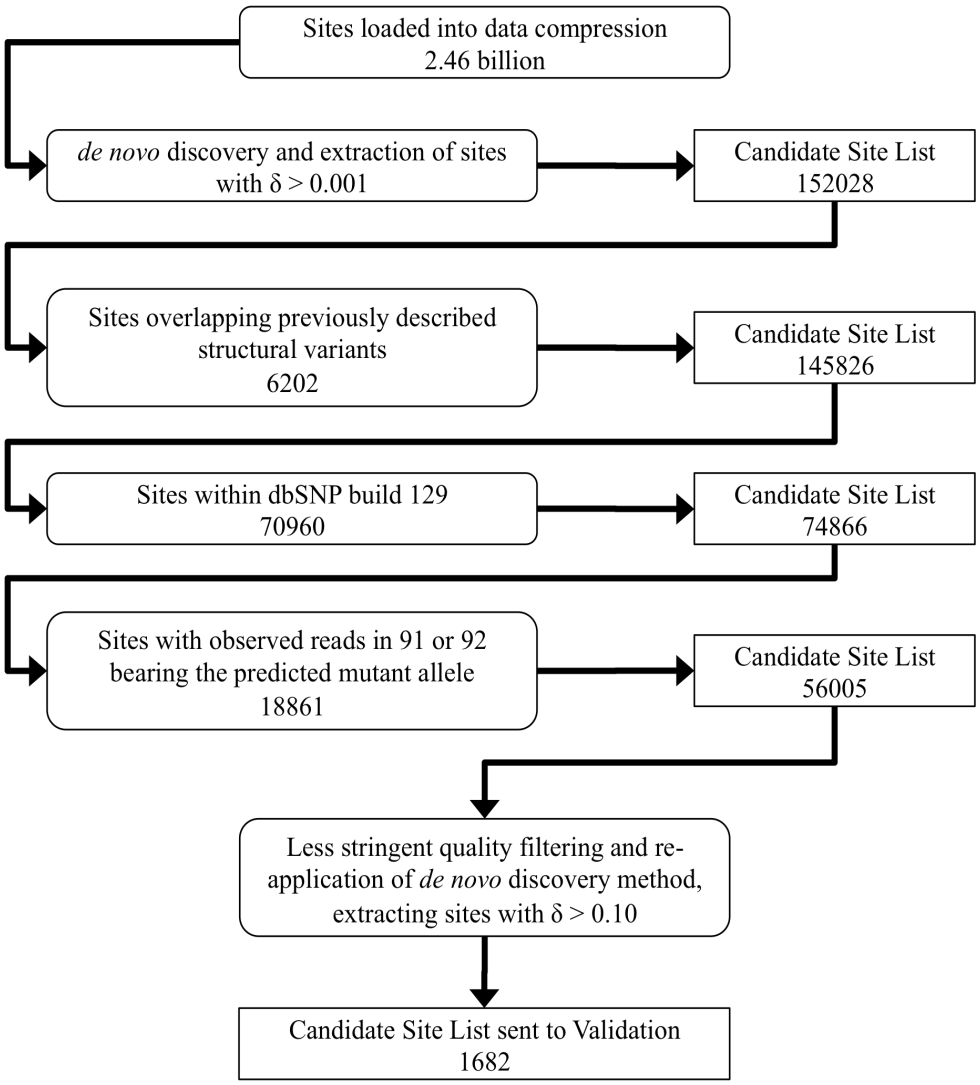



**Figure 3-3.** Generation of Validation Set for YRI Family

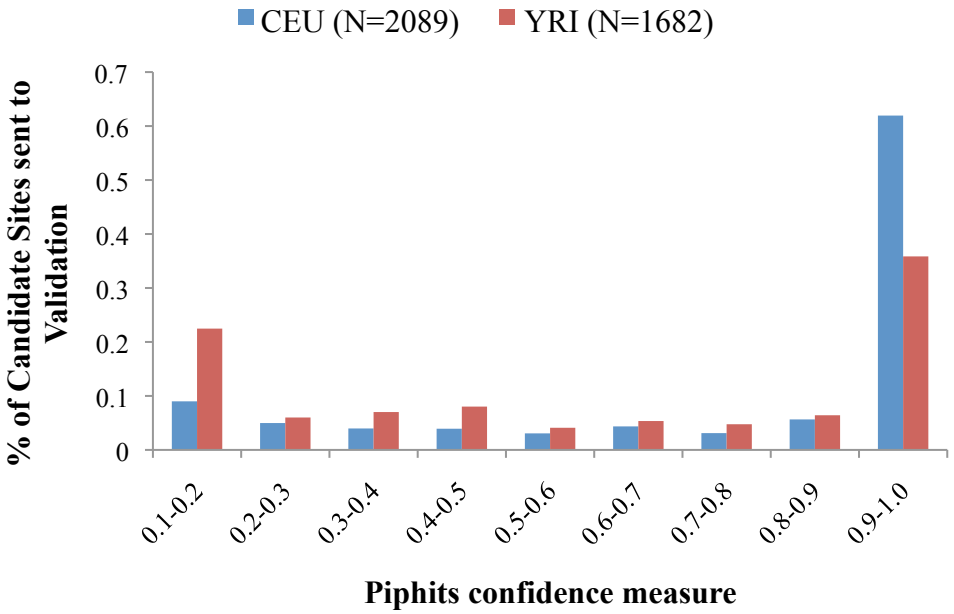



**Figure 3-4.** Piphits confidence measure distribution for final candidate site list.

### *Experimental validation of the predicted mutations*

This list of candidate sites was merged with those from BI and WTSI to form a single list of possible *de novo* mutation sites. Created independently, it should not be surprising that there is a fair degree of overlap between the predicted site lists. The overlap of these sites is shown in Figure 3-5. The higher overlap with between the Piphits-produced list and the WTSI group than exists between either list and the BI list may be a result of the similar filtering approach used. The BI list was produced without filtering of candidate sites.

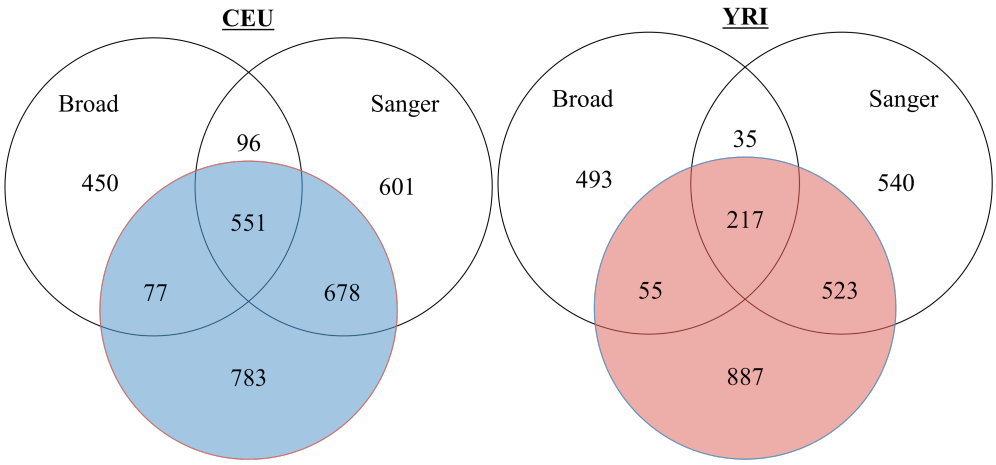



**Figure 3-5.** Overlap of CEU and YRI candidate sites with WTSI and BI groups. In total 3236 CEU sites and 2750 YRI sites were included in experimental validation.

The union of all three lists, totaling 3,236 and 2,750 sites from the CEU and YRI families respectively, formed the basis of a targeted DNA resequencing exercise aimed at querying the specific genotype of each individual in order to test the validity of the mutation predictions. Working in collaboration with members of the Awadalla Lab at the University of Montreal (UdeM), a single strategy was used to validate all sites across the three groups. First, probes were designed for 120 total bases surrounding each candidate mutation site. Of the original candidate lists, probes were successfully designed for 5,838 total sites: 3,135 sites for the CEU family and 2,703. The Agilent SureSelect Target Enrichment System[13] was performed by the UdeM group to produce amplified RNA solution for each individual. Applied Biosystems resequencing by oligo-ligation (SOLiD) of the amplified RNA sample was performed to produce approximately 32 million 50-base pair reads. The reads produced were then aligned to the reference genome using Bowtie[14], a

BWT-based alignment tool which is capable of handling color-space reads as produced by the ABI SOLiD machine. The resulting alignment was summarized using simply the number of read observations matching the mutant allele and the number matching the reference. This complete process, from sequence capture to alignment is represented in Figure 3-6. Samples were treated in this manner for each family member within the original trios, as well as extended members of the CEU family. By including family members from the extended CEU pedigree (shown in Figure 3-7), the germ/cell-line or somatic status of each mutation could be identified. The full alignment results for all individuals sampled are shown in Table 3-4. Currently SOLiD sequencing is known to produce roughly 50% unmappable reads, so the average of ~45% of reads aligned per individual is not far from this mark.

In tandem, the WTSI group also performed experimental validation of these predicted sites using MAQ alignment of Illumina sequencing reads at extremely high coverage levels, with 95% of the sites covered in a typical sample and a mean coverage of 3,000x. The WTSI group also sample the extended members of the CEU family as well as a blood sample of the child of the YRI pedigree, individual 40, to facilitate distinguishing germline from somatic mutation events. The validation data from the UdeM group and the WTSI group were merged to locate the true de novo variants predicted in the three lists by classifying each site as ‘germline’, ‘somatic’, ‘inherited’, ‘false-positive’ according to the rules given in Table 3-5. If sufficient alignment data was not available for a predicted sites, it was labeled as ‘no call’.

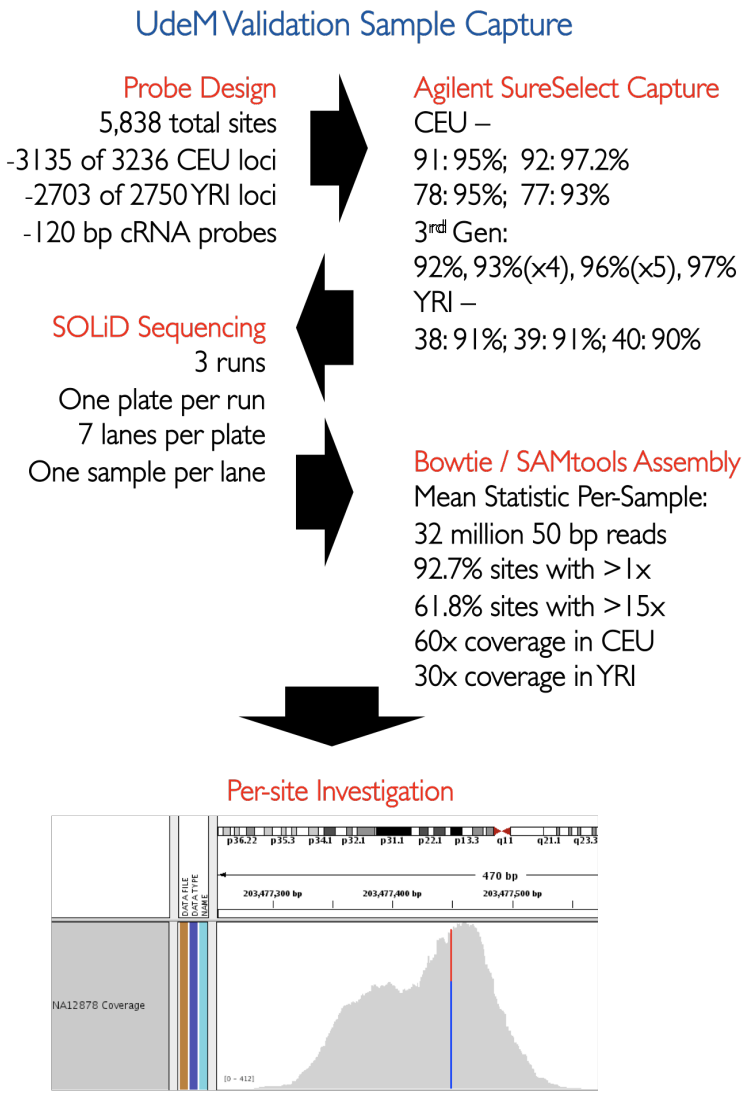



**Figure 3-6.** UdeM validation capture procedure used for each sample. The capture success numbers are shown for each sample specifically. 3rd Gen in the CEU family denotes the children of 78. Picture at the bottom is the Integrative Genomic Viewer (IGV) depiction of the validation reads from sample 78 at one predicted *de novo* site. The height of the graph is the amount of read coverage and the red/blue column shows validated heterozygosity at the putative *de novo* position.

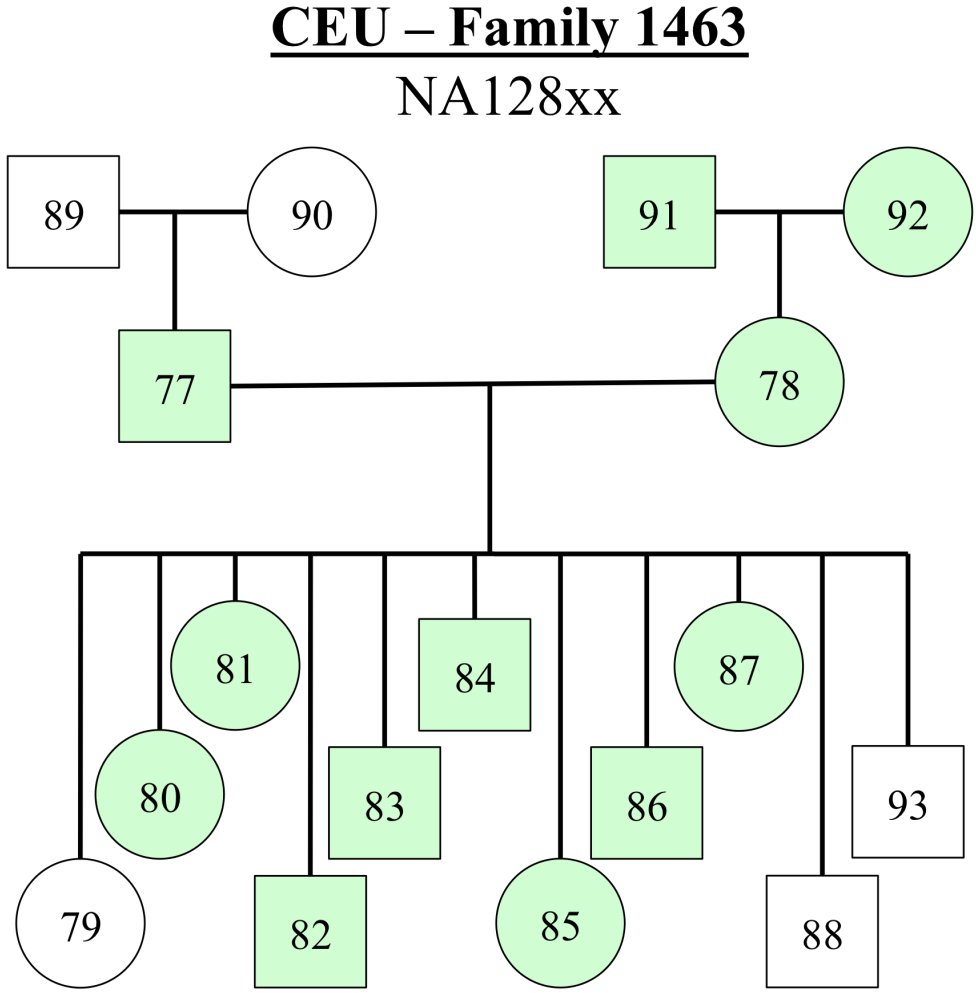



**Figure 3-7.** Validation pedigree for CEU family. All individuals highlighted were sampled by the UdeM validation. The WTSI validation sampled the same individuals except did not sample 77.

**Table 3-4.** Bowtie alignment results of UdeM validation reads. For individuals 80 through 87 (3rd Gen), the sum of reads sequenced from all individuals and the average coverage across individuals are reported.

| **CEU Individual** | **91** | **92** | **78** | **3rd Gen (80-87)** | **77** |
|---|---|---|---|---|---|
| Total Reads | 36 million | 43 million | 37 million | 222 million | 31 million |
| % Reads Aligned | 57% | 43% | 51% | 49% | 32% |
| % Sites Captured | 92% | 94% | 92% | 97% | 90% |
| % Sites > 15x | 63% | 67% | 63% | 67% | 47% |
| Number of Sites Captured | 3,013 | 3,098 | 3,035 | 3,179 | 2,970 |
| Mean Read Depth of Candidate Sites | 61.6x | 56.1x | 46.3x | 35.9x | 23.4x |

| **YRI Individual** | **38** | **39** | **40** |
|---|---|---|---|
| Total Reads | 41 million | 36 million | 30 million |
| % Reads Aligned | 30% | 42% | 55% |
| % Sites Captured | 89% | 90% | 89% |
| % Sites > 15x | 49% | 53% | 49% |
| Number of Sites Captured | 2,457 | 2,462 | 2,446 |
| Mean Read Depth of Candidate Sites | 32.0x | 38.4x | 33.1x |

**Table 3-5.** Validation classification rules for occurrence of mutant allele.

| Classification | Parents | Offspring | 3rd Gen (CEU) /Blood (YRI) |
|---|---|---|---|
| germline | No | Yes | Yes |
| somatic | No | Yes | No |
| inherited | Yes | Yes | Yes |
| false positive | No | No | No |

# Results

## *Validated Mutations*

Using the merged results of the UdeM and WTSI validation methods, across the three sets of predicted mutations, a total of 1,945 Mendelian errors were validated including 1,141 in the CEU family and 804 within the YRI family. These Mendelian errors originate either from a *de novo* germline or somatic mutation and constitute ~33% of all sites investigated. Figure 3-8 shows the aggregate percentage of sites validated to each classification. The genomic distribution of these validated *de novo* mutations across the first ten chromosomes is shown in Figure 3-9, and Figure 3-10 gives the contribution by each group to the four validation classifications.

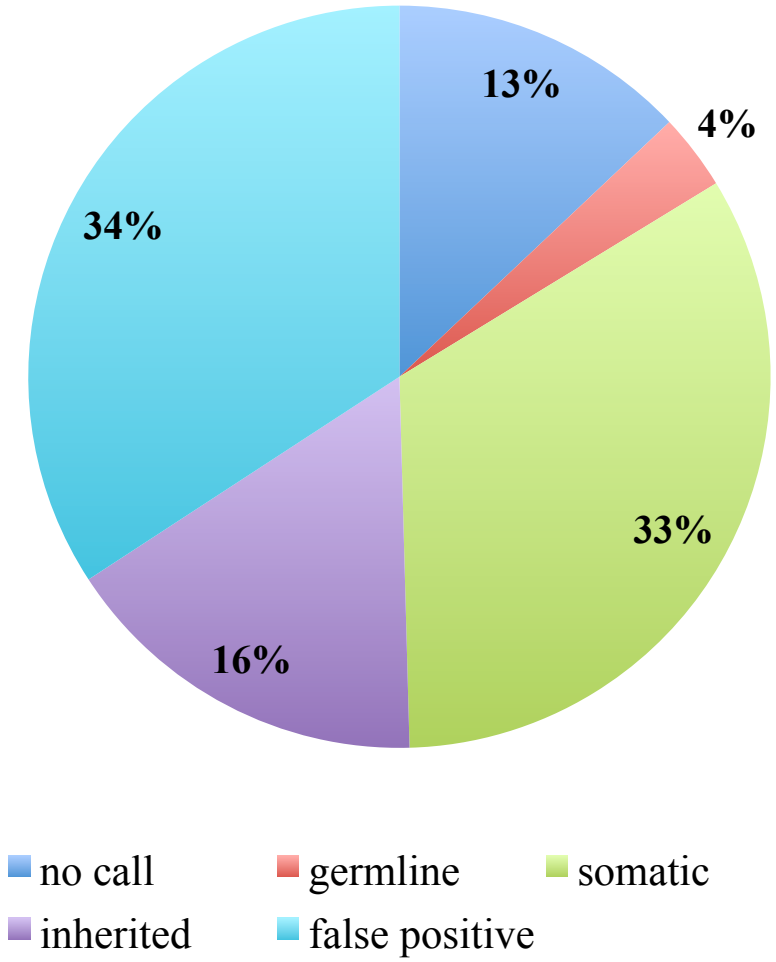



**Figure 3-8.** Validation classification results across both families for all prediction lists (N=5,986).

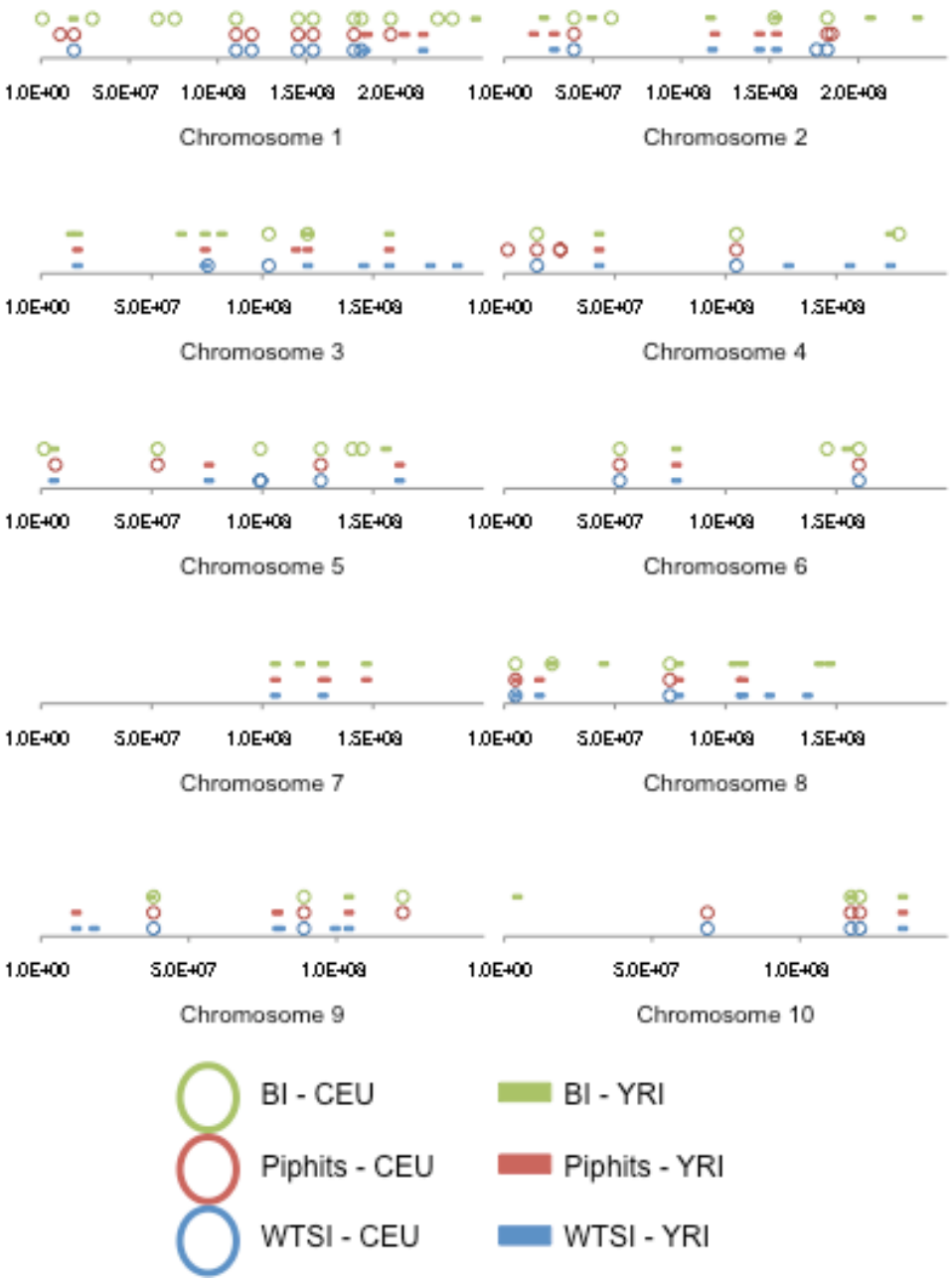



**Figure 3-9.** Illustration of genomic locations of validated *de novo* mutations for the YRI and CEU families.

Examining these 1,945 validated mutations across both families, the first notable pattern to observe is the large number of somatic events relative to the number of germline mutations estimated. A total of 1,720, roughly 90%, of the validated *de novos* were not germline in origin. The high sequence capture success rate and high level of sequencing coverage within the CEU 3rd generation or YRI blood sample across the two validation methods suggests there are few of these which are miss-classified. There are two possible sources for the non-germline mutations. The first is the possibility that mutation events occurred within somatic cells of the human donors during development while the second is that they are the result of in *vitro* evolution of the sampled cells within a culture. Either origin can have important implications. Assuming for the moment that each of these non-germline validated mutations truly originate in somatic tissues, this finding would suggest an approximate average of 300 mutations per individual occurred within the cell lineage from germline to the somatic tissue sampled (~1,800 mutations divided amongst three family members in two families). In any particular individual, this begs the question how many variable sites will be observed if several somatic tissues were sampled whose progenitor stem cells split early in development. Would 300 mutations be expected along each lineage? What are the implications for genetic diagnostics for human disease? These ideas have recently gained more attention and it is possible a field of human somatic evolutionary genomics will emerge from the discussion[15]. Alternatively, cell-culture origin for this high number of mutations could have implications on the utility of immortalized cell lines for making general inferences on the genetics of the donor

individual or population. Further investigations and sequencing studies will be necessary to determine the relative mutation rates within human cell cultures and in somatic cell lineages during development.

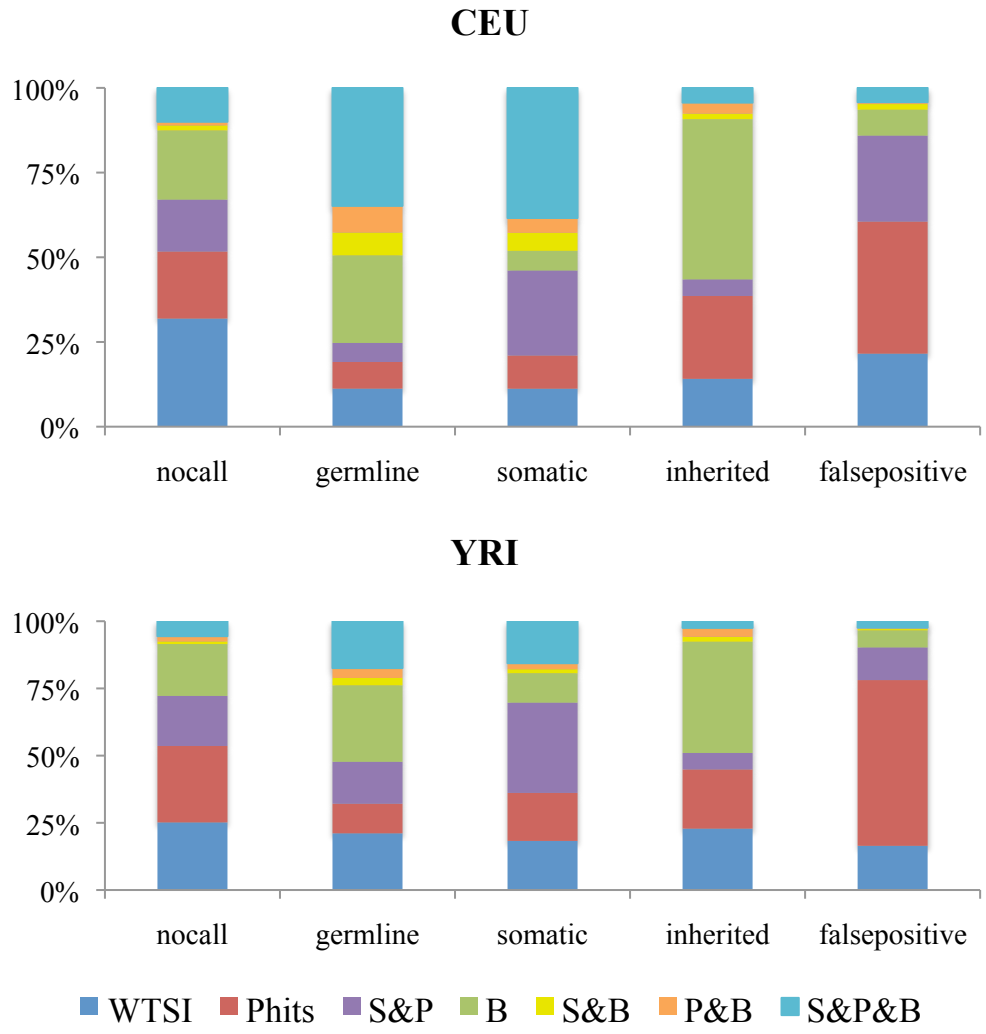



**Figure 3-10.** Relative contribution to each validation class for each family by each group. WTSI (S) denotes Wellcome Trust Sanger Institute, Piphits (P) probabilistic inference of pedigree using high-throughput sequencing, B Broad Institute.

The number of validated germline mutations totaled 198, with 109 in the YRI trio and 89 in the CEU. Using these values with the number of sites investigated within each family, one can estimate the per-base, per-generation human mutation rate. Simply by dividing the number of validated germline mutations by the total number of sites examined, which in this case would truly be the length of the HG18 reference genome used for alignment, $2.834 \times 10^{9}$, gives a rate estimation of $3.846 \times 10^{-8}$ for the YRI trio and $3.140 \times 10^{-8}$ in the CEU family. Although slightly high, these values are in accordance with recently reported estimates[10,11,16]. Dividing the scaled population mutation rate reported by the Piphits EM algorithm by these rate estimates, the effective population size from which the parents of the two trios arose can be approximated. Using $\theta = 9.51 \times 10^{-4}$ for the CEU trio gives an effective population size of 30,282. For the YRI trio the estimate for the population mutation rate was $1.19 \times 10^{-3}$, giving an effective population size estimate of 30,941. These estimates are approximately three times the magnitude of previous reports[17].

For all validated mutations, the pattern of base substitution is given in Table 3-6. Based on these values, the transition-transversion rate is 1.7 for the CEU trio and 1.5 for the YRI trio. This rate is widely held in humans to be approximately 2.0[18,19], indicating the validated sets either retain some false positives or are missing additional mutations.

**Table 3-6.** Base substitution patterns for validated *de novo* mutations.

| CEU | A | C | G | T |
|---|---|---|---|---|
| A | | 8 | 14 | 3 |
| C | 6 | | 5 | 14 |
| G | 22 | 3 | | 3 |
| T | 8 | 16 | 7 | |

| YRI | A | C | G | T |
|---|---|---|---|---|
| A | | 3 | 10 | 2 |
| C | 4 | | 3 | 17 |
| G | 17 | 4 | | 3 |
| T | 8 | 12 | 6 | |

## *Piphits false positives*

Of the sites which were not validated as *de novo* mutations, 3,019 or ~50% were validated as having no mutation event, either because the site was not variable in the 78 or 40 individual or because one of the parents was truly heterozygous and the allele was inherited. Because the three candidate mutation lists were produced by different methods with possibly different propensities for error, I will here focus primarily on the contribution of the Piphits method to the overall amount of false positives mutation predictions and discuss some of the causes of error. Validation was attempted for a total of 2,089 CEU sites and 1,682 YRI sites predicted by the Piphits method specifically. Included in these counts are 1,005 and 817 false positive sites that contributed to the overall false positive rate. There were nearly 500 for each family sites that were predicted only by Piphits that were among these false positives. The relative percentages for each category of the Piphits predicted site list for the two families is given in Figure 3-11.

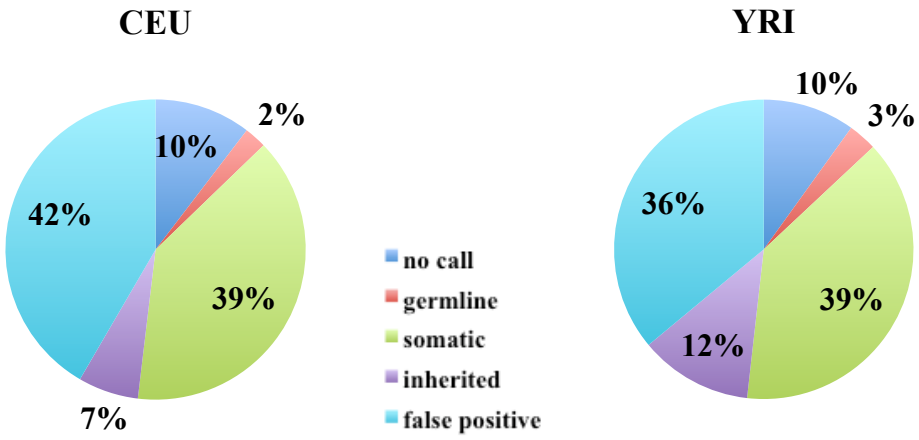



**Figure 3-11.** Proportion of Piphits predicted mutation sites for each validation result.

One clear class of mistakes that was made in the Piphits method is the prediction of a mutation events when no event, either by falsely predicting a variant site in the trio offspring leading to a false-positive classification, or missing a heterozygote genotype in a parent, leading to a classification of 'inherited.' Considered jointly, the overall amount of false positive mutation calls in this prediction-validation experiment is the sum of the inherited and false-positive classification across the two families: 1,816. This is actually larger than the number of true positives defined as the sum of the validated germline and somatic mutations for the two families: 1,468. However, this result was somewhat expected given the original decision to over-call the number of mutations with the goal of maximizing sensitivity. Indeed, many low-confidence mutation sites were included in the validation step. The correlation between false-positive fraction and the confidence measure, $\delta$, is shown in Figure 3-12, indicating that the complete story is not told by

looking at the false positive and true positive calls alone. While many false positive calls remain at the highest levels of $\delta$, the percentage relative to the number of true positive mutation calls is lower than at lower confidence levels. This can more clearly be seen in Figure 3-13.

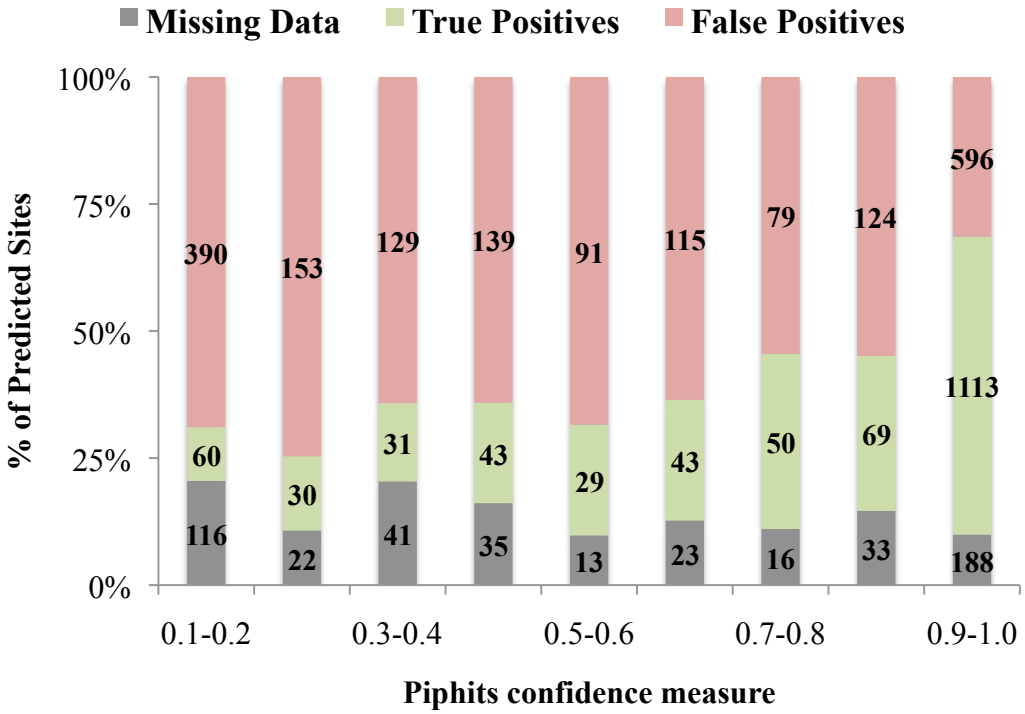



**Figure 3-12.** Correlation of validation results across both families with the Piphits mutation confidence measure $\delta$.

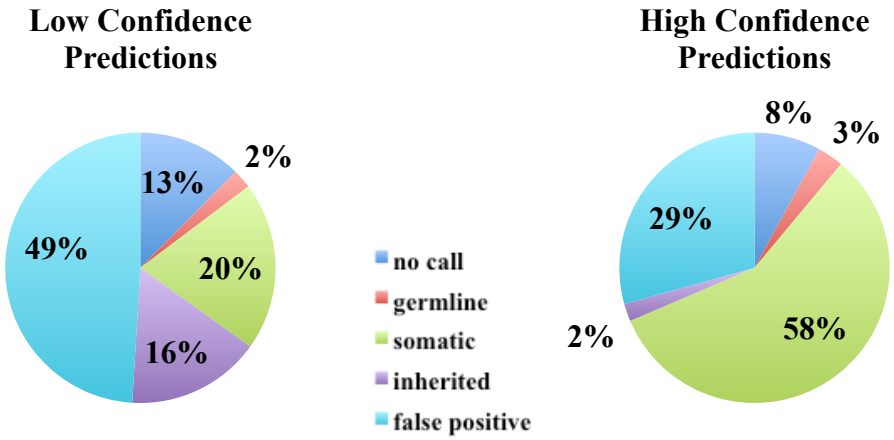



**Figure 3-13.** Validation classifications of low- ($\delta$<0.90) and high- ($\delta$>0.90) confidence Piphits mutation predictions pooled across both trios.

The issue remains that a substantial fraction of the sites predicted with high $\delta$ values were validated as false-positives. What is the cause of these misplaced predictions? As with the EM algorithm over-estimation, the inability of the quality filtering to effectively reduce the input data set to one with reads primarily mapped with equal confidence may be to blame. As an example, consider the data for a particular site that was predicted to be a mutation only by the Piphits method, and not by the BI or WTSI groups, and was classified as a false positive by the validation. The site is genomic coordinate 8:77023179 in the YRI trio, where the first number is the chromosome and the second is the chromosome position. Examination of this site reveals that the reads supporting the original predication of a heterozygous call in individual 40 were likely misaligned to a region of low complexity. The original multiple sequence alignment supporting this call is shown in Figure 3-14. On the basis of visual inspection the reads carrying the additional

'A' alleles are not likely to be correctly mapped to this region. When evaluating any particular site, Piphits is unaware of the genomic context such as what is seen in this alignment due to the statistically independent treatment of each locus and the decoupling of specific genomic position from the read observations during the compression algorithm. During quality filtering, several reads were discarded based on MapQ and BaseQ which matched the reference and which matched the putative mutating allele. These qualities are shown in Figure 3-15. In this case the MapQ filter was unable to cull the erroneous reads at either stringency level. The issue is exacerbated by the relatively small high-quality coverage of the reference allele. Since only seven reference alleles were observed passing the quality filter, the three high-quality mutant alleles were sufficient to support a heterozygous call at this site.

**Figure 3-14.** Integrative Genomics Viewer (IGV, http://www.broadinstitute.org/igv) multiple sequence alignment for the 1kGP reads at 8:77023179 for YRI individual 40. The gray columns indicate the read depth at each site, and the horizontal gray bars represent the individual sequencing reads. Only bases are shown that do not match the reference sequence at the bottom of the figure. Where the gray column is blue and green indicates a consensus heterozygous position, with the green portion corresponding to the 'A' allele and the blue with the reference, 'C'. This site was validated as a germline mutation by both the combined UdeM and WTSI validation experiments. IGV available at http://www.broadinstitute.org/igv.

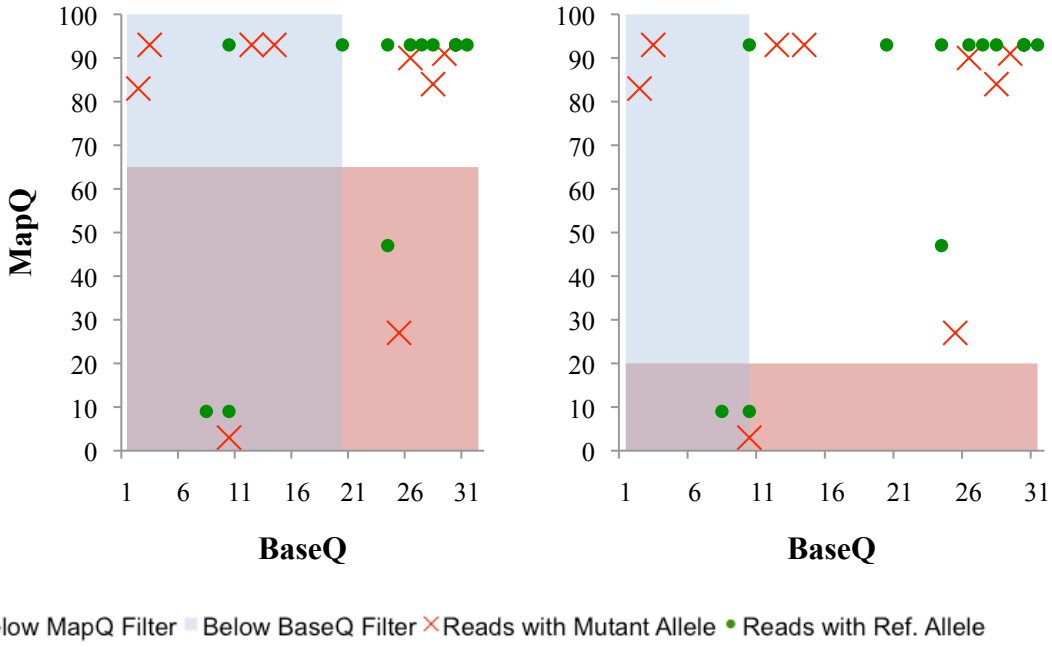



**Figure 3-15.** Mapping and Base qualities for reads carrying the reference and mutant allele at site 8:77023179 in YRI individual 40. The left panel shows the stringent quality filters used in the first pass, the right panel shows the lowered quality filters as described in the Methods section. In both cases sufficient reference and non-reference reads exist to support a heterozygous call.

## *Piphits false negatives*

In addition to the over-calling of the Piphits method, equally as interesting are the sites where the method was not sensitive to the signal of mutation. There were 62 sites study-wide where both BI and WTSI predicted a mutation that was validated while Piphits did not. To learn the potential source of these unidentified mutations, site 1:110099347 within the CEU family was investigated in detail. A picture of the original multiple sequence alignment for this site is given in Figure 3-16. The alignment appears clean at this validated heterozygous position in individual 78. However, as shown in Figure 3-17,

the mapping and base qualities for the majority of reads bearing the mutant allele at this site were found to be below the stringent filter levels given in Table 3-1. Because only four reads carrying the mutant allele and 25 matching the reference were retained, the model inference was that these four reads were the result of sequencing errors. Therefore the calculated $\delta$ for this site fell below the original threshold (0.001) and was not included in the downstream analysis. Reducing the quality filter would have increased the mutant allele coverage to a sufficient level to allow inference of a mutation event, however taking this strategy at this site would likely have produced more false positive results as occurred with site 8:77023179.

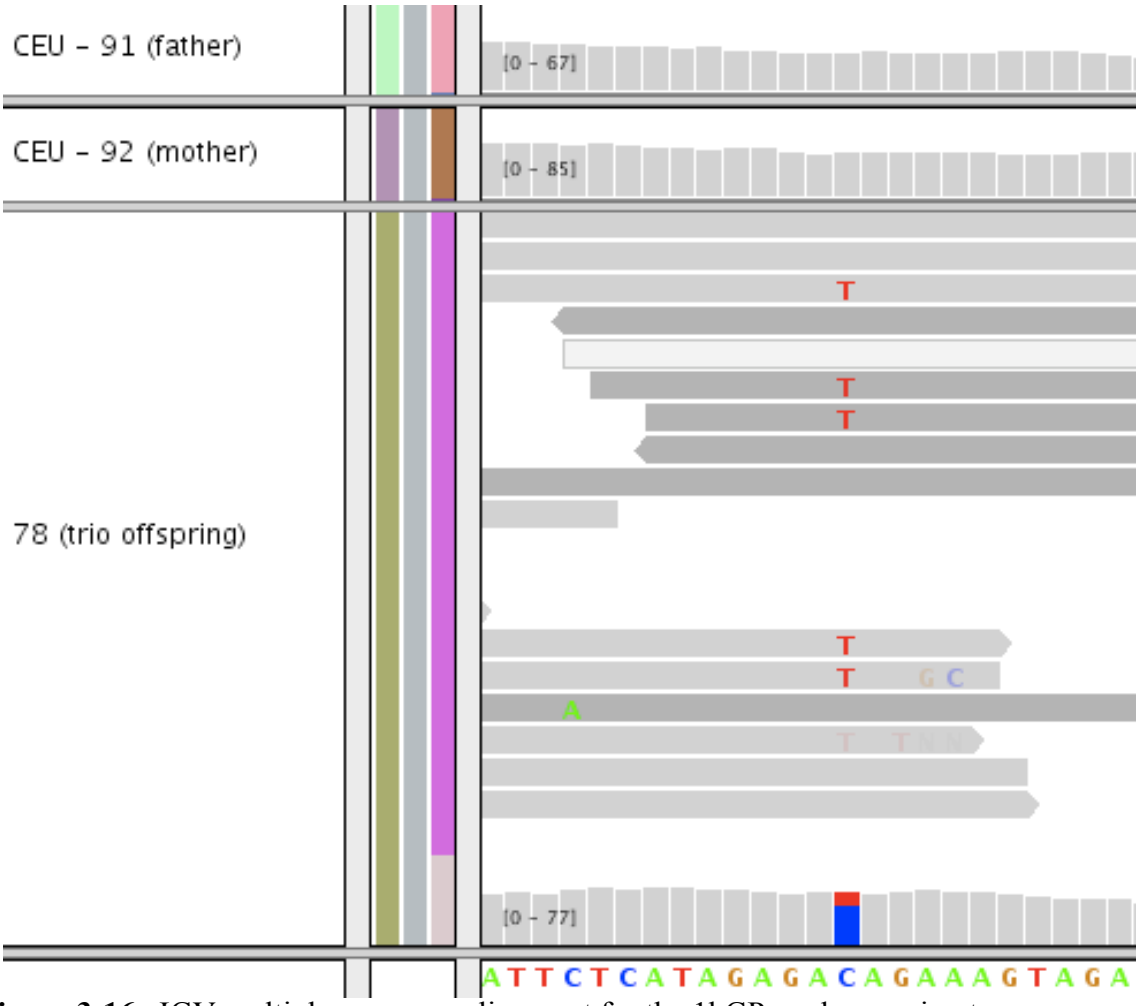



**Figure 3-16.** IGV multiple sequence alignment for the 1kGP reads mapping to 1:110099347 for individual 78, child of the CEU trio.

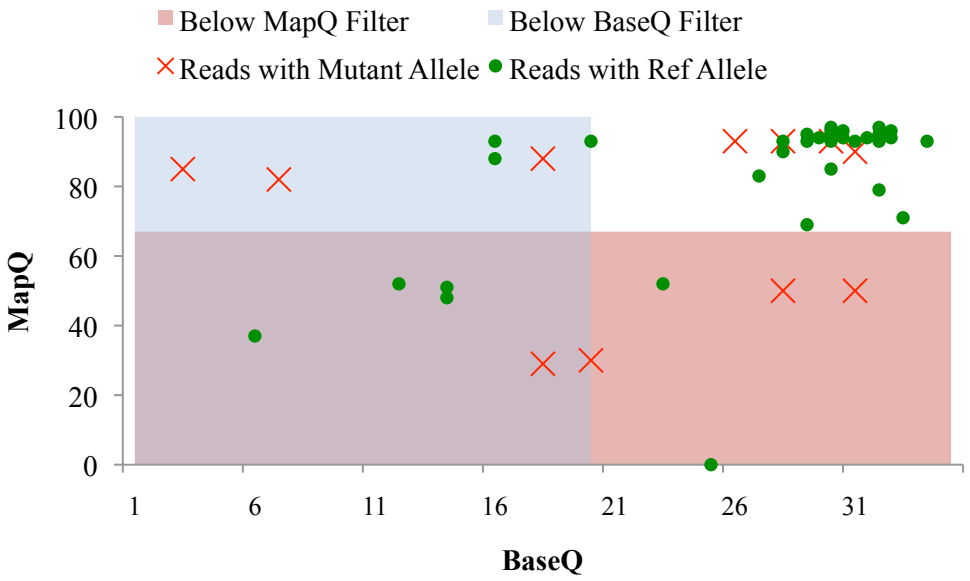



**Figure 3-17.** Original mapping and base qualities for reads carrying the reference and mutant allele at 1:110099347 in CEU individual 78.

In both the false positive and true positive cases highlighted, the quality filtering contributed towards both increase the amount of false positive and false negative calls. Given access to the full, unfiltered data at these sites, the model would infer a mutation for both 1:110099347 and 8:77023179. Indeed the elevated mutation rate in the EM algorithm is likely due to inferring a mutation at sites like 8:77023179. However it is clear that the data can be over-filtered, removing valid mutant alleles from the analysis as was observed for site 1:110099347. In either case, an argument exists for considering the MapQ metric as an unreliable predictor of the validity of the read alignments. This is likely an over reaction since these are marginal cases under examination, and the full reliability of the MapQ metric should be judged from a broader scope. Indeed, probabilistically

incorporating the quality metrics into the data compression algorithm could allow the removal of the quality filtering process and improve the model inference by taking advantage of any information included within the quality scores. The MAQ consensus genotyping model takes advantages of these scores and was incorporated into the WTSI and BI methods for predicting the mutations in this data set.

### *Piphits true negatives*

There were 17 sites across the data set at which the Piphits method performed well, passing over sites that the WTSI and BI groups both predicted as mutants when the validation showed otherwise. In each of these cases a homozygous parent was predicted by the other groups while Piphits correctly inferred heterozygosity for the parent, and so did not predict a mutation event. For example, site 6:113689210 in the YRI family was validated as a heterozygous 'TC' in both validation experiments. The original multiple-sequencing alignment for this site is shown in Figure 3-17. Considered independently, the father is apparently homozygous for the reference allele, while the daughter, individual 78, is clearly heterozygous. However when considered jointly, the few copies of the 'C' allele in the father allow the inference of an inherited allele, an event which was confirmed in the validation. The ability to infer the presence of an under-sampled parental allele rather than erroneously call a mutation is one of the strengths of the pedigree approach, and likely is

reflected in the low incidence of sites with validated heterozygous parents being assigned a high value of $\delta$ (Figure 3.12).

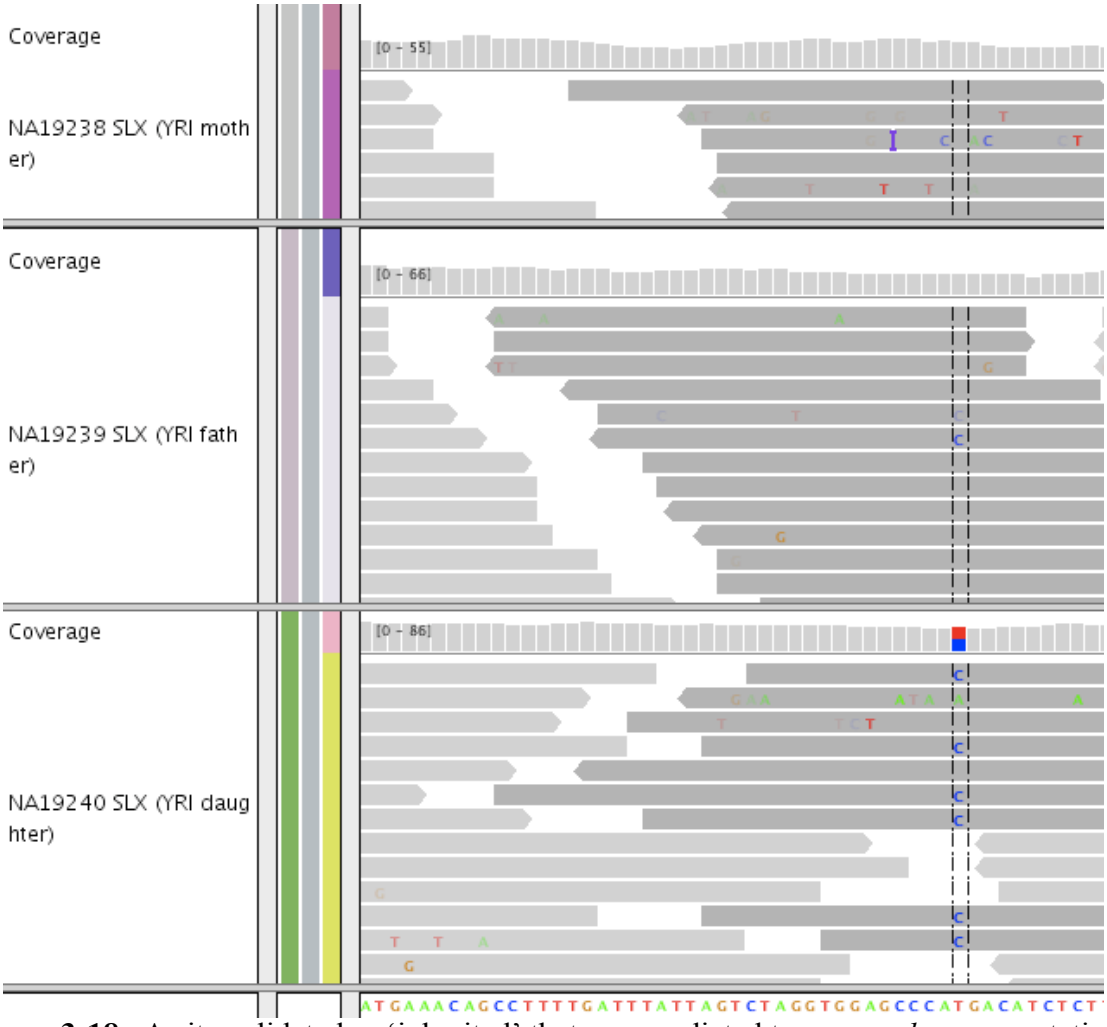



**Figure 3-18.** A site validated as 'inherited' that was predicted to carry a *de novo* mutation by the Broad and Sanger groups, but not by Piphits: 6:113689210 in the YRI family.

## Conclusion

Through collaboration the two trios making up Pilot 2 of the 1,000 Genome Project were queried for *de novo* mutations taking the form of Mendelian errors. The utility of this study can be seen in some of the observations made both within the process of locating and validating the mutation predictions as well as within the validation results themselves. First, it is clear that alignment error continues to play an important role in the analysis of high-throughput sequencing data. This was seen initially in the results of the attempted EM algorithm, and again within the false positive inspection. Closely tied to this is the need for improved calibration of the mapping and base qualities such that theses metrics consistently reflect the confidence of any particular base call or read mapping. The second conclusion of note is the relative abundance of validated "somatic" variants to germline variants. Whether the excess of these non-germline variants originate in cell-culture or true somatic tissue is yet to be determined, but in either case there will be implications on future study designs and applications of theoretical and applied human genetics. The validated results from the two families were used to generate independent rough estimates of the human germline mutation rate, effective population sizes coupled with the Piphits maximum-likelihood estimate of the population mutation rate, and the transition-transversion ratio. The per-base, per-generation mutation rate as well as the effective size estimates were larger than those typically found by other studies, while the transition-transversion ratio was likely under-estimated. These suggest this particular data set requires more refinement before fully reliable estimates can be produced. Finally, with respect to the Piphits method

in particular, further work should be done to improve the handling of alignment error and non-uniformity in the confidence of read observations. This will most likely have effects on the compression algorithm employed and the independent treatment of errors on neighboring sites may require more attention. From a broader perspective, this study has shown the ability of statistical methods applied to the high throughput sequencing of nuclear family to discover both novel, inherited variation and *de novo* mutation events. The notable contribution by the effort of each group, using independent discovery methods, to the overall validated result highlights the value of collaboration and inclusion of multiple prediction lists.

# References


1. Manolio, T.A. et al. Finding the missing heritability of complex diseases. in *Nature* Vol. 461 747-53 (2009).

2. Lango, H. et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* **57**, 3129-35 (2008).

3. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. in *Trends Genet* Vol. 17 502-10 (2001).

4. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? in *Am J Hum Genet* Vol. 69 124-37 (2001).

5. Li, H. et al. The Sequence Alignment/Map format and SAMtools. in *Bioinformatics* Vol. 25 2078-9 (2009).

6. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**, 1851-8 (2008).

7. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-94 (1998).

8. Shendure, J. & Ji, H. Next-generation DNA sequencing. in *Nature Biotechnology* Vol. 26 1135-45 (2008).

9. Chakravarti, A. Population genetics--making sense out of sequence. in *Nat Genet* Vol. 21 56-60 (1999).

10. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. in *Genetics* Vol. 156 297-304 (2000).

11. Xue, Y. et al. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. *Current Biology*, 1-5 (2009).

12. Tenesa, A. et al. Recent human effective population size estimated from linkage disequilibrium. in *Genome Research* Vol. 17 520-6 (2007).

13. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27**, 182-9 (2009).

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

15. Frank, S. Evolution in Health and Medicine Sackler Colloquium: Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. in *Proc Natl Acad Sci USA* (2009).

16. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107**, 961-8 (2010).

17. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. in *Nat Rev Genet* Vol. 10 195-205 (2009).

18. Venter, J.C. et al. The sequence of the human genome. in *Science* Vol. 291 1304-51 (2001).

19. Lander, E.S. et al. Initial sequencing and analysis of the human genome. in *Nature* Vol. 409 860-921 (2001).

# Chapter IV

# Discovery of Somatic Mutations within the Sequenced Exome of an Acute Lymphoblastic Leukemia

# Introduction

Using while blood cells sampled from a child diagnosed with acute lymphoblastic leukemia (ALL), the full exome was captured and resequenced with the goal of uncovering novel pathological genetic variation. After the individual was treated and no longer presented with disease symptoms, a sample of healthy blood cells was taken and subjected to the same experimental protocol. Comparing the two samples provides an opportunity to locate genetic variants that may be involved in the carcinogenic phenotype. However at sites that these samples differ, how does one discern which allele is the mutation? By sampling the two healthy parents, the resulting four samples can be compared and it becomes possible to distinguish germline-inherited variants from newly acquired variants that occurred specifically in the leukemia or in the healthy sample. The pedigree relating the four samples is shown in Figure 4-1. This chapter presents functionally-annotated *de novo* point mutations found uniquely in the leukemia sample that may contribute to the cancer phenotype, discovered using three different methods. The first is a traditional single-nucleotide polymorphism (SNP) discovery approach, inferring genotypes independently and comparing the variants located in the leukemia sample with those found within the other three. The second and third uses the Piphits method presented in Chapter II to locate mutations via Mendelian error detection using the relatedness of the samples to jointly infer the genotypes at each locus. The Piphits method is applied both in a liberal and in a stringent mode to locate recurrent and novel mutations respectively.

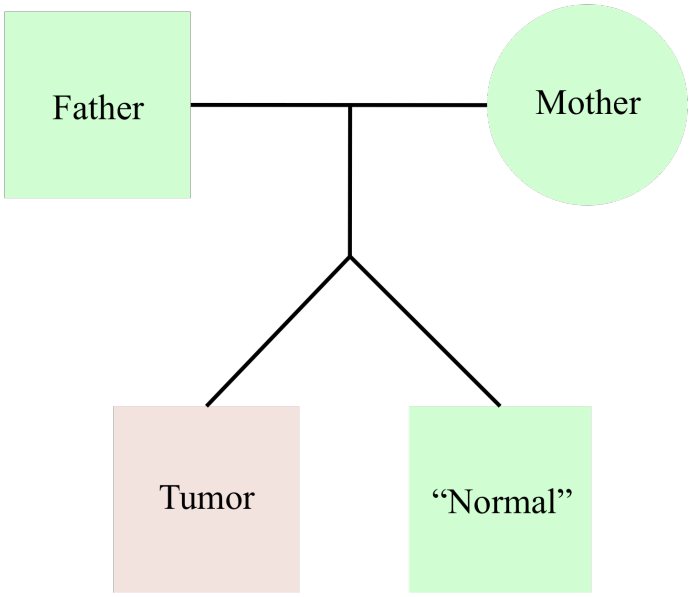



**Figure 4-1.** Pedigree relating the blood samples acquired. The samples from the three green-shaded individuals were identified as healthy, and the pink sample was cancerous.

## Methods

The genetic material sequenced from each blood sample was acquired using a targeted exon-capture technology commercialized by NimbleGen Systems, Inc.[1]. Although loci from many non-coding regions may be involved in disease susceptibility, capturing them requires contending with low complexity and highly repetitive regions of the genome. Focusing on the small fraction of the genome that is protein-coding serves as an optimization of sequencing and analysis resources towards regions that are most likely to produce results that have a clearly understood functional role. Ideally the entirety of a genome is examined for medically informative variation, but an argument can be made that

the current gap which exists between producing DNA sequence and garnering informative analysis can best be bridged by the exclusive sequencing of exons.

Specifically, the regions targeted by the sequence captured included the exons listed in the April 30th, 2008 build of the collaborative consensus coding sequence database (CCDS)[2], which lists exon coordinates relative to the HG18 human genomic reference build from UCSC. The captured DNA was sequenced using ABI SOLiD technology[3], resulting in millions of 50-base pair reads for each sample. The reads were aligned to the HG18 reference genome using the program Bowtie[4] which is an implementation of the Burrows-Wheeler Transform method, applied to color-space reads generated by the SOLiD machine. The specific details of this assembly are given in Table 4-1. Although sequence was only captured for gene regions, it is possibly many reads were misaligned outside of these regions given the number of bases across the genome that had coverage is up to five time the number of coding bases, approximately 30 megabases. Due to the spurious nature of these alignments, it is likely many false signals of mutation will be found outside of genic regions, and should be avoided.

**Table 4-1.** Quartet Bowtie alignment results.

| Statistic | Mother | Father | Leukemia | Normal |
|---|---|---|---|---|
| Total Reads | 45,858,958 | 43,085,010 | 50,461,237 | 61,872,173 |
| Fraction mapped to HG18 | .617 | 0.677 | 0.554 | 0.513 |
| Mean Read Depth | 9.3x | 11.8x | 11.1x | 15.0x |
| Bases covered by at least 1x | 153 million | 124 million | 126 million | 106 million |

The primary goal for this study is to locate potential driver or passenger point mutations within the leukemia sample genome. Driver point mutations are distinguished from passengers in that they contribute to the development of the cancer while passenger mutations are benign, yet get carried to higher clonal frequencies as the leukemia cells divide. Clearly differentiating the two requires intimate knowledge of the biochemical roles the gene products carrying the mutations play. While literature searching can be helpful in this regard, making this distinction with certainly will not be attempted within this text. Both driver and passenger types of mutations will be somatic in origin and must be differentiated from inherited germline mutations. Applying the Piphits method introduced in Chapter II to this data set provided an opportunity to probabilistically distinguish somatic and germline mutations. Conceptually the relationship between the four samples is identical to a pedigree with a single pair of monozygotic twins, with the

leukemia and normal child blood samples corresponding to two independent somatic samplings of the same zygote, each separated from the germline by an undetermined number of somatic cell divisions. The model representation of this application is shown in Figure 4-2. It is important to note the algorithm specifically is unaware of the expectation for spontaneous somatic mutations to accumulate in the leukemia sample compared with the others. It is therefore providing an unbiased query into the three possible locations of the somatic mutations: within the germline of either parent, inherited equally by both of the child samples, within the cell lineage leading to the somatic leukemia sample, and within the cell lineage leading to the normal sample. In general when the most likely genotypes are different between the two child samples, a somatic mutation in one or the other tissue will be inferred. When the two samples are heterozygous for the same alleles, while both parents are homozygous, a germline *de novo* mutation will be the most likely inference.

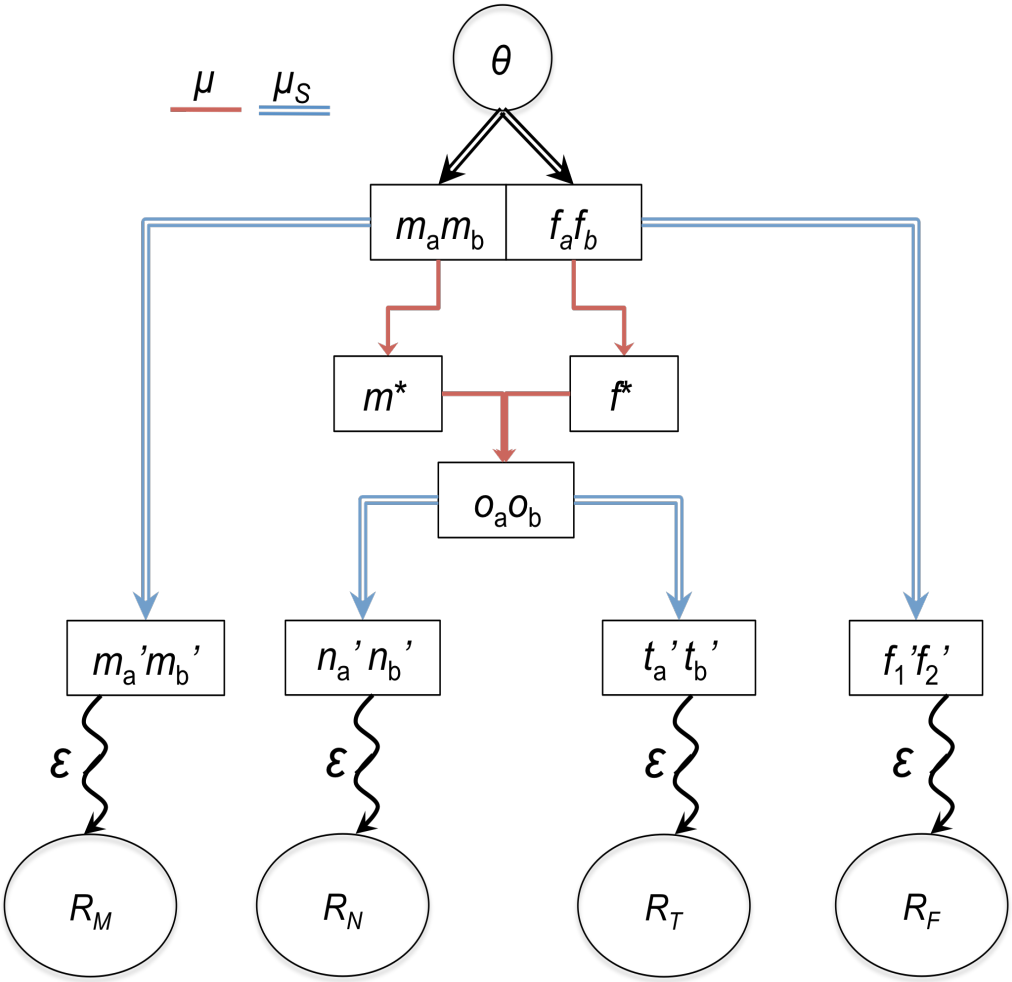



**Figure 4-2.** Pedigree model for the four blood samples. The sample of four parental alleles is drawn from a population with population mutation parameter $\theta$. The alleles inherited by the single zygote from the mother and father are $m^*$ and $f^*$ respectively. The child zygote is formed from the union of these two alleles, with the possibility of germline mutation. The true genotype of the blood samples taken are represented by $m'$, $f'$, $n'$, and $t'$, and the reads observed overlapping a particular site for each sample are given by $R_M$, $R_F$, $R_N$, and $R_T$. Although the possibility for somatic mutation exists between the mother and father germline cells and their respective somatic blood samples, this will be largely undetectable given the possibility of observing up to three different alleles in the original sample of four parental chromosomes.

*Somatic mutation discovery by testing the inheritance of single nucleotide polymorphisms (SNP-method)*

The first approach used to locate *de novo* mutations was based on typical genotyping methodology by constructing a consensus genotype from the aligned sequencing reads. This was performed by the running the alignment data through the SAMtools software, utilizing the MAQ genotype consensus model. To begin, high quality SNPs were discovered in the two parental samples. These SNPs were required to have a minimum *phred*-scaled quality of 30, with the non-reference allele observed on at least two reads. This resulted in the identification of 37,885 SNPs in the father sample, 44,934 SNPs in the mother sample, 42,530 SNPs in the leukemia sample, and 36,258 SNPs in the normal sample. Each variant discovered within the two child samples must either be inherited from one of the parents or the result of a *de novo* mutation. To determine the inheritance of the child sample polymorphisms, the original parent alignments were queried and the child SNPs were categorized as being either 'inherited', 'missing', or '*de novo'*. For those classified as inherited, the parental snip quality was greater than zero at the same genomic coordinate with non-reference allele observations matching that of the child SNP. The parental pileup had zero or less than 15x coverage at the genomic position for SNPs placed in the 'missing' category. As shown in Chapter II, mutations cannot reliably be called in when the depth in the parents is not sufficient to infer homozygosity or heterozygosity. All other SNPs, those with parental coverage greater than 15 and with SNP quality equal to zero, were classified as *de novo*. After performing this test for both child

SNP lists against both parental pileup files, 1,120 sites within the leukemia and 1,305 within the normal. At the 196 positions where the two lists intersect with the same genotype, the position was inferred as carrying a germline mutation. After this analysis 917 unique SNPs remained in the leukemia sample. To identify the potential for these positions to play functional role in cancer, each remaining SNP was annotated according to the Consensus Coding Sequence (CCDS) database[2] and the potential for deleterious effects was tested using the Sorting Tolerant From Intolerant (SIFT) algorithm[5]. SIFT predicts the potential effects of amino acid substitutions on protein function by classifying a mutation as either 'tolerated' or 'damaging' with an associated score, and has recently been successfully applied to locating deleterious somatic mutations within leukemia cells[6]. Annotation using SIFT revealed 416 SNPs were located within non-genic regions and so considered artifacts of spurious read alignments outside of the targeted exome, leaving 500 sites that were tentatively called somatic leukemia mutations. This culling process is illustrated in Figure 4-3.

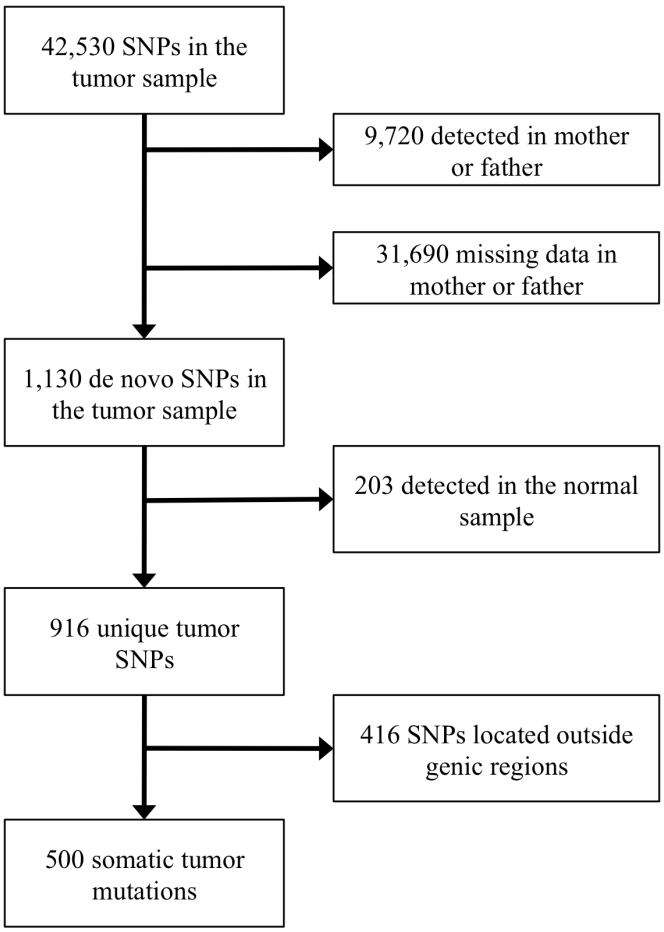



**Figure 4-3.** Procedure for culling somatic mutations from the leukemia SNP list.

### *Somatic mutation discovery using two applications of Piphits*

The Piphits method was used in two different manners to locate somatic mutations: one to target possible recurrent somatic mutations and one to uncover novel driver mutations. In the first, termed Piphits-CGC, Piphits was executed using the results of the Expectation-Maximization algorithm, which, as was observed in Chapter III, produced somatic and germline mutation rate estimates that were several orders of magnitude higher than expected. These estimates are shown in Table 4-2. Using these high rates, Piphits is

very sensitive to a small signal of mutation. In this mode a lower frequency of reads bearing a non-reference allele relative to the reference is necessary to allow inference of mutation. Using abnormally high mutation rates within the Phips method will produce many more false positive predictions, but can also serve to capture true mutation sites that have read patterns that carry a very small signal of mutation. Rather than querying the entire sequenced exome, Piphits was applied in this hyper-sensitive mode only to genes that had been previously descried to harbor recurrent cancer-related mutations. Sites having a posterior probability of mutation, $\delta$, greater than 0.01 were kept for further analysis. The putative cancer genes examined were defined by the February 2, 2010 version of the Cancer Gene Census[7] (CGC) list hosted by the Wellcome Trust Sanger Institute. This query resulted in 238 putative mutations within cancer-related genes. By requiring a minimum of 15x coverage in both parents and allowing at most one parental read matching the putative mutant allele, this list was further culled to 14 putative recurrent mutations within cancer-related genes.

**Table 4-2.** Prior model parameters used for the applications of Piphits-CGC and Piphits-EW.

| Pipits method priors on model parameters | Piphits-CGC | Piphits-EW |
|---|---|---|
| Germline mutation rate $\mu$ | $1.44 \times 10^{-5}$ | $2.00 \times 10^{-8}$ |
| Somatic mutation rate $\mu_S$ | $5.78 \times 10^{-5}$ | $4.00 \times 10^{-7}$ |
| Population mutation rate $\theta$ | $5.30 \times 10^{-4}$ | $1.00 \times 10^{-3}$ |
| Sequencing error rate $\varepsilon$ | $5.29 \times 10^{-4}$ | $5.00 \times 10^{-3}$ |

In the second application of Piphits to this data set, termed Piphits-EW, the algorithm was executed in a similar method to that described in Chapter III applied to sites available exome-wide. As introduced in Chapter II, the probability of mutation measure $\delta$ was calculated for all sites loaded into the compression exome-wide. In this mode parameters were chosen to roughly approximate expected rates, resulting in a more conservative set of mutation predictions. For all examined sites the prior value on the germline mutation rate was set to $2.0 \times 10^{-8}$, the somatic mutation rate to $4.0 \times 10^{-7}$, the sequencing error rate to 0.005, and the population mutation parameter was held constant at 0.001. The somatic rate was elevated to the germline rate due to the greater number of opportunities for somatic mutation within the pedigree and the expectation that malignant cells would carry somatic mutations. After executing Piphits, putative mutations were selected which had $\delta$ values greater than 0.01. The same rules were applied to this list as the previous one, requiring a minimum of 15x coverage in both parental samples and a

maximum of one read carrying the mutation were observed in either parent. This process resulted in a list of 10 candidate somatic mutations.

After the candidate mutation lists were produced by these three approaches, from this point on referred to solely as the SNP-method, Piphits-CGC, and Piphits-EW respectively, they were treated equally with the same set of rules for selecting potential driver mutations. First sites located within dbSNP[8] were segregated and included only if the observed mutant allele did not match an allele of high population frequency, tested by observing the mean weighted average of allele frequencies across all HapMap samples[9,10]. The SIFT tool was used heavily to annotate potentially harmful variants, although both 'tolerated' and 'deleterious' mutants are presented. A larger focus is placed on the current body of literature describing the genes harboring putative mutations, tolerated, deleterious, and synonymous changes alike. The literature support for any given gene referenced is drawn from four primary sources: the Online Mendelian Inheritance in Man database[11] (OMIM), GeneCards[12], the Gene Ontology database (GO)[13] queried using the g:Profiler tool[14], and COSMIC[15], the Catalog Of Somatic Mutations In Cancer. In general the strategy for selecting driver mutation candidates is to discuss generally the number of mutations discovered within each class and specifically mention affected genes that have literature support for playing a role in one or another form of cancer. Slight emphasis will be placed on nonsynonymous (missense) mutations over synonymous (silent) ones. Many genes will be suggested which may play no specific role within ALL, but have been associated with other forms of cancer, giving the expert reader the ability to draw the most

informed conclusions, as well as leaving open the door for creating novel connections between ALL and other cancers.

## Results

By far the most prolific method used for predicting somatic mutations was the SNP-calling method. Pooled together, Piphits-CGC and Piphits-EW predicted a total of 24 somatic mutations while the SNP-method produced ten times as many. Whether this discrepancy is primarily the cause of lack of specificity within the SNP method or sensitivity with the Piphits approaches shall be discussed in subsequent sections with the suspicion that neither method is performing optimally. The mutations will be presented beginning with the SNP-method results and gradually shifting to the Piphits predictions, noting where the approaches overlap. Using literature searches, GO categorizations, and the SIFT predictions as a guide, sites within genes not part of the CGC were individually selected from all classes of mutations for more detailed analysis. This process was not exhaustive and certainly may leave significant mutations undescribed, but all mutations identified are available for future analysis listed in the Appendix.

*SNP-method mutation predictions for genes listed by the Cancer Gene Census.*

As shown in Figure 4-3 there were 500 putative somatic mutations following the removal of predicted variants within non-genic regions. In the broadest terms, this body of predicted *de novo* mutation was made up of 150 synonymous and 310 nonsynonymous SNPs, with an additional 9 in introns and 31 located within untranslated regions (UTR). Sites overlapping with dbSNP were set aside for the moment, as were the 9 intronic sites, leaving 28 UTR sites, 127 synonymous, and 280 nonsynonymous positions. Initially mutations located within genes within the CGC list were examined. Eight genes carrying a set of 11 putative mutations were found within the list including *BMPR1A*, *MLLT4*, *NOTCH2, NR4A3*, *PDE4DIP*, *RBM15*, *RET*, and *SETD2*. A description of each of these genes and all others specifically mentioned are included in the Appendix of this chapter with literature references. Details for each of these mutations are given in Table 4-3. The relative confidence in each mutation call is given by the SNP quality metric. Since the CGC was also the list of genes specifically targeted by Piphits-CGC, some overlap between predictions is expected. The mutations predicted in *MLLT4*, *NOTCH2*, *NR4A3*, and *SETD2* were predicted by both methods, with the Piphits-CGC $\delta$ value greater than 0.40. Of these four the predicted deleterious mutations in *MLLT4*, *NOTCH2*, and *SETD2* are perhaps the most interesting. Abnormalities within the Myeloid/Lymphoid or Mixed Lineage Leukemia gene (*MLL*) are associated with a poor prognosis and are frequently seen in pediatric ALL patients[16,17]. Deregulation of *NOTCH2* has been suggested to play a critical

role in the malignancy of chronic lymphocytic leukemia cells[18]. Finally *SETD2*, a histone methyltransferase involved in transcriptional elongation[19], has contributed to renal carcinoma when subjected to inactivating mutations[20]. Dual, independent predictions of the same mutations within genes clinically relevant to ALL is encouraging for the utility of this mutation discovery procedure.

**Table 4-3.** SNP-method predicted mutations at sites within genes within the Cancer Gene Census list. Coord. denotes genomic position, Base Subs the reference followed by mutant allele, SIFT Pred. prediction of mutation effect on protein: Del. Deleterious, Tol. Tolerated, Q the MAQ genotype consensus SNP quality.

| Coord. | Gene Symbol | Base Subs. | SNP Type | SIFT Pred. | Q |
|---|---|---|---|---|---|
| 6:168092005 | MLLT4 | C/A | Missense | Del. | 47 |
| 1:120260690 | NOTCH2 | G/A | Missense | Del. | 37 |
| 3:47137269 | SETD2 | C/A | Missense | Del. | 54 |
| 1:120374070 | NOTCH2 | T/C | Missense | Tol. | 81 |
| 1:143565938 | PDE4DIP | T/C | Missense | Tol. | 54 |
| 1:143592964 | PDE4DIP | G/T | Missense | Tol. | 42 |
| 1:143727234 | PDE4DIP | G/T | Missense | Tol. | 108 |
| 10:88671362 | BMPR1A | C/T | Silent | None | 42 |
| 9:101646889 | NR4A3 | A/G | Silent | None | 53 |
| 1:110686028 | RBM15 | G/A | Silent | None | 32 |
| 10:42933850 | RET | C/A | Silent | None | 77 |

## *SNP-method prediction of nonsense mutations*

There were 13 nonsense mutations predicted among the 281 undescribed nonsynonymous mutations, creating novel stop codon which prematurely truncates protein

translation. Due to the major effects these mutations can play on protein function, each of these 13 was the subject of a literature search using the tools mentioned above. This inquiry led to the identification of three notable mutations within putative genes of interest *STARD8*, *WNT4*, and *YY1AP1*. *STARD8*, also known as Deleted in Liver Cancer-3 (*DLC3*), is thought to be a leukemia suppressor gene whose down-regulation or deletion has been implicated in prostate, breast[21], and colorectal cancers[22]. The canonical *WNT* signaling pathway is a fundamental player in eukaryotic organisms[23], important in the processes of cell differentiation, proliferation, and potentially the oncogenesis of stem cells[24]. *YY1AP1* is a co-activator of the *YY1* gene[25], which has been found to be differentially expressed in ALL paitents[26]. These three nonsense mutations along with the other 10 predictions are shown in Table 4-4.

**Table 4-4.** Putative nonsense mutations predicted by the SNP-method. Codon is given in the same orientation as the transcription of the gene. Lower-case base in the Codon Substitution column is the mutant allele. Mutations in bold face are mentioned in the text.

| Coord. | Gene Symbol | Base Subs. | Codon Subs. | Residue # | Q |
|---|---|---|---|---|---|
| 10:17948677 | MRC1L1 | G/A | TGG-Tga | 628 | 108 |
| 6:155168298 | RBM16 | C/T | CAA-tAA | 323 | 93 |
| 3:19915283 | EFHB | C/A | GAG-tAG | 391 | 50 |
| **X:67860890** | **STARD8** | **G/A** | **TGG-Tga** | **991** | **48** |
| **1:153905101** | **YY1AP1** | **G/A** | **CAA-tAA** | **320** | **44** |
| 2:71642552 | DYSF | G/A | TGG-TGa | 775 | 37 |
| 1:153227483 | FLAD1 | G/A | TGG-TGa | 217 | 36 |
| 2:40510531 | SLC8A1 | C/A | GAA-tAA | 132 | 35 |
| **1:22328852** | **WNT4** | **G/A** | **CAG-tAG** | **53** | **34** |
| 15:26680796 | HERC2P2 | C/T | CGA-tGA | 103 | 34 |
| 7:23591599 | CLK2P | G/T | TGC-TGa | 362 | 32 |
| 4:48688324 | AC020593.1 | G/A | TGG-TaG | 111 | 32 |
| 1:175516839 | FAM5B | G/A | TGG-TaG | 635 | 31 |

## *Mutations within untranslated regions predicted by the SNP-method*

In addition to coding mutations, variants located in UTR's were considered important due to the possibility of cancer-related post-transcriptional regulation of gene expression by micro-RNAs[27,28]. The SNP-method predicted 28 mutations within these regions of interest, and six were tentatively identified as relevant after literature review.

These included *AIM2*[29], *MST1*[30,31], *SP140*[32], *NUP214*[33], *APITD1*[34], and *DGKG*. The mutations in *AIM2*, *APITD1*, and *MST1* are located in the 3' UTR and the rest are in the 5' UTR. *SP140* and *NUP214* are perhaps particularly interesting since they have specifically previously been linked with leukemia[32,35]. These and all other *de novo* mutations predicted by the SNP-method in UTR are shown in Table 4-5.

**Table 4-5.** Mutations predicted by the SNP-method within untranslated regions. Bold-face indicates literature search revealed an association with cancer.

| Coord. | Gene Symbol | Base Subs. | UTR | Q |
|---|---|---|---|---|
| 5:162862865 | *MAT2B* | T/C | 5' UTR | 151 |
| 9:123114435 | *GSN* | C/T | 5' UTR | 58 |
| 16:33687838 | *AC133561.4* | C/T | 5' UTR | 50 |
| **2:230798735** | ***SP140*** | **T/C** | **5' UTR** | **43** |
| 11:33139524 | *CSTF3* | C/T | 5' UTR | 42 |
| 20:1112506 | *C20orf46* | C/A | 5' UTR | 41 |
| 1:16684733 | *KIAA1922* | C/T | 5' UTR | 38 |
| 19:19487793 | *NDUFA13* | C/G | 5' UTR | 38 |
| 4:8262328 | *SH3TC1* | G/T | 5' UTR | 38 |
| 3:42924598 | *ZNF662* | G/C | 5' UTR | 37 |
| 1:217413841 | *LYPLAL1* | G/T | 5' UTR | 34 |
| **9:132990876** | ***NUP214*** | **G/A** | **5' UTR** | **34** |
| **3:187520970** | ***DGKG*** | **G/A** | **5' UTR** | **31** |
| 5:64483394 | *ADAMTS6* | G/T | 3' UTR | 109 |
| 19:49582472 | *ZNF806* | A/G | 3' UTR | 91 |
| 15:26715679 | *HERC2P2* | C/A | 3' UTR | 76 |
| 4:174489801 | *HMGB2* | G/A | 3' UTR | 67 |
| 7:48935428 | *CDC14* | G/A | 3' UTR | 64 |
| **1:157299041** | ***AIM2*** | **T/C** | **3' UTR** | **62** |

**Table 4-5.** Continued

| | | | | |
|---|---|---|---|---|
| 1:89291610 | *GBP1* | A/G | 3' UTR | 58 |
| 16:29383411 | *SULT1A4* | A/G | 3' UTR | 54 |
| 2:119914126 | *SCTR* | C/A | 3' UTR | 50 |
| 2:63973990 | *VPS54* | G/A | 3' UTR | 47 |
| **1:16848695** | ***MST1*** | **G/A** | **3' UTR** | **44** |
| 2:198719939 | *PLCL1* | C/A | 3' UTR | 42 |
| 9:69724143 | *CBWD5* | T/C | 3' UTR | 41 |
| **1:10425059** | ***APITD1*** | **C/T** | **3' UTR** | **32** |
| 2:24898033 | *CENPO* | A/G | 3' UTR | 31 |

## *Mutations at dbSNP loci predicted by the SNP-method*

Predicted mutation events that occurred at genomic positions overlapping with dbSNP were considered valid only when the observed mutant allele did not match one of the dbSNP alleles known to occur at high frequencies, referencing the weighted allele frequencies of HapMap populations reported by the SIFT algorithm. The rationale behind this strategy was to consider rare mutant alleles at described polymorphic sites potential recurrent mutations. If the mutant allele was of high frequency, inheritance from an unidentified heterozygous parent is the more likely source. Of the 56 candidate somatic mutation sites overlapping dbSNP, 22 had an allele that was not known to be common. Within these 22 sites were 11 missense and 8 silent mutations, and 3 located within the UTR of a gene. Using the GeneCards and OMIM databases, literature search revealed one missense mutation and one silent mutation within two relevant genes: *POTEC*[36] and *MUC17*[37] respectively. Neither of these genes are located in CGC although *MUC17* is

represented in COSMIC as part of a sequenced lung cancer genome[38]. Expression of *POTEC* has been used as a diagnostic for prostate cancer[36], and *MUC17* has been shown to be over expressed in pancreatic cancer[37]. These two mutations as well as the other rare-allele putative mutations overlapping dbSNP positions are shown in Table 4-6. The Piphits-CGC method predicted a silent mutation within the *AKAP9* gene at a site found within dbSNP. The mutant allele, G, is reported to have a frequency of 0.38, so this site does not meet the same criteria for reporting as those within *POTEC* and *MUC17*. However since *AKAP9* is within the CGC and has been found to associate with many forms of cancer[39-41], its mention here is warranted.

**Table 4-6.** Predicted mutations at sites discordant with dbSNP most frequent population alleles. SIFT Pred. denotes prediction of mutation effect on protein: Del. Deleterious, Tol. Tolerated. In bold font, *POTEC* and *MUC17*, likely play active roles in leukemia formation[36,37].

| Coord. | Gene Symbol | Base Subs. | SNP Type | SIFT Pred. | dbSNP ID | Population Allele Frequency | Q |
|---|---|---|---|---|---|---|---|
| 10:95710491 | *PIPSL* | T/C | 5' UTR | None | rs12571819 | NA | 182 |
| 4:321658 | *ZNF141* | T/C | 5' UTR | None | rs3749520 | NA | 56 |
| 8:7754120 | *SPAG11A* | G/T | 5' UTR | None | rs2853655 | NA | 44 |
| 5:149192436 | *PPARGC1B* | G/C | Missense | Del. | rs7732671 | A(0.77), G(0.23) | 133 |
| 1:147170201 | *DRD5P2* | A/G | Missense | Del. | rs4004791 | C(0.78), T(0.22) | 76 |
| **15:19331099** | ***POTEC*** | A/G | **Missense** | **Del.** | **rs4578610** | **T(0.69), C(0.31)** | **63** |

**Table 4-6.** Continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2:234302651 | *UGT1A8* | T/C | Missense | None | rs6431625 | G(0.59), T(0.41) | 59 |
| 19:60042760 | *KIR2DS4* | A/T | Missense | Tol. | rs4806589 | G(0.48), A(0.52) | 162 |
| 14:77210108 | *ALKBH1* | T/A | Missense | Tol. | rs6494 | C(0.82), T(0.18) | 94 |
| 11:56224788 | *OR9G9* | G/A | Missense | Tol. | rs591369 | C(0.78), T(0.22) | 87 |
| 7:74196333 | *GTF2IRD2B* | C/A | Missense | Tol. | rs2539034 | G(0.83), C(0.17) | 87 |
| 9:106401420 | *OR13C5* | G/C | Missense | Tol. | rs6479260 | A(0.62), G(0.38) | 74 |
| 20:1843889 | *SIRPA* | G/C | Missense | Tol. | rs16997190 | G(0.48) ,A(0.52) | 69 |
| 16:73501222 | *WDR59* | A/G | Missense | Tol. | rs5023505 | T(0.69), C(0.31) | 63 |
| **7:100465092** | ***MUC17*** | T/C | **Silent** | **None** | **rs11769823** | **G(0.76) ,A(0.24)** | **228** |
| 17:34180293 | *PIP5K2B* | G/A | Silent | None | rs228290 | T(0.69), C(0.31) | 184 |
| 1:200832571 | *SYT2* | G/A | Silent | None | rs504261 | T(0.74), C(0.26) | 104 |
| 16:16278820 | *NOMO3* | C/A | Silent | None | rs393246 | T(0.69), C(0.31) | 80 |
| 17:24914089 | *AC104564.11-1* | C/G | Silent | None | rs721479 | T(0.69), C(0.31) | 80 |
| 11:6587409 | *ILK* | G/A | Silent | None | rs1043390 | C(0.78), T(0.22) | 63 |
| 6:26478551 | *BTN3A2* | C/T | Silent | None | rs34878490 | C(0.83), G(0.17) | 51 |
| 1:89424666 | *GBP4* | A/G | Silent | None | rs608339 | A(0.32), C(0.68) | 42 |

## *Specific effects of a coding mutation predicted exclusively by Piphits-CGC*

Rather than continue the endless listing of genes found within different categories with minimal information provided for each, the focus will instead turn to the putative effects of a single mutation discovered exclusively by Piphits-CGC, including support for

the mutation call and reasons the other methods used did not predict this mutation. Another nonsense mutation, it has the potential for serious loss-of-function effects in the resulting protein product. Because this mutation was predicted by the Piphits-CGC method, has been previously described to play a role in cancer.

In this case the gene is the interleukin 6-signal transducer (*IL6ST)*, also commonly referred to as glycoprotien 130 subunit (GP130)[42,43]. The protein product of this gene acts as a shared signal transducer for several cytokines, intercellular signaling molecules between immune cells, and is central to a signaling cascade which has been used as a model for the cytokine signaling system in general[44]. The cytokines sharing *IL6ST* most prominently referenced include interleukin 6 (*IL6*), interleukin 11 (*IL11*), leukemia inhibitory factor (*LIF*), and oncostatin M (*OSM*), along with up to ten others[45-47]. *IL6* has been the subject of much inquiry and debate with reports differing on its ability as a promoter or suppressor of leukemia growth[48-50]. However, the increased activity of *IL6* has been recently verified to mediate proliferation and prevent inflammatory-induced apoptosis of normal and premalignant cells alike[51,52]. This would suggest a deleterious effect of a down-stream signal transducer of *IL6* would not be associated with tumorogenisis. Conflicting reports exist also for *OSM*, which was originally described as a growth regulator[53], now has been shown to both support or inhibit leukemia growth depending on cell type[54]. *LIF*, which is potentially a recently formed paralog of *OSM*[55], is primarily involved in embryo development and implantation during human pregancy[56], but also may support leukemia growth[57] and cell proliferation when expressed in adult epithileal cells[58].

The 'C' to 'A' substitution predicted in the first position of the codon at the 508th amino acid of the *IL6ST* gene results in a codon change from 'GAA' to 'TAA' in coding orientation, replacing the code for glutamic acid. The complete *IL6ST* protein includes 918 amino acids with the 22-residue transmembrane domain beginning at the 640th amino acid[42]. In addition, a tyrosine located at position 759 has been shown to be a significant site for phosphorylation and binding with *OSM* and *LIF*[59]. As a trans-membrane signal transducer, the truncation of this gene premature to the translation of these regions would have severe effects on function, possibly producing a completely non-functioning protein product. The precise effect this would have upon the signaling pathways mentioned above within the leukemic phenotype of the leukemia sample is difficult to determine, particularly given the heterozygous genotype of the leukemia sample. One functioning copy of the *IL6ST* gene may provide sufficient dosage to the system to promote the oncogenic effects of these cytokines in epithileal cells in particular.

The level of confidence in the *IL6ST* mutation prediction has yet to be addressed. The Piphits-CGC method predicted the mutation with a confidence measure of $\delta$=0.866. The observed read data structure is seen below, in Table 4-6. Although the coverage level in the leukemia sample was low at this site, with only eight reads, two of the eight or 25% contained an 'A', the non-reference allele. As seen in other applications of Piphits, sufficient coverage to confidently call homozygosity in other individuals in the pedigree lend support to a mutation inference for a potentially heterozygous individual when the proportion of mutant reads is greater than 0.20, even when the total read depth at the site

for the polymorphic individual is very low. The base calls of the two heterozygous supporting reads were observed to be of high quality. Although this detailed assessment of the alignment data supports the genotype calls, the low coverage of the mutant allele leaves room for doubt. More experimental evidence for heterozygosity of the leukemia sample will be necessary before any final verdict can be reached on the role the truncation of *IL6ST* may play as a driver mutation in ALL.

**Table 4-7.** Observed read data supporting mutation call at 5:55283865, E508* nonsense mutation in *IL6ST*, ubiquitous transmembrane signal transducer in the interleukin-6 cytokine signaling pathway.

| Observed Nucleotide at a Single Site | Adenine (A) | Cytosine (C) | Guanine (G) | Thymine (T) |
|---|---|---|---|---|
| Mother Sample Reads | 0 | 28 | 0 | 0 |
| Father Sample Reads | 0 | 37 | 0 | 0 |
| Normal Sample Reads | 0 | 22 | 0 | 0 |
| Leukemia Sample Reads | 2 | 6 | 0 | 0 |

# Discussion

## *Discrepancies between the predicted mutation lists from each method*

The greatest cause for concern in this mutation finding experiment was the lack of overlap in the three mutation calls in general, and the paucity of Piphits-produced calls relative to the abundance of mutations predicted by the SNP-method. Piphits-CGC

produced a total of Piphits-EW produced 27 mutation predictions. The *IL6ST* mutation can serve as an example since it was only called by the Piphits-CGC method. Why did the SNP-method fail to call the mutation when specifically examining the CGC genes? Why Piphits-EW fail to predict this mutation? Answering these questions will provide a framework for assessing the confidence of the mutation calls presented here as a whole.

The SNP-method applied a simple heuristic to call the mutations based on observed coverage levels and the SNP-quality metric for each site individually across the four samples. In brief, a leukemia somatic mutation was predicted when all of the following events were observed for a single site: the leukemia sample was heterozygous with a SNP-quality greater than 30 and at least two observations were made of the non-reference allele, both parents were found to have at least 15x coverage of the reference allele and the SNP-quality was equal to 0, and the normal sample did not share the SNP. In the particular case of the *IL6ST* mutation, the SNP-quality was equal to 19 in the leukemia sample. Therefore application of the hard SNP-quality cutoff eliminated this site from downstream analysis. In addition to the high-quality base calls at this site, six additional bases were called matching the reference: five with quality 0 and one with quality 3. By incorporating these bases into the quality calculation, the overall consensus-quality, from which the SNP-quality is derived, was penalized. Piphits ignored these low-quality reads during data compression so they were not factored into the genotype calculation. Unifying the calls made by the two methods would amount to either removing the stringent SNP-quality filter from the SNP-method or removing the base-quality filtering from Piphits. In either case

the overall quality of the body of predictions made by either method would have suffered at the expense of unification at a few sites. It is of little utility to apply two separate methods that abandon specificity in the service of concordance of results.

Both executions of Piphits used the same threshold for base quality cut-offs. Therefore where one had the opportunity to call a mutation, the other did as well. The primary distinction between Piphits-CGC and Piphits-EW execution was the selection of prior parameters, as show in Table 4-2. With a somatic mutation rate approximately 100-fold smaller and an error rate 10-fold higher the Piphits-EW confidently inferred the two observed 'A' bases were the result of sequencing errors ($\delta \approx 0.0005$). In general the strategy for applying such drastically different prior rate parameters was two-fold. First, the rates used for Piphits-CGC offered an empirical test of using the EM-produced parameters to infer mutation within the original data set producing the estimates. Second, by restricting the list of sites examined, sensitivity was kept high in these regions of particular interest without suffering the consequences of a poor specificity exome-wide. Before culling the Piphits-CGC candidate mutation list using the CGC list, the method had predicted mutations at over 34,000 sites with $\delta$>0.01. In order to most efficiently locate mutations within the remainder of the exome, Piphits-EW resorted to the more conservative parameter estimates used in earlier chapters. As designed, the sensitivity to putative mutation sites such as the one discussed in *IL6ST* was sacrificed in the application of Piphits-EW in favor of producing a more reliable list of exome-wide mutation candidates.

*Assessment of SNP-method compared to Piphits in predicting mutations in the quartet sample*

A natural question to ask is why the effort of multiple executions of Piphits is necessary, particularly given the meager reward of 24 predicted mutations out of the complete sequenced human exome. In one sense this could be a highly desirable result assuming the true number of mutations within the data set is close to this number. However the amount of putative mutation calls inferred by the SNP-method and the number of somatic mutations expected within leukemia genomes in general suggest otherwise[60-62]. If the goal is to capture as many of the true mutations as possible while keeping the candidate list to a manageable size from which biologically relevant predictions can be easily be garnered, certainly the SNP-method seemed to out-perform the implementation of Piphits. In short, the root of the drawback of Piphits when applied to this data set in particular, and to potentially any larger pedigree studied with high-throughput sequencing, is the departure of the systematic sampling of chromosomes at true heterozygous sites from the probabilistic model assumed by Piphits. This effect may be small when dealing with trio data, but grows linearly with the number of somatic samplings taken, and seems to dwarf the true signal of mutation in this quartet sample. The SNP-method, relying primarily on observation of coverage and the SNP-quality metric, places minimal model-based restrictions on the sampling pattern alleles, so inconsistent samplings for one member of the pedigree do not affect genotype inference in the other family members. This is illustrated in Figure 4-4 were the confidence metric is plotted against the

observed frequency of the putative mutant allele in the leukemia sample. The MAQ consensus genotype model, which is the source of the SNP-quality metric, confidently calls mutations when there are large departures from the 50% sampling of heterozygous alleles, and instead places more weight on the specific qualities of each sampled base. This particular distinction was observed in the different treatment of the alignment data at the *IL6ST* nonsense mutation prediction, where the presence of 6 poor-quality reads rather than the incongruous sampling of chromosome alleles prevented the heterozygous call. Until sequencing and alignment technologies improve towards producing the ideal data set, which in some ways is what is modeled by the current implementation of Piphits, calling of mutations may best be accomplished by relying on base and mapping qualities to provide more information for the confidence in genotype calls rather than specifically testing the distribution of observed reads without fully taking quality into account.

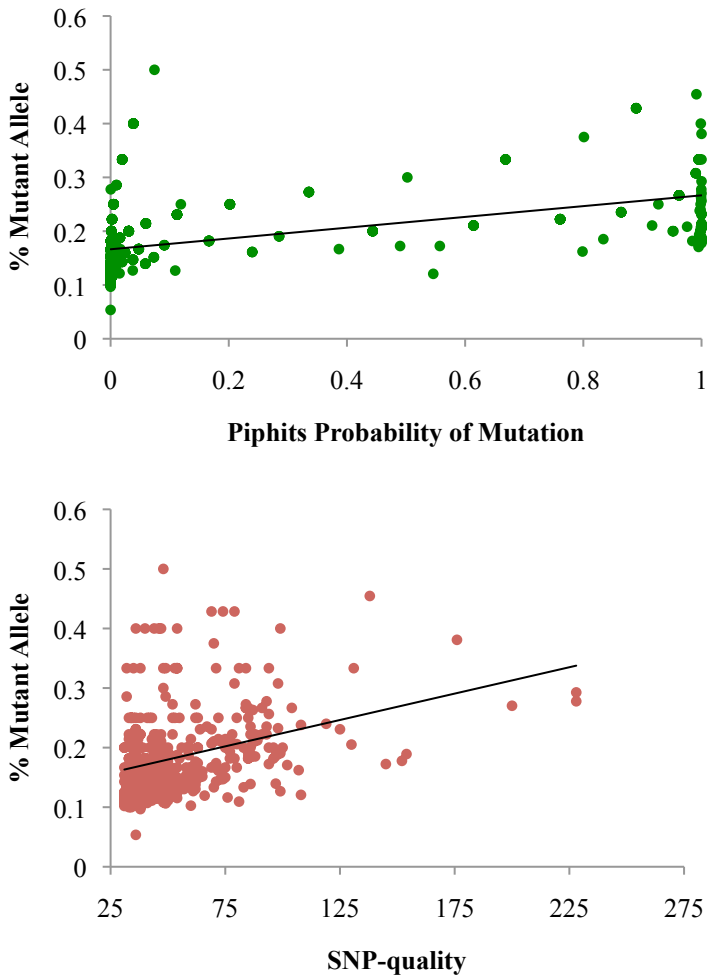



**Figure 4-4.** SNP-quality and Piphits probability of mutation plotted against the percent coverage in the leukemia sample. The sites included were the 500 somatic mutations predicted by the SNP-method. While the Piphits method clusters sites at the maximum and minimum confidence values based purely on the fit of the observed data to the modeled expectation of read sampling, the SNP-quality relies upon the sequence base qualities to present a spread of values across the confidence metric that is less reliant on the observed sampling pattern of reads.

# Conclusion

This study has made two primary accomplishments. The first is the presentation of a list of putative, annotated *de novo* somatic mutations within coding sequences of an ALL genome, largely void of diversity that is either inherited from the parents of the patient or is shared with a healthy blood sample. Rather than a final definitive list of predicted driver mutations, the mutations predicted serve as a resource for deeper inquiry into the specific mutational features of this particular leukemia. This investigation can be expedited by taking advantage of the SIFT annotations presented with each prediction for distinguishing tolerated amino acid substitutions from those that are more likely to be damaging. The second utility of this study has been to test the current formulation of the Piphits method against an expanded pedigree in which somatic and germline mutations can be distinguished. Although the two applications of Piphits were able to make some potentially informative predictions, the larger indication is that there is work to be done in the implementation to make it as effective as standard genotyping procedures which take advantage of the full range of quality information associated with read observations, rather than placing stringent requirements on sampling patterns as the primary indication of heterozygosity. Despite these implementation issues, the fundamental premise of Piphits remains valid, that by taking advantage of the relatedness of individuals within a pedigree when inferring genotypes, the potential exists for increased power to locate departures from

Mendelian inheritance and infer germline and somatic mutations, and to directly, jointly estimate their per-nucleotide rates.

# References


1. Hodges, E. et al. Genome-wide in situ exon capture for selective resequencing. in *Nat Genet* Vol. 39 1522-7 (2007).

2. Pruitt, K.D. et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* **19**, 1316-23 (2009).

3. McKernan, K.J. et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. in *Genome Research* Vol. 19 1527-41 (2009).

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

5. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).

6. Mardis, E.R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. in *N Engl J Med* Vol. 361 1058-66 (2009).

7. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).

8. Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **38**, D5-16 (2010).

9. Consortium, I.H. A haplotype map of the human genome. in *Nature* Vol. 437 1299-320 (2005).

10. Consortium, I.H. et al. A second generation human haplotype map of over 3.1 million SNPs. in *Nature* Vol. 449 851-61 (2007).

11. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. in *Nucleic Acids Research* Vol. 33 D514-7 (2005).

12. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics (Oxford, England)* **14**, 656-64 (1998).

13. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

14. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research* **35**, W193-200 (2007).

15. Forbes, S.A. et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Research* **38**, D652-7 (2010).

16. Mitterbauer-Hohendanner, G. & Mannhalter, C. The biological and clinical significance of MLL abnormalities in haematological malignancies. *Eur J Clin Invest* **34 Suppl 2**, 12-24 (2004).

17. Meyer, C. et al. New insights to the MLL recombinome of acute leukemias. *Leukemia* **23**, 1490-9 (2009).

18. Hubmann, R. et al. NOTCH2 links protein kinase C delta to the expression of CD23 in chronic lymphocytic leukaemia (CLL) cells. *Br J Haematol* (2009).

19. Al Sarakbi, W. et al. The mRNA expression of SETD2 in human breast cancer: correlation with clinico-pathological parameters. *BMC Cancer* **9**, 290 (2009).

20. Dalgliesh, G.L. et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-3 (2010).

21. Durkin, M.E., Ullmannova, V., Guan, M. & Popescu, N.C. Deleted in liver cancer 3 (DLC-3), a novel Rho GTPase-activating protein, is downregulated in cancer and inhibits tumor cell growth. *Oncogene* **26**, 4580-9 (2007).

22. Mokarram, P. et al. Distinct high-profile methylated genes in colorectal cancer. *PLoS ONE* **4**, e7012 (2009).

23. Kitazoe, M., Futami, J., Nishikawa, M., Yamada, H. & Maeda, Y. Polyethylenimine-cationized beta-catenin protein transduction activates the Wnt

canonical signaling pathway more effectively than cationic lipid-based transduction. *Biotechnology journal* (2010).

24. Cheng, X.Y. & O'Neill, H.C. Oncogenesis and cancer stem cells: current opinions and future directions. *J Cell Mol Med* (2009).

25. Wang, C.-Y. et al. YY1AP, a novel co-activator of YY1. *J Biol Chem* **279**, 17750-5 (2004).

26. Grubach, L. et al. Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia. *Eur J Haematol* **81**, 112-22 (2008).

27. McManus, M.T. MicroRNAs and cancer. *Semin Cancer Biol* **13**, 253-8 (2003).

28. Audic, Y. & Hartley, R.S. Post-transcriptional regulation in cancer. *Biol Cell* **96**, 479-98 (2004).

29. Liu, G. et al. AIM-2: a novel tumor antigen is expressed and presented by human glioma cells. *J Immunother* **27**, 220-6 (2004).

30. Willett, C.G. et al. Macrophage-stimulating protein and its receptor in non-small-cell lung tumors: induction of receptor tyrosine phosphorylation and cell migration. *Am J Respir Cell Mol Biol* **18**, 489-96 (1998).

31. Welm, A.L. et al. The macrophage-stimulating protein pathway promotes metastasis in a mouse model for breast cancer and predicts poor prognosis in humans. *Proc Natl Acad Sci USA* **104**, 7570-5 (2007).

32. Bloch, D.B., de la Monte, S.M., Guigaouri, P., Filippov, A. & Bloch, K.D. Identification and characterization of a leukocyte-specific component of the nuclear body. *J Biol Chem* **271**, 29198-204 (1996).

33. Graux, C. et al. Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute lymphoblastic leukemia. *Nat Genet* **36**, 1084-9 (2004).

34. Krona, C. et al. A novel 1p36.2 located gene, APITD1, with tumour-suppressive properties and a putative p53-binding domain, shows low expression in neuroblastoma tumours. *Br J Cancer* **91**, 1119-30 (2004).

35. Hagemeijer, A. & Graux, C. ABL1 rearrangements in T-cell acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **49**, 299-308 (2010).

36. Bera, T.K. et al. POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer. *Proc Natl Acad Sci USA* **99**, 16975-80 (2002).

37. Moniaux, N., Junker, W.M., Singh, A.P., Jones, A.M. & Batra, S.K. Characterization of human mucin MUC17. Complete coding sequence and organization. *J Biol Chem* **281**, 23676-85 (2006).

38. Pleasance, E.D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-90 (2010).

39. Lee, J.-H., Lee, E.S., Kim, Y.-S., Won, N.H. & Chae, Y.-S. BRAF mutation and AKAP9 expression in sporadic papillary thyroid carcinomas. *Pathology* **38**, 201-4 (2006).

40. Frank, B. et al. Association of a common AKAP9 variant with breast cancer risk: a collaborative analysis. *J Natl Cancer Inst* **100**, 437-42 (2008).

41. Truong, T. et al. International Lung Cancer Consortium: Coordinated association study of 10 potential lung cancer susceptibility variants. *Carcinogenesis* (2010).

42. Hibi, M. et al. Molecular cloning and expression of an IL-6 signal transducer, gp130. *Cell* **63**, 1149-57 (1990).

43. Bravo, J., Staunton, D., Heath, J.K. & Jones, E.Y. Crystal structure of a cytokine-binding region of gp130. *EMBO J* **17**, 1665-74 (1998).

44. Hirano, T., Nakajima, K. & Hibi, M. Signaling mechanisms through gp130: a model of the cytokine system. *Cytokine Growth Factor Rev* **8**, 241-52 (1997).

45. Hirano, T., Matsuda, T. & Nakajima, K. Signal transduction through gp130 that is shared among the receptors for the interleukin 6 related cytokine subfamily. *Stem Cells* **12**, 262-77 (1994).

46. Taga, T. Gp130, a shared signal transducing receptor component for hematopoietic and neuropoietic cytokines. *J Neurochem* **67**, 1-10 (1996).

47. Boulanger, M.J. & Garcia, K.C. Shared cytokine signaling receptors: structural insights from the gp130 system. *Adv Protein Chem* **68**, 107-46 (2004).

48. Knüpfer, H. & Preiss, R. Significance of interleukin-6 (IL-6) in breast cancer (review). *Breast Cancer Res Treat* **102**, 129-35 (2007).

49. Heikkilä, K., Ebrahim, S. & Lawlor, D.A. Systematic review of the association between circulating interleukin-6 (IL-6) and cancer. *Eur J Cancer* **44**, 937-45 (2008).

50. Bromberg, J. & Wang, T.C. Inflammation and cancer: IL-6 and STAT3 complete the link. *Cancer Cell* **15**, 79-80 (2009).

51. Grivennikov, S. et al. IL-6 and Stat3 are required for survival of intestinal epithelial cells and development of colitis-associated cancer. *Cancer Cell* **15**, 103-13 (2009).

52. Bollrath, J. et al. gp130-mediated Stat3 activation in enterocytes regulates cell survival and cell-cycle progression during colitis-associated tumorigenesis. *Cancer Cell* **15**, 91-102 (2009).

53. Zarling, J.M. et al. Oncostatin M: a growth regulator produced by differentiated histiocytic lymphoma cells. *Proc Natl Acad Sci USA* **83**, 9739-43 (1986).

54. Kim, H., Jo, C., Jang, B.G., Oh, U. & Jo, S.A. Oncostatin M induces growth arrest of skeletal muscle cells in G1 phase by regulating cyclin D1 protein level. *Cell Signal* **20**, 120-9 (2008).

55. Rose, T.M. et al. The genes for oncostatin M (OSM) and leukemia inhibitory factor (LIF) are tightly linked on human chromosome 22. *Genomics* **17**, 136-40 (1993).

56. Wånggren, K. et al. Leukaemia inhibitory factor receptor and gp130 in the human Fallopian tube and endometrium before and after mifepristone treatment and in the human preimplantation embryo. *Mol Hum Reprod* **13**, 391-7 (2007).

57. Kamohara, H., Ogawa, M., Ishiko, T., Sakamoto, K. & Baba, H. Leukemia inhibitory factor functions as a growth factor in pancreas carcinoma cells: Involvement of regulation of LIF and its receptor expression. *Int J Oncol* **30**, 977-83 (2007).

58. García-Tuñón, I. et al. OSM, LIF, its receptors, and its relationship with the malignance in human breast carcinoma (in situ and in infiltrative). *Cancer Invest* **26**, 222-9 (2008).

59. Anhuf, D. et al. Signal transduction of IL-6, leukemia-inhibitory factor, and oncostatin M: structural receptor requirements for signal attenuation. *J Immunol* **165**, 2535-43 (2000).

60. Mardis, E.R. et al. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *New England Journal of Medicine* **361**, 1058-1066 (2009).

61. Fox, E.J., Salk, J.J. & Loeb, L.A. Cancer genome sequencing--an interim analysis. in *Cancer Res* Vol. 69 4948-50 (2009).

62. Radtke, I. et al. Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci USA* **106**, 12944-9 (2009).

# Chapter V

# Concluding Remarks

Through efforts that began at least 25 centuries ago with the Greek philosopher Anaximander of Miletus, humans have been questioning and striving to comprehend their own origins: life contemplating life. This endeavor has proceeded with punctuated advancement over the millennia, but has seen a dramatic acceleration in the past 200 years. What was once described as a "living filament" in the writings of Erasmus Darwin is now precisely understood to be the sequence of nucleotides comprising the DNA molecules carried by all organisms. The theory of evolution in general and the field of population genetics in specific, has revolutionized the capability of humans to not only comprehend their origin as a species but the origin of the ways in which people can exhibit so many differences while retaining many fundamental similarities. This is observable between and within human populations of all scales, ranging from individual family units to entire nations or ethnicities. Even the relationships between populations of cells making up an individual human can be described in these terms.

Through equally astounding advancements in chemistry and physics, technology has reached a point that the process of decoding the blueprint for the physical manifestation of life has become as routine as a naturalist's expedition to collect biological samples was 150 years ago, arguably even more so. In large part this thesis has been concerned with obtaining biological information from the raw output of this technology. More specifically, the primary focus has been on discovering genetic variation through high-throughput sequencing of entire genomes of DNA. In Chapter I, the most recent developments in describing a personal genome via resequencing were reviewed. Many computational

challenges have arisen due to the methodology driving these advancements and the pace at which they are occurring. In Chapter II, a framework was introduced for discovering genetic variation and origins of *de novo* mutations by probabilistically inferring a pedigree using high-throughput sequencing, or Piphits. Through simulation studies Piphits was shown to reliably predict the locations of mutation events within the sequenced family members as well as produce direct estimates for the rates of mutation and sequencing error. Chapter III reported on the first application of this method within the context of an international effort to create an unprecedented catalog of human genetic variation, The 1,000 Genomes Project. Although the Piphits-EM algorithm was unable to produce reliable estimates for the human mutation rate or effective population size, the validation experiments preformed showed that Piphits was successful at locating many true *de novo* mutations. In addition, this project brought to light the potentially large amount of somatic variation contained within any individual human, emphasizing the idea that the human genome is not merely a single sequence of nucleic acids, but is a catalog of many variations surrounding a common theme. Finally, in Chapter IV Piphits was applied in an extended manor, attempting to discover medically significant somatic mutations within a resequenced leukemia sample when placed in context of a pedigree. Samples were taken from the patient's father and mother to distinguish inherited from *de novo* variation, and a blood sample was taken from the patient after treatment to distinguish somatic, leukemia-specific variation from acquired, germline mutations. In this study Piphits was shown to be unable to efficiently predict mutations from the particular data set in question when

compared with a method that did not jointly infer the genotypes across the pedigree. In this case making use of the full complement of data, of both high and poor quality, was shown to be more effective at predicting genotype than probabilistically modeling the sample of observed nucleotides. However, the possibility remains of producing a method that takes advantages of these ideas while at the same time making the use of relatedness between resequenced individuals to increase the power for determining genotype and inferring the presence and origin of *de novo* mutation.

Conscious awareness of genetic variation will play an increasing role in the lives of people in medically advanced parts of the world. Linking genetic variation to disease and disease susceptibility is already becoming a priority modern medicine. However doubt exists as to whether the methods are yet available allow patients and health-care providers to reliably make informed decisions when applying genotype information to the prognosis, diagnosis, or treatment of human illness. Although the technology is available to quickly reproduce the primary DNA sequence of a patient, the body of work in this text makes evident that the methods and procedures are not in place for consistently producing genetic information with the fidelity necessary for making significant decisions about a patient's health. The personal genome is a concept far removed from the mind of a Nineteenth century naturalist, but the full, prudent application of the variation contained within a single personal genome to human medicine may be just as removed from the current body of knowledge in human genetics. The current efforts underway to more completely describe the genetic variation contained within humans as a species may be the first steps in another

era of punctuated development in our collective understanding of what it means to be human.

# APPENDIX

**Table A-1.** Coding somatic mutations predicted by the Chapter IV SNP-method

| Coord. | Gene Symbol | Base Subs. | SNP Type | SIFT Pred. | Q |
|---|---|---|---|---|---|
| 2:130589320 | *A26C1B* | G/A | Missense | Del. | 59 |
| 4:171246170 | *AADAT* | A/G | Missense | Tol. | 50 |
| 2:215618894 | *ABCA12* | C/A | Missense | Del. | 34 |
| 1:94282842 | *ABCA4* | C/A | Missense | Del. | 74 |
| 2:203968725 | *ABI2* | C/A | Missense | Tol. | 54 |
| 16:19451270 | *AC012621.2* | A/G | Missense | Tol. | 94 |
| 4:48688324 | *AC020593.1* | G/A | Missense | Tol. | 32 |
| 16:28381818 | *AC138894.2-2* | C/T | Silent | None | 46 |
| 14:22619550 | *ACIN1* | T/C | Silent | None | 44 |
| 2:135375897 | *ACMSD* | A/G | Missense | Del. | 87 |
| 14:73111651 | *ACOT2* | G/A | Missense | Del. | 39 |
| 3:58495755 | *ACOX2* | A/G | Missense | Del. | 31 |
| 2:158152056 | *ACVR1C* | A/G | Missense | Del. | 48 |
| 5:7760890 | *ADCY2* | A/G | Missense | Del. | 43 |
| 16:3955817 | *ADCY9* | G/A | Missense | Tol. | 62 |
| 5:148660894 | *AFAP1L1* | G/A | Nonsense | Del. | 33 |
| 4:74529644 | *AFP* | G/A | Missense | Del. | 41 |
| 1:100102672 | *AGL* | G/A | Missense | Tol. | 62 |
| 11:62046597 | *AHNAK* | C/A | Missense | Del. | 36 |
| 6:151711605 | *AKAP12* | C/A | Missense | Tol. | 49 |
| 6:135302211 | *ALDH8A1* | A/G | Missense | Del. | 44 |
| 18:54354599 | *ALPK2* | A/G | Missense | Tol. | 31 |
| 12:45757705 | *AMIGO2* | C/A | Missense | None | 48 |
| 4:74163232 | *ANKRD17* | C/T | Missense | Tol. | 32 |
| 9:43114347 | *ANKRD20A1* | C/T | Silent | None | 76 |
| 9:67544715 | *ANKRD20A3* | T/A | Silent | None | 54 |
| 4:125812661 | *ANKRD50* | A/G | Silent | None | 31 |
| 14:19993997 | *APEX1* | C/T | Missense | Tol. | 35 |
| 22:34452654 | *APOL5* | C/A | Missense | Del. | 44 |
| 9:33375828 | *AQP7* | C/T | Missense | Tol. | 46 |
| 20:47044529 | *ARFGEF2* | G/A | Missense | Del. | 57 |
| 19:922878 | *ARID3A* | G/T | Missense | Tol. | 44 |
| 3:35745848 | *ARPP-21* | C/A | Silent | None | 45 |
| 2:9436547 | *ASAP2* | G/A | Missense | Del. | 32 |
| 6:101270064 | *ASCC3* | G/A | Missense | Del. | 79 |
| 1:153716897 | *ASH1L* | A/T | Missense | Tol. | 63 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 2:190243609 | *ASNSD1* | A/T | Silent | None | 99 |
| 1:175123833 | *ASTN1* | G/T | Silent | None | 37 |
| 2:175684574 | *ATF2* | A/G | Missense | Tol. | 85 |
| 3:194603323 | *ATP13A4* | G/A | Missense | Tol. | 73 |
| 3:143127167 | *ATP1B3* | A/T | Missense | Tol. | 51 |
| 16:28806360 | *ATP2A1* | C/A | Missense | Tol. | 34 |
| 12:109262876 | *ATP2A2* | C/A | Missense | Del. | 37 |
| 10:7882048 | *ATP5C1* | G/A | Silent | None | 35 |
| 17:19187331 | *B9D1* | A/G | Missense | Del. | 62 |
| 21:14363201 | *BC048201* | T/C | Missense | Del. | 32 |
| 11:118277051 | *BCL9L* | T/C | Missense | Del. | 63 |
| 4:104233263 | *BDH2* | T/C | Missense | Tol. | 39 |
| 10:88671362 | *BMPR1A* | C/T | Missense | Tol. | 42 |
| 1:149284974 | *BNIPL* | C/A | Missense | Tol. | 86 |
| 22:48567181 | *BRD1* | A/G | Missense | Tol. | 51 |
| 6:33056188 | *BRD2* | G/A | Missense | Del. | 38 |
| 5:137516098 | *BRD8* | T/A | Silent | None | 37 |
| 6:36306326 | *BRPF3* | C/A | Missense | Del. | 107 |
| 16:3580163 | *BTBD12* | C/T | Missense | Tol. | 33 |
| 10:124447621 | *C10orf120* | C/A | Silent | None | 55 |
| 12:86904223 | *C12orf50* | T/C | Missense | Del. | 50 |
| 12:86944450 | *C12orf50* | A/G | Nonsense | Del. | 54 |
| 1:181139953 | *C1orf14* | T/C | Silent | None | 38 |
| 1:245779028 | *C1orf150* | C/T | Missense | Tol. | 40 |
| 1:116468428 | *C1orf161* | C/A | Missense | Tol. | 43 |
| 1:116477280 | *C1orf161* | T/A | Nonsense | Del. | 53 |
| 1:154968888 | *C1orf66* | A/G | Missense | Tol. | 35 |
| 13:23793617 | *C1QTNF9* | C/T | Missense | Tol. | 125 |
| 20:26032134 | *C20orf191* | T/C | Missense | Tol. | 50 |
| 20:25997834 | *C20orf91* | A/T | Silent | None | 86 |
| 20:26011537 | *C20orf91* | C/A | Missense | Del. | 35 |
| 22:45022841 | *C22orf40* | G/A | Nonsense | Del. | 45 |
| 2:232166579 | *C2orf57* | G/C | Missense | Tol. | 39 |
| 2:218940481 | *C2orf62* | G/A | Missense | Tol. | 38 |
| 3:8642315 | *C3orf32* | C/A | Silent | None | 62 |
| 4:57534092 | *C4orf14* | A/G | Silent | None | 47 |
| 4:170895401 | *C4orf27* | G/T | Silent | None | 32 |
| 4:166097847 | *C4orf39* | C/T | Missense | Del. | 70 |
| 5:41197713 | *C6* | A/G | Missense | Del. | 39 |
| 6:99888028 | *C6orf168* | C/A | Silent | None | 31 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 8:86429302 | *CA1* | A/G | Missense | Tol. | 33 |
| 19:53834670 | *CA11* | T/C | Missense | Tol. | 37 |
| 19:51803661 | *CALM3* | A/G | Missense | Tol. | 131 |
| 5:40889576 | *CARD6* | A/G | Missense | Tol. | 38 |
| 2:113967997 | *CBWD2* | G/A | Missense | Tol. | 83 |
| 19:38064592 | *CCDC123* | G/A | Missense | Tol. | 62 |
| 21:29356398 | *CCT8* | T/C | Silent | None | 38 |
| 7:151773751 | *CCT8L1* | C/T | Missense | Del. | 152 |
| 1:206139168 | *CD34* | A/C | Missense | None | 49 |
| 2:173932273 | *CDCA7* | T/A | Silent | None | 47 |
| 3:45109870 | *CDCP1* | C/A | Missense | Del. | 77 |
| 8:95255241 | *CDH17* | A/G | Missense | Tol. | 75 |
| 1:178227969 | *CEP350* | G/A | Missense | Tol. | 84 |
| 10:50540752 | *CHAT* | G/A | Missense | Tol. | 36 |
| 8:42706593 | *CHRNB3* | A/G | Missense | Del. | 42 |
| 1:16254553 | *CLCNKB* | A/G | Missense | Tol. | 34 |
| 1:153507293 | *CLK2* | A/T | Silent | None | 36 |
| 7:23591599 | *CLK2P* | G/T | Nonsense | Del. | 32 |
| 8:87752352 | *CNGB3* | T/C | Silent | None | 36 |
| 9:43793529 | *CNTNAP3B* | G/A | Silent | None | 62 |
| 9:43848488 | *CNTNAP3B* | C/T | Nonsense | Del. | 85 |
| 2:189582170 | *COL3A1* | C/A | Missense | Tol. | 91 |
| 9:136814288 | *COL5A1* | C/T | Missense | Tol. | 60 |
| 2:189608070 | *COL5A2* | C/A | Missense | Tol. | 37 |
| 4:1378350 | *CRIPAK* | G/A | Silent | None | 96 |
| 11:111287491 | *CRYAB* | C/A | Missense | Tol. | 97 |
| 1:36712053 | *CSF3R* | G/A | Missense | Del. | 48 |
| 21:44018600 | *CSTB* | A/C | Silent | None | 35 |
| 7:117138928 | *CTTNBP2* | T/C | Silent | None | 44 |
| 5:127021277 | *CTXN3* | A/G | Silent | None | 31 |
| 19:15864351 | *CYP4F2* | C/T | Missense | Tol. | 89 |
| 12:31135932 | *DDX11* | C/T | Missense | Tol. | 55 |
| 9:134526430 | *DDX31* | C/G | Missense | Del. | 94 |
| 5:134181125 | *DDX46* | G/A | Silent | None | 93 |
| 12:12865591 | *DDX47* | C/A | Missense | None | 96 |
| 12:122659309 | *DDX55* | G/A | Missense | Tol. | 33 |
| 11:46347629 | *DGKZ* | C/T | Missense | Del. | 145 |
| 22:45448245 | *DIP* | G/A | Missense | Tol. | 200 |
| 6:170435986 | *DLL1* | C/T | Missense | Del. | 48 |
| 22:37265287 | *DMC1* | C/G | Missense | Del. | 75 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 3:57468506 | *DNAH12L* | A/G | Missense | Del. | 33 |
| 17:7642315 | *DNAH2* | C/A | Missense | Tol. | 32 |
| 16:20882762 | *DNAH3* | C/A | Silent | None | 31 |
| 5:13772178 | *DNAH5* | C/A | Missense | Del. | 42 |
| 11:73348248 | *DNAJB13* | G/A | Missense | Del. | 38 |
| 6:32046148 | *DOM3Z* | A/G | Missense | Tol. | 60 |
| 2:74599769 | *DQX1* | C/A | Missense | None | 39 |
| 2:71642552 | *DYSF* | G/A | Missense | Tol. | 37 |
| 3:19915283 | *EFHB* | C/A | Missense | Del. | 50 |
| 1:15625061 | *EFHD2* | A/G | Nonsense | Del. | 34 |
| 1:36126648 | *EIF2C1* | T/C | Silent | None | 71 |
| 1:22208884 | *ELA3A* | T/A | Missense | Tol. | 43 |
| 6:145998273 | *EPM2A* | C/A | Missense | Del. | 31 |
| 1:43081264 | *ERMAP* | A/C | Missense | Del. | 40 |
| 4:5851235 | *EVC* | G/A | Missense | Tol. | 40 |
| 1:11073669 | *EXOSC10* | G/T | Missense | Del. | 64 |
| 11:107886613 | *EXPH5* | T/C | Silent | None | 44 |
| 5:159591748 | *FABP6* | G/T | Missense | Del. | 34 |
| 5:175858551 | *FAF2* | A/G | Missense | Del. | 38 |
| 20:25715343 | *FAM182B* | T/A | Silent | None | 49 |
| 1:177300226 | *FAM20B* | A/T | Silent | None | 39 |
| 1:175516839 | *FAM5B* | G/A | Missense | Tol. | 31 |
| 5:150928385 | *FAT2* | T/C | Silent | None | 31 |
| 5:150928615 | *FAT2* | C/A | Missense | Tol. | 44 |
| 4:126556036 | *FAT4* | A/G | Missense | Del. | 36 |
| 5:147801789 | *FBXO38* | A/G | Silent | None | 37 |
| 1:157540546 | *FCER1A* | A/G | Silent | None | 47 |
| 5:176453320 | *FGFR4* | C/G | Missense | Tol. | 40 |
| 1:153227483 | *FLAD1* | G/A | Missense | Tol. | 36 |
| 1:153228718 | *FLAD1* | T/C | Missense | Del. | 44 |
| 1:150542737 | *FLG* | G/A | Missense | Del. | 99 |
| 1:150543761 | *FLG* | G/C | Silent | None | 32 |
| 2:215948648 | *FN1* | T/C | Silent | None | 44 |
| 6:109092174 | *FOXO3* | G/A | Silent | None | 47 |
| 4:191111249 | *FRG1* | A/T | Missense | Del. | 84 |
| 5:121216133 | *FTMT* | C/A | Silent | None | 39 |
| 6:37550355 | *FTSJD2* | G/A | Silent | None | 54 |
| 5:161046559 | *GABRA6* | G/A | Missense | Tol. | 38 |
| 4:46728487 | *GABRB1* | T/C | Missense | Tol. | 67 |
| 11:62157528 | *GANAB* | C/A | Missense | Del. | 43 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 1:152051196 | *GATAD2B* | A/G | Silent | None | 47 |
| 10:104126427 | *GBF1* | C/A | Missense | Del. | 44 |
| 5:154285817 | *GEMIN5* | C/A | Missense | Tol. | 37 |
| 17:70750817 | *GGA3* | A/G | Nonsense | Del. | 65 |
| 7:149805618 | *GIMAP8* | A/G | Silent | None | 83 |
| 12:56149635 | *GLI1* | C/A | Missense | None | 48 |
| 9:126692559 | *GOLGA1* | T/C | Missense | Del. | 80 |
| 15:70740717 | *GOLGA6B* | G/A | Missense | Del. | 138 |
| 15:72155321 | *GOLGA6C* | C/T | Missense | Del. | 56 |
| 6:117220327 | *GPRC6A* | T/C | Missense | Del. | 35 |
| 5:175957757 | *GPRIN1* | C/T | Silent | None | 36 |
| 11:105309884 | *GRIA4* | C/T | Missense | Del. | 56 |
| 4:2979348 | *GRK4* | G/T | Silent | None | 41 |
| 3:143018311 | *GRK7* | C/A | Silent | None | 62 |
| 2:144481263 | *GTDC1* | A/G | Missense | Del. | 32 |
| 20:30135507 | *HCK* | A/G | Silent | None | 32 |
| 15:26680796 | *HERC2P2* | C/T | Silent | None | 34 |
| 21:37053942 | *HLCS* | A/G | Missense | None | 44 |
| 1:184218084 | *HMCN1* | C/A | Missense | Del. | 72 |
| 10:100894137 | *HPSE2* | C/T | Missense | Del. | 58 |
| 6:122776482 | *HSF2* | T/C | Missense | Tol. | 35 |
| 4:128971304 | *HSPA4L* | A/G | Missense | None | 37 |
| 1:43690514 | *HYI* | C/T | Silent | None | 36 |
| 15:76248379 | *IDH3A* | C/T | Silent | None | 41 |
| 6:160417002 | *IGF2R* | A/G | Missense | Tol. | 58 |
| 6:160445888 | *IGF2R* | A/G | Missense | Tol. | 70 |
| 3:152643574 | *IGSF10* | T/C | Silent | None | 53 |
| 2:213622769 | *IKZF2* | A/G | Missense | Tol. | 32 |
| 1:24319968 | *IL22RA1* | C/A | Missense | Del. | 42 |
| 2:218737465 | *IL8RA* | G/T | Silent | None | 59 |
| 7:64283236 | *INTS4* | C/G | Silent | None | 119 |
| 15:20864717 | *KIAA0393* | T/C | Missense | Tol. | 47 |
| 16:27668541 | *KIAA0556* | C/T | Missense | Tol. | 130 |
| 3:115238220 | *KIAA1407* | A/G | Missense | Del. | 47 |
| 2:61198755 | *KIAA1841* | A/G | Missense | Del. | 39 |
| 5:137549549 | *KIF20A* | A/C | Missense | Tol. | 52 |
| 15:67520165 | *KIF23* | C/T | Silent | None | 47 |
| 20:30381643 | *KIF3B* | C/T | Missense | Tol. | 56 |
| 12:56249427 | *KIF5A* | C/G | Silent | None | 54 |
| 16:56363007 | *KIFC3* | T/C | Missense | Tol. | 38 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 19:60017144 | *KIR2DL4* | C/T | Missense | Tol. | 31 |
| 6:129875255 | *LAMA2* | C/A | Missense | Del. | 38 |
| 3:49134129 | *LAMB2* | C/T | Missense | Del. | 59 |
| 4:16509183 | *LDB2* | G/T | Missense | Del. | 69 |
| 1:53500420 | *LRP8* | G/A | Missense | Del. | 36 |
| 1:233954020 | *LYST* | C/A | Silent | None | 37 |
| 4:6663795 | *MAN2B2* | C/T | Missense | Del. | 78 |
| 6:136976011 | *MAP3K5* | C/T | Silent | None | 34 |
| 14:70276615 | *MAP3K9* | C/T | Missense | Tol. | 49 |
| 15:39897559 | *MAPKBP1* | C/T | Missense | Tol. | 41 |
| 2:8934751 | *MBOAT2* | A/G | Missense | Del. | 35 |
| 18:13875460 | *MC2R* | C/A | Missense | Tol. | 52 |
| 1:29395329 | *MECR* | C/T | Missense | Tol. | 41 |
| 1:17176277 | *MFAP2* | C/A | Missense | Tol. | 31 |
| 1:40205411 | *MFSD2* | C/A | Missense | Del. | 34 |
| 7:151576489 | *MLL3* | C/T | Missense | Del. | 49 |
| 6:168092005 | *MLLT4* | C/A | Silent | None | 47 |
| 10:17948677 | *MRC1L1* | G/A | Silent | None | 108 |
| 6:43133892 | *MRPL2* | G/A | Silent | None | 36 |
| 3:137353357 | *MSL2* | A/G | Silent | None | 45 |
| 7:100466016 | *MUC17* | C/T | Silent | None | 97 |
| 7:100467210 | *MUC17* | A/G | Missense | Del. | 154 |
| 7:100467986 | *MUC17* | C/A | Missense | Tol. | 73 |
| 3:196959757 | *MUC4* | C/A | Missense | Del. | 33 |
| 17:45950042 | *MYCBPAP* | C/A | Missense | Tol. | 39 |
| 8:1993265 | *MYOM2* | T/C | Silent | None | 44 |
| 20:62307515 | *MYT1* | G/A | Missense | Tol. | 32 |
| 1:1678609 | *NADK* | C/T | Silent | None | 37 |
| 1:199954368 | *NAV1* | C/A | Missense | Tol. | 36 |
| 1:39272734 | *NDUFS5* | C/A | Missense | Del. | 39 |
| 16:66749319 | *NFATC3* | G/T | Missense | Tol. | 61 |
| 11:129244756 | *NFRKB* | T/C | Silent | None | 228 |
| 1:115630300 | *NGF* | C/A | Missense | None | 88 |
| 3:175475884 | *NLGN1* | C/A | Missense | Tol. | 42 |
| 7:150327376 | *NOS3* | C/A | Missense | Del. | 40 |
| 1:120260690 | *NOTCH2* | G/A | Silent | None | 37 |
| 1:120374070 | *NOTCH2* | T/C | Silent | None | 81 |
| 1:151928330 | *NPR1* | C/A | Silent | None | 91 |
| 17:35506317 | *NR1D1* | A/G | Silent | None | 44 |
| 9:101646889 | *NR4A3* | A/G | Missense | Del. | 53 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 9:86739026 | *NTRK2* | A/C | Silent | None | 85 |
| 1:227680058 | *NUP133* | A/G | Missense | Del. | 39 |
| 11:7463750 | *OLFML1* | C/T | Nonsense | Del. | 98 |
| 11:7463758 | *OLFML1* | A/G | Missense | Del. | 58 |
| 11:6847723 | *OR10A2* | T/C | Missense | Tol. | 228 |
| 19:15766086 | *OR10H5* | C/T | Missense | Tol. | 52 |
| 19:15766179 | *OR10H5* | C/T | Missense | Del. | 52 |
| 1:246045256 | *OR14A16* | G/A | Missense | Del. | 63 |
| 1:246179626 | *OR2L8* | C/T | Missense | Del. | 46 |
| 1:246704103 | *OR2T3* | A/G | Nonsense | Del. | 176 |
| 1:246803853 | *OR2T34* | T/C | Missense | Del. | 70 |
| 1:246592296 | *OR2T4* | A/C | Silent | None | 71 |
| 19:62140 | *OR4F17* | C/G | Silent | None | 45 |
| 1:156936837 | *OR6K2* | G/A | Missense | None | 44 |
| 11:55814474 | *OR8H1* | G/A | Missense | None | 94 |
| 11:57703856 | *OR9Q1* | C/A | Missense | Del. | 56 |
| 1:52634587 | *ORC1L* | T/A | Silent | None | 34 |
| 4:146282153 | *OTUD4* | G/C | Missense | Del. | 50 |
| 1:111759115 | *OVGP1* | A/G | Missense | Del. | 51 |
| 1:39807899 | *PABPC4* | T/C | Missense | Del. | 36 |
| 19:44571481 | *PAF1* | G/A | Missense | Tol. | 38 |
| 19:19542631 | *PBX4* | G/A | Missense | Del. | 87 |
| 5:140454671 | *PCDHB2* | C/A | Silent | None | 86 |
| 5:140835904 | *PCDHGA12* | A/G | Silent | None | 76 |
| 17:34148223 | *PCGF2* | C/A | Missense | Tol. | 71 |
| 12:20698330 | *PDE3A* | T/A | Silent | None | 33 |
| 1:143565938 | *PDE4DIP* | T/C | Missense | Del. | 54 |
| 1:143592964 | *PDE4DIP* | G/T | Missense | Del. | 42 |
| 1:143727234 | *PDE4DIP* | G/T | Silent | None | 108 |
| 2:10842547 | *PDIA6* | G/A | Missense | Tol. | 35 |
| 3:181016362 | *PEX5L* | T/A | Missense | Tol. | 33 |
| 3:48562377 | *PFKFB4* | G/T | Missense | Tol. | 86 |
| 21:44549694 | *PFKL* | G/A | Silent | None | 69 |
| 1:120079501 | *PHGDH* | C/T | Missense | Tol. | 79 |
| 1:202683265 | *PIK3C2B* | G/A | Missense | Del. | 54 |
| 1:9706988 | *PIK3CD* | G/A | Silent | None | 40 |
| 1:199519485 | *PKP1* | C/T | Silent | None | 77 |
| 2:160512307 | *PLA2R1* | T/C | Missense | Del. | 50 |
| 2:106370477 | *PLGLA1* | A/G | Missense | Tol. | 66 |
| 3:147407207 | *PLSCR4* | A/G | Missense | Del. | 33 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 22:41328808 | *POLDIP3* | A/T | Missense | Del. | 33 |
| 6:30678974 | *PPP1R10* | T/C | Missense | Tol. | 35 |
| 1:13034134 | *PRAMEF23* | C/A | Nonsense | Del. | 38 |
| 1:13203324 | *PRAMEF3* | G/A | Missense | Del. | 32 |
| 1:13241215 | *PRAMEF5* | G/T | Silent | None | 63 |
| 6:105883524 | *PREP* | C/T | Missense | Tol. | 72 |
| 11:75741284 | *PRKRIR* | A/C | Silent | None | 34 |
| 6:3989094 | *PRPF4B* | C/A | Missense | Tol. | 44 |
| 4:119435622 | *PRSS12* | G/T | Missense | Tol. | 47 |
| 19:47951116 | *PSG8* | G/C | Missense | Tol. | 35 |
| 3:185500878 | *PSMD2* | T/C | Silent | None | 86 |
| 1:28350038 | *PTAFR* | G/A | Missense | Tol. | 84 |
| 1:200386172 | *PTPN7* | T/C | Missense | Tol. | 69 |
| 1:200393550 | *PTPN7* | G/A | Missense | Del. | 63 |
| 1:31238030 | *PUM1* | C/A | Missense | Del. | 51 |
| 1:178418021 | *QSOX1* | C/T | Missense | Tol. | 31 |
| 8:117939775 | *RAD21* | A/G | Missense | Del. | 32 |
| 11:36571047 | *RAG2* | C/A | Silent | None | 49 |
| 2:108748255 | *RANBP2* | T/C | Silent | None | 88 |
| 1:110686028 | *RBM15* | G/A | Missense | Tol. | 32 |
| 6:155168298 | *RBM16* | C/T | Silent | None | 93 |
| 12:112877360 | *RBM19* | G/A | Missense | Del. | 62 |
| 1:89221886 | *RBMXL1* | C/G | Missense | Tol. | 98 |
| 1:8347417 | *RERE* | C/T | Missense | Del. | 60 |
| 10:42933850 | *RET* | C/A | Silent | None | 77 |
| 5:158520938 | *RNF145* | C/A | Missense | Tol. | 79 |
| 1:33175131 | *RNF19B* | G/T | Missense | Tol. | 55 |
| 4:84014874 | *SEC31A* | C/T | Missense | Tol. | 56 |
| 3:124130057 | *SEMA5B* | T/C | Silent | None | 84 |
| 18:59721207 | *SERPINB2* | C/A | Missense | Del. | 52 |
| 3:47137269 | *SETD2* | C/A | Missense | Tol. | 54 |
| 16:57109819 | *SETD6* | A/T | Missense | Tol. | 43 |
| 21:31995323 | *SFRS15* | A/G | Missense | Tol. | 36 |
| 6:134533955 | *SGK1* | G/A | Missense | Del. | 44 |
| 6:134536321 | *SGK1* | T/A | Missense | Del. | 93 |
| 4:42097789 | *SHISA3* | C/A | Missense | Del. | 36 |
| 3:166182822 | *SI* | T/C | Missense | Del. | 54 |
| 6:32038238 | *SKIV2L* | T/C | Silent | None | 43 |
| 5:127538193 | *SLC12A2* | A/G | Missense | Tol. | 55 |
| 2:230622730 | *SLC16A14* | C/A | Silent | None | 45 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 9:4562380 | *SLC1A1* | C/A | Silent | None | 32 |
| 2:172350220 | *SLC25A12* | C/A | Missense | Del. | 47 |
| 3:142177860 | *SLC25A36* | C/T | Missense | Del. | 38 |
| 5:149340657 | *SLC26A2* | T/C | Missense | Tol. | 42 |
| 6:88275480 | *SLC35A1* | C/A | Silent | None | 37 |
| 6:118694966 | *SLC35F1* | G/A | Silent | None | 37 |
| 9:107187544 | *SLC44A1* | G/A | Silent | None | 31 |
| 1:95102909 | *SLC44A3* | C/T | Missense | Tol. | 45 |
| 1:44249137 | *SLC6A9* | C/A | Missense | Tol. | 41 |
| 2:40510531 | *SLC8A1* | C/A | Missense | Del. | 35 |
| 2:102691042 | *SLC9A2* | G/A | Silent | None | 33 |
| 5:168071901 | *SLIT3* | T/C | Silent | None | 32 |
| 5:168113572 | *SLIT3* | C/A | Nonsense | Del. | 69 |
| 14:91012010 | *SMEK1* | C/A | Silent | None | 47 |
| 4:90968358 | *SNCA* | C/A | Silent | None | 58 |
| 18:17457849 | *SNRPD1* | A/T | Missense | Del. | 39 |
| 2:70368776 | *SNRPG* | G/A | Missense | Del. | 37 |
| 1:16598524 | *SPATA21* | C/A | Missense | Tol. | 60 |
| 4:124068241 | *SPATA5* | G/T | Missense | Tol. | 59 |
| 17:4336292 | *SPNS3* | C/T | Missense | Del. | 80 |
| 19:45699926 | *SPTBN4* | C/T | Silent | None | 44 |
| 1:24849113 | *SRRM1* | A/G | Missense | Del. | 72 |
| 3:188243791 | *ST6GAL1* | C/T | Missense | Del. | 37 |
| 3:52531395 | *STAB1* | C/T | Missense | Tol. | 42 |
| X:67860890 | *STARD8* | G/A | Missense | Tol. | 48 |
| 15:41687445 | *STRC* | C/T | Missense | Tol. | 100 |
| 11:59316167 | *STX3* | T/C | Missense | Tol. | 48 |
| 13:52152215 | *SUGT1* | C/T | Silent | None | 38 |
| 1:115257200 | *SYCP1* | C/T | Silent | None | 43 |
| 20:57903986 | *SYCP2* | T/C | Missense | Tol. | 32 |
| 6:152694541 | *SYNE1* | A/G | Missense | Del. | 42 |
| 6:152717509 | *SYNE1* | T/C | Missense | Tol. | 36 |
| 6:33516673 | *SYNGAP1* | C/A | Silent | None | 32 |
| 5:68696997 | *TAF9* | C/A | Silent | None | 60 |
| 6:159377862 | *TAGAP* | T/C | Missense | Tol. | 54 |
| 17:25911777 | *TBC1D29* | C/T | Missense | Del. | 38 |
| 5:179247887 | *TBC1D9B* | C/A | Missense | None | 33 |
| 1:119229579 | *TBX15* | A/C | Missense | Del. | 45 |
| 12:102897967 | *TDG* | T/A | Missense | Del. | 75 |
| 18:3447624 | *TGIF* | C/A | Missense | Del. | 53 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 6:155603516 | *TIAM2* | C/A | Silent | None | 54 |
| 5:114984168 | *TICAM2* | A/G | Missense | Del. | 42 |
| 12:55101522 | *TIMELESS* | C/G | Missense | Tol. | 45 |
| 20:2521019 | *TMC2* | A/G | Silent | None | 65 |
| 2:96283568 | *TMEM127* | G/A | Missense | Del. | 33 |
| 19:52241734 | *TMEM160* | G/A | Missense | Tol. | 51 |
| 1:222056456 | *TP53BP2* | C/T | Silent | None | 85 |
| 8:110168958 | *TRHR* | T/C | Missense | Del. | 59 |
| 4:154436089 | *TRIM2* | C/A | Silent | None | 43 |
| 1:117462798 | *TRIM45* | C/A | Silent | None | 81 |
| X:53129181 | *TSPYL2* | T/C | Missense | Tol. | 104 |
| 6:43328517 | *TTBK1* | G/C | Silent | None | 50 |
| 17:44223883 | *TTLL6* | G/C | Missense | Del. | 40 |
| 3:133876833 | *UBA5* | G/A | Silent | None | 48 |
| 1:200568786 | *UBE2T* | T/C | Silent | None | 38 |
| 1:19355992 | *UBR4* | C/A | Silent | None | 36 |
| 2:234286881 | *UGT1A8* | G/A | Silent | None | 80 |
| 19:18837102 | *UPF1* | G/A | Missense | Tol. | 38 |
| 1:159278177 | *USF1* | G/A | Silent | None | 34 |
| 6:41882432 | *USP49* | C/T | Silent | None | 33 |
| 15:48572282 | *USP8* | T/C | Silent | None | 99 |
| 5:72900111 | *UTP15* | C/A | Missense | Del. | 43 |
| 6:144851519 | *UTRN* | T/C | Silent | None | 31 |
| 6:30991757 | *VARSL* | A/T | Silent | None | 44 |
| 8:100637989 | *VPS13B* | A/G | Missense | Tol. | 52 |
| 1:12259714 | *VPS13D* | A/G | Missense | Del. | 34 |
| 1:12324505 | *VPS13D* | C/T | Missense | Del. | 39 |
| 6:33343948 | *VPS52* | G/T | Missense | Tol. | 75 |
| 14:96392281 | *VRK1* | C/A | Silent | None | 32 |
| 20:43186193 | *WFDC12* | C/A | Missense | Tol. | 31 |
| 1:22328852 | *WNT4* | G/A | Silent | None | 34 |
| X:52860911 | *XAGE5* | T/C | Missense | Del. | 62 |
| 1:153905101 | *YY1AP1* | G/A | Missense | Tol. | 44 |
| 3:102867051 | *ZBTB11* | T/C | Silent | None | 43 |
| X:119271349 | *ZBTB33* | T/G | Silent | None | 39 |
| 1:22700572 | *ZBTB40* | A/T | Missense | None | 50 |
| 1:202083300 | *ZC3H11A* | C/A | Silent | None | 74 |
| 10:31648179 | *ZEB1* | G/A | Silent | None | 44 |
| 3:15090568 | *ZFYVE20* | C/A | Missense | Del. | 34 |
| 14:67314623 | *ZFYVE26* | T/C | Missense | Del. | 102 |

**Table A-1**. Continued

| | | | | | |
|---|---|---|---|---|---|
| 16:25174187 | *ZKSCAN2* | G/A | Missense | Del. | 104 |
| 10:278061 | *ZMYND11* | T/C | Missense | Del. | 32 |
| 16:88327305 | *ZNF276* | A/C | Missense | Del. | 85 |
| 19:57160296 | *ZNF350* | C/A | Missense | Del. | 94 |
| 14:73440425 | *ZNF410* | T/C | Missense | Del. | 50 |
| 19:62776817 | *ZNF416* | A/G | Silent | None | 34 |
| 4:4355449 | *ZNF509* | T/C | Silent | None | 53 |
| 7:63804281 | *ZNF588* | C/T | Silent | None | 46 |
| 12:123063202 | *ZNF664* | C/G | Missense | None | 41 |
| 1:149527164 | *ZNF687* | C/A | Nonsense | Del. | 44 |
| 19:49583501 | *ZNF806* | A/G | Missense | None | 55 |

**Table A-2.** Coding somatic mutations predicted by Piphits-CGC in Chapter IV

| Coord. | Gene Symbol | Base Subs. | SNP Type | SIFT Pred. | $\delta$ |
|---|---|---|---|---|---|
| 15:89147848 | BLM | T/C | Missense | Del. | 0.272562456 |
| 15:89135046 | BLM | A/G | Missense | Tol. | 0.022893994 |
| 10:88649829 | BMPR1A | G/A | Missense | Tol. | 0.272562456 |
| 8:42917689 | HOOK3 | T/G | Missense | Del. | 0.011581061 |
| 5:55283865 | IL6ST | C/A | Missense | None | 0.866228165 |
| 6:168092005 | MLLT4 | C/A | Missense | Del. | 0.999997238 |
| 2:47876566 | MSH6 | G/T | Missense | Del. | 0.011581106 |
| 12:55392842 | NACA | T/C | Missense | Del. | 0.856976345 |
| 14:50289081 | NIN | A/G | Missense | Tol. | 0.044757531 |
| 1:120260690 | NOTCH2 | G/A | Missense | Del. | 0.409460291 |
| 9:101646889 | NR4A3 | A/G | Silent | None | 0.999918397 |
| 3:47137269 | SETD2 | C/A | Missense | Del. | 0.749769779 |
| 3:47137229 | SETD2 | G/T | Missense | Del. | 0.272562456 |
| 1:36529651 | THRAP3 | A/T | Missense | Del. | 0.996961247 |

**Table A-3.** Coding, somatic mutations predicted by Piphits-EW in Chapter IV

| Coord. | Gene Symbol | Base Subs. | SNP Type | SIFT Pred. | Pr.Mutation |
|---|---|---|---|---|---|
| 11:6847723 | *E2F8* | T/C | Silent | None | 1 |
| 7:91479864 | AKAP9 | A/G | Silent | None | 0.927 |
| 7:100465092 | MUC17 | T/G | Silent | None | 0.926 |
| 19:15864351 | CYP42 | C/T | Missense | Tol. | 0.373 |

**Table A-4.** Glossary of Genes Referenced

| Gene | Description |
|---|---|
| ***A26C1B*** | ANKRD26-like family C member 1B (Chimeric POTE-actin protein) A5A3E0 |
| ***AADAT*** | Kynurenine/alpha-aminoadipate aminotransferase mitochondrial Precursor (KAT/AadAT)(EC 2.6.1.7)(Kynurenine aminotransferase II)(Kynurenine--oxoglutarate aminotransferase II)(Kynurenine--oxoglutarate transaminase II)(2-aminoadipate transaminase)(EC 2.6.1.39)(2-aminoadipate aminotransferase)(Alpha-aminoadipate aminotransferase)(AadAT) Q8N5Z0 |
| ***ABCA12*** | ATP-binding cassette sub-family A member 12 (ATP-binding cassette transporter 12)(ATP-binding cassette 12) Q86UK0 |
| ***ABCA4*** | Retinal-specific ATP-binding cassette transporter (ATP-binding cassette sub-family A member 4)(RIM ABC transporter)(RIM protein)(RmP)(Stargardt disease protein) P78363 |
| ***ABI2*** | Abl interactor 2 (Abelson interactor 2)(Abi-2)(Abl-binding protein 3)(AblBP3)(Arg-binding protein 1)(ArgBP1) Q9NYB9 |
| ***AC012621.2*** | Centrosomal protein of 110 kDa (Cep110) O43303 |
| ***AC020593.1*** | PGAP2-interacting protein Q9H720 |
| ***AC104564.11-1*** | Abhydrolase domain-containing protein UNQ6510/PRO21435 Precursor (EC 3.1.1.-) Q6UXT9 |
| ***AC133561.4*** | Transporter Fragment :UniProtKB/TrEMBL;Acc:A6NF70 |
| ***AC138894.2-2*** | NPIP-like protein LOC440350 Precursor O75200 |
| ***ACIN1*** | Apoptotic chromatin condensation inducer in the nucleus (Acinus) Q9UKV3 |
| ***ACMSD*** | 2-amino-3-carboxymuconate-6-semialdehyde decarboxylase (EC 4.1.1.45) Q8TDX5 |
| ***ACOT2*** | Acyl-coenzyme A thioesterase 2, mitochondrial Precursor (Acyl-CoA thioesterase 2)(EC 3.1.2.2)(Acyl-coenzyme A thioester hydrolase 2a)(Long-chain acyl-CoA thioesterase 2)(ZAP128)(CTE-Ia) P49753 |
| ***ACOX2*** | Peroxisomal acyl-coenzyme A oxidase 2 (EC 1.17.99.3)(3-alpha,7-alpha,12-alpha-trihydroxy-5-beta-cholestanoyl-CoA 24-hydroxylase)(3-alpha,7-alpha,12-alpha-trihydroxy-5-beta-cholestanoyl-CoA oxidase)(Trihydroxycoprostanoyl-CoA oxidase)(THCA-CoA oxidase)(THCCox) Q99424 |

**Table A-4.** Continued

| | |
|---|---|
| ***ACVR1C*** | Activin receptor type-1C Precursor (EC 2.7.11.30)(ACTR-IC)(Activin receptor-like kinase 7)(ALK-7) Q8NER5 |
| ***ADAMTS6*** | A disintegrin and metalloproteinase with thrombospondin motifs 6 Precursor (ADAMTS-6)(ADAM-TS 6)(ADAM-TS6)(EC 3.4.24.-) Q9UKP5 |
| ***ADCY2*** | adenylate cyclase 2 (brain) |
| ***ADCY9*** | Adenylate cyclase type 9 (EC 4.6.1.1)(Adenylate cyclase type IX)(ATP pyrophosphate-lyase 9)(Adenylyl cyclase 9) O60503 |
| ***AFAP1L1*** | Actin filament-associated protein 1-like 1 Q8TED9 |

| | |
|---|---|
| ***AFP*** | Alpha-fetoprotein Precursor (Alpha-1-fetoprotein)(Alpha-fetoglobulin) P02771 |
| ***AGL*** | Glycogen debranching enzyme (Glycogen debrancher) Includes 4-alpha-glucanotransferase(EC 2.4.1.25)(Oligo-1,4-1,4-glucantransferase);Amylo-alpha-1,6-glucosidase(Amylo-1,6-glucosidase)(EC 3.2.1.33)(Dextrin 6-alpha-D-glucosidase) P35573 |
| ***AHNAK*** | Neuroblast differentiation-associated protein AHNAK (Desmoyokin) Q09666 |
| ***AIM2*** | Interferon-inducible protein AIM2 (Absent in melanoma 2) O14862 |
| ***AKAP12*** | A-kinase anchor protein 12 (A-kinase anchor protein 250 kDa)(AKAP 250)(Gravin)(Myasthenia gravis autoantigen) Q02952 |
| ***AKAP9*** | A-kinase anchor protein 9 (Protein kinase A-anchoring protein 9)(PRKA9)(A-kinase anchor protein 450 kDa)(AKAP 450)(A-kinase anchor protein 350 kDa)(AKAP 350)(hgAKAP 350)(AKAP 120-like protein)(Protein hyperion)(Protein yotiao)(Centrosome- and Golgi-localized PKN-associated protein)(CG-NAP) |
| ***ALDH8A1*** | Aldehyde dehydrogenase family 8 member A1 (EC 1.2.1.-)(Aldehyde dehydrogenase 12) Q9H2A2 |
| ***ALKBH1*** | Alkylated DNA repair protein alkB homolog 1 Q13686 |
| ***ALPK2*** | Alpha-protein kinase 2 (EC 2.7.11.-)(Heart alpha-protein kinase) Q86TB3 |
| ***AMIGO2*** | Amphoterin-induced protein 2 Precursor (AMIGO-2)(Alivin-1)(Differentially expressed in gastric adenocarcinomas)(DEGA) Q86SJ2 |
| ***ANKRD17*** | ankyrin repeat domain 17 |
| ***ANKRD20A1*** | Ankyrin repeat domain-containing protein 20A3 Q5VUR7 |
| ***ANKRD20A3*** | Ankyrin repeat domain-containing protein 20A1 Q5TYW2 |
| ***ANKRD50*** | Ankyrin repeat domain-containing protein 50 Q9ULJ7 |
| ***APEX1*** | DNA-(apurinic or apyrimidinic site) lyase (EC 4.2.99.18)(Apurinic-apyrimidinic endonuclease 1)(AP endonuclease 1)(APEX nuclease)(APEN)(Protein REF-1) P27695 |
| ***APITD1*** | Centromere protein S (CENP-S)(Apoptosis-inducing TAF9-like domain-containing protein 1) Q8N2Z9 |
| ***APOL5*** | Apolipoprotein L5 (Apolipoprotein L-V)(ApoL-V) Q9BWW9 |

**Table A-4.** Continued

| | |
|---|---|
| ***AQP7*** | Aquaporin-7 (AQP-7)(Aquaporin-7-like)(Aquaporin adipose)(AQPap) O14520 |
| ***ARFGEF2*** | Brefeldin A-inhibited guanine nucleotide-exchange protein 2 (Brefeldin A-inhibited GEP 2) Q9Y6D5 |
| ***ARID3A*** | AT-rich interactive domain-containing protein 3A (ARID domain-containing protein 3A)(Dead ringer-like protein 1)(B-cell regulator of IgH transcription)(Bright)(E2F-binding protein 1) Q99856 |

| | |
|---|---|
| ***ARPP-21*** | cAMP-regulated phosphoprotein 21 (ARPP-21)(Thymocyte cAMP-regulated phosphoprotein) Q9UBL0 |
| ***ASAP2*** | Arf-GAP with SH3 domain, ANK repeat and PH domain-containing protein 2 (Development and differentiation-enhancing factor 2)(Pyk2 C-terminus-associated protein)(PAP)(Paxillin-associated protein with ARFGAP activity 3)(PAG3) O43150 |
| ***ASCC3*** | Activating signal cointegrator 1 complex subunit 3 (EC 3.6.1.-)(ASC-1 complex subunit p200)(Trip4 complex subunit p200)(Helicase, ATP binding 1) Q8N3C0 |
| ***ASH1L*** | Probable histone-lysine N-methyltransferase ASH1L (EC 2.1.1.43)(Absent small and homeotic disks protein 1 homolog)(ASH1-like protein)(huASH1)(Lysine N-methyltransferase 2H) Q9NR48 |
| ***ASNSD1*** | Asparagine synthetase domain-containing protein 1 (HCV NS3-transactivated protein 1) Q9NWL6 |
| ***ASTN1*** | Astrotactin-1 Precursor O14525 |
| ***ATF2*** | activating transcription factor 2 |
| ***ATP13A4*** | Probable cation-transporting ATPase 13A4 (EC 3.6.3.-)(P5-ATPase isoform 4) Q4VNC1 |
| ***ATP1B3*** | Sodium/potassium-transporting ATPase subunit beta-3 (Sodium/potassium-dependent ATPase subunit beta-3)(ATPB-3)(CD298 antigen) P54709 |
| ***ATP2A1*** | Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 (SERCA1)(EC 3.6.3.8)(Calcium pump 1)(Calcium-transporting ATPase sarcoplasmic reticulum type, fast twitch skeletal muscle isoform)(SR Ca(2+)-ATPase 1)(Endoplasmic reticulum class 1/2 Ca(2+) ATPase) O14983 |
| ***ATP2A2*** | Sarcoplasmic/endoplasmic reticulum calcium ATPase 2 (SERCA2)(EC 3.6.3.8)(Calcium pump 2)(Calcium-transporting ATPase sarcoplasmic reticulum type, slow twitch skeletal muscle isoform)(SR Ca(2+)-ATPase 2)(Endoplasmic reticulum class 1/2 Ca(2+) ATPase) P16615 |
| ***ATP5C1*** | ATP synthase subunit gamma, mitochondrial Precursor (F-ATPase gamma subunit) P36542 |
| ***B9D1*** | B9 domain-containing protein 1 Q9UPM9 |

**Table A-4.** Continued

| | |
|---|---|
| ***BCL9L*** | B-cell CLL/lymphoma 9-like protein (B-cell lymphoma 9-like protein)(BCL9-like protein)(Protein BCL9-2) Q86UU0 |
| ***BDH2*** | 3-hydroxybutyrate dehydrogenase type 2 (EC 1.1.1.30)(R-beta-hydroxybutyrate dehydrogenase)(Dehydrogenase/reductase SDR family member 6)(Oxidoreductase UCPA) Q9BUT1 |

| | |
|---|---|
| ***BLM*** | Bloom syndrome protein (EC 3.6.1.-)(RecQ protein-like 3)(DNA helicase, RecQ-like type 2) :UniProtKBSwiss-Prot;Acc:P54132 ENSFM00500000269713 BLOOM SYNDROME Bloom Syndrome |
| ***BMPR1A*** | Bone morphogenetic protein receptor type-1A Precursor (EC 2.7.11.30)(Serine/threonine-protein kinase receptor R5)(SKR5)(Activin receptor-like kinase 3)(ALK-3)(CD292 antigen) P36894 |
| ***BNIPL*** | Bcl-2/adenovirus E1B 19 kDa-interacting protein 2-like protein Q7Z465 |
| ***BRD1*** | Bromodomain-containing protein 1 (BR140-like protein) O95696 |
| ***BRD2*** | Bromodomain-containing protein 2 (Protein RING3)(O27.1.1) P25440 |
| ***BRD8*** | Bromodomain-containing protein 8 (p120)(Skeletal muscle abundant protein 2)(Skeletal muscle abundant protein)(Thyroid hormone receptor coactivating protein 120kDa)(TrCP120) Q9H0E9 |
| ***BRPF3*** | bromodomain and PHD finger containing, 3 |
| ***BTBD12*** | BTB/POZ domain-containing protein 12 Q8IY92 |
| ***BTN3A2*** | Butyrophilin subfamily 3 member A2 Precursor P78410 |
| ***C10orf120*** | Uncharacterized protein C10orf120 Q5SQS8 |
| ***C12orf50*** | Uncharacterized protein C12orf50 Q8NA57 |
| ***C12orf50*** | Uncharacterized protein C12orf50 Q8NA57 |
| ***C1orf14*** | Uncharacterized protein C1orf14 Q9BZQ2 |
| ***C1orf150*** | Putative uncharacterized protein C1orf150 Q5JQS6 |
| ***C1orf161*** | Uncharacterized protein C1orf161 Q8N8X9 |
| ***C1orf161*** | Uncharacterized protein C1orf161 Q8N8X9 |
| ***C1orf66*** | UPF0431 protein C1orf66 Q96FB5 |
| ***C1QTNF9*** | Complement C1q tumor necrosis factor-related protein 9 Precursor P0C862 |
| ***C20orf191*** | Putative nuclear receptor corepressor 1-like protein C20orf191 Q9H4R4 |
| ***C20orf46*** | Transmembrane protein C20orf46 Q9NUR3 |
| ***C20orf91*** | Protein FAM182A Q5T1J6 |
| ***C22orf40*** | UPF0595 protein C22orf40 Q6NVV7 |
| ***C2orf57*** | Uncharacterized protein C2orf57 Q53QW1 |
| ***C2orf62*** | Uncharacterized protein C2orf62 Q7Z7H3 |
| ***C3orf32*** | Uncharacterized protein C3orf32 Q9Y2M2 |

**Table A-4.** Continued

| | |
|---|---|
| ***C4orf14*** | Uncharacterized protein C4orf14 Q8NC60 |
| ***C4orf27*** | UPF0609 protein C4orf27 Q9NWY4 |
| ***C4orf39*** | chromosome 4 open reading frame 39 |
| ***C6*** | Complement component C6 Precursor P13671 |
| ***C6orf168*** | Uncharacterized protein C6orf168 Q5TGI0 |

| | |
|---|---|
| ***CA1*** | Carbonic anhydrase 1 (EC 4.2.1.1)(Carbonic anhydrase I)(CA-I)(Carbonate dehydratase I)(Carbonic anhydrase B)(CAB) P00915 |
| ***CA11*** | Carbonic anhydrase-related protein 11 Precursor (CA-RP XI)(CARP XI)(CA-XI)(Carbonic anhydrase-related protein 2)(CARP-2)(CA-RP II) O75493 |
| ***CALM3*** | Calmodulin (CaM) P62158 |
| ***CARD6*** | Caspase recruitment domain-containing protein 6 Q9BX69 |
| ***CBWD2*** | COBW domain-containing protein 2 (Cobalamin synthetase W domain-containing protein 2) Q8IUF1 |
| ***CBWD5*** | COBW domain-containing protein 5 (Cobalamin synthetase W domain-containing protein 5) Q5RIA9 |
| ***CCDC123*** | coiled-coil domain containing 123 |
| ***CCT8*** | chaperonin containing TCP1, subunit 8 (theta) |
| ***CCT8L1*** | Putative T-complex protein 1 subunit theta-like 1 A6NM43 |
| ***CD34*** | Hematopoietic progenitor cell antigen CD34 Precursor (CD34 antigen) P28906 |
| ***CDC14*** | CDC14 cell division cycle 14 homolog A |
| ***CDCA7*** | Cell division cycle-associated protein 7 (Protein JPO1) Q9BWT1 |
| ***CDCP1*** | CUB domain-containing protein 1 Precursor (Transmembrane and associated with src kinases)(Membrane glycoprotein gp140)(Subtractive immunization M plus HEp3-associated 135 kDa protein)(SIMA135)(CD318 antigen) Q9H5V8 |
| ***CDH17*** | Cadherin-17 Precursor (Liver-intestine cadherin)(LI-cadherin)(Intestinal peptide-associated transporter HPT-1) Q12864 |
| ***CENPO*** | Centromere protein O (CENP-O)(Interphase centromere complex protein 36) Q9BU64 |
| ***CEP350*** | Centrosome-associated protein 350 (Cep350)(Centrosome-associated protein of 350 kDa) Q5VT06 |
| ***CHAT*** | Choline O-acetyltransferase (Choline acetylase)(CHOACTase)(ChAT)(EC 2.3.1.6) P28329 |
| ***CHRNB3*** | Neuronal acetylcholine receptor subunit beta-3 Precursor Q05901 |
| ***CLCNKB*** | Chloride channel protein ClC-Kb (Chloride channel Kb)(ClC-K2) P51801 |
| ***CLK2*** | Dual specificity protein kinase CLK2 (EC 2.7.12.1)(CDC-like kinase 2) P49760 |
| ***CLK2P*** | CDC-like kinase 2, pseudogene |

**Table A-4.** Continued

| | |
|---|---|
| ***CNGB3*** | Cyclic nucleotide-gated cation channel beta-3 (CNG channel beta-3)(Cyclic nucleotide-gated channel beta-3)(Cone photoreceptor cGMP-gated channel subunit beta)(Cyclic nucleotide-gated cation channel modulatory subunit) Q9NQW8 |
| ***CNTNAP3B*** | Contactin-associated protein-like 3B Precursor (Cell recognition molecule Caspr3b) Q96NU0 |

| | |
|---|---|
| ***COL3A1*** | Collagen alpha-1(III) chain Precursor P02461 |
| ***COL5A1*** | Collagen alpha-1(V) chain Precursor P20908 |
| ***COL5A2*** | Collagen alpha-2(V) chain Precursor P05997 |
| ***CRIPAK*** | Cysteine-rich PAK1 inhibitor (CRIPak) Q8N1N5 |
| ***CRYAB*** | Alpha-crystallin B chain (Alpha(B)-crystallin)(Rosenthal fiber component)(Heat shock protein beta-5)(HspB5)(Renal carcinoma antigen NY-REN-27) P02511 |
| ***CSF3R*** | Granulocyte colony-stimulating factor receptor Precursor (G-CSF-R)(CD114 antigen) Q99062 |
| ***CSTB*** | Cystatin-B (Stefin-B)(Liver thiol proteinase inhibitor)(CPI-B) P04080 |
| ***CSTF3*** | Cleavage stimulation factor 77 kDa subunit (CSTF 77 kDa subunit)(CstF-77)(CF-1 77 kDa subunit) Q12996 |
| ***CTTNBP2*** | Cortactin-binding protein 2 (CortBP2) Q8WZ74 |
| ***CTXN3*** | Cortexin-3 (Kidney and brain-expressed protein) Q4LDR2 |
| ***CYP4F2*** | Cytochrome P450 4F2 (EC 1.14.13.30)(CYPIVF2)(Leukotriene-B(4) omega-hydroxylase)(Leukotriene-B(4) 20-monooxygenase)(Cytochrome P450-LTB-omega) P78329 |
| ***DDX11*** | Probable ATP-dependent RNA helicase DDX11 (EC 3.6.1.-)(DEAD/H box protein 11)(CHL1-related protein 1)(hCHLR1)(Keratinocyte growth factor-regulated gene 2 protein)(KRG-2) Q96FC9 |
| ***DDX31*** | Probable ATP-dependent RNA helicase DDX31 (EC 3.6.1.-)(DEAD box protein 31)(Helicain) Q9H8H2 |
| ***DDX46*** | Probable ATP-dependent RNA helicase DDX46 (EC 3.6.1.-)(DEAD box protein 46)(PRP5 homolog) Q7L014 |
| ***DDX47*** | Probable ATP-dependent RNA helicase DDX47 (EC 3.6.1.-)(DEAD box protein 47) Q9H0S4 |
| ***DDX55*** | ATP-dependent RNA helicase DDX55 (EC 3.6.1.-)(DEAD box protein 55) Q8NHQ9 |
| ***DGKG*** | Diacylglycerol kinase gamma (DAG kinase gamma)(EC 2.7.1.107)(Diglyceride kinase gamma)(DGK-gamma) P49619 |
| ***DGKZ*** | Diacylglycerol kinase zeta (DAG kinase zeta)(EC 2.7.1.107)(Diglyceride kinase zeta)(DGK-zeta) Q13574 |
| ***DIP*** | GRAM domain-containing protein 4 (Death-inducing protein) Q6IC98 |
| ***DLL1*** | Delta-like protein 1 Precursor (Drosophila Delta homolog 1)(Delta1)(H-Delta-1) O00548 |
| ***DMC1*** | Meiotic recombination protein DMC1/LIM15 homolog Q14565 |

**Table A-4.** Continued

| | |
|---|---|
| ***DNAH12L*** | Axonemal dynein heavy chain 12-like protein (Axonemal dynein heavy chain 7-like protein) Q6ZTR8 |
| ***DNAH2*** | Dynein heavy chain 2, axonemal (Axonemal beta dynein heavy chain 2)(Ciliary dynein heavy chain 2)(Dynein heavy chain domain-containing protein 3) Q9P225 |

| | |
|---|---|
| ***DNAH3*** | Dynein heavy chain 3, axonemal (Axonemal beta dynein heavy chain 3)(HsADHC3)(Ciliary dynein heavy chain 3)(Dnahc3-b) Q8TD57 |
| ***DNAH5*** | Dynein heavy chain 5, axonemal (Axonemal beta dynein heavy chain 5)(Ciliary dynein heavy chain 5) Q8TE73 |
| ***DNAJB13*** | DnaJ homolog subfamily B member 13 (Testis spermatocyte apoptosis-related gene 6 protein)(Testis and spermatogenesis cell-related protein 6)(Testis spermatogenesis apoptosis-related gene 6 protein)(Testis spermatogenesis apoptosis-related gene 3 protein) P59910 |
| ***DOM3Z*** | Protein Dom3Z (Dom-3 homolog Z) O77932 |
| ***DQX1*** | ATP-dependent RNA helicase DQX1 (EC 3.6.1.-)(DEAQ box polypeptide 1) Q8TE96 |
| ***DRD5P2*** | Seven transmembrane helix receptor :UniProtKB/TrEMBL;Acc:Q8NH22 |
| ***DYSF*** | Dysferlin (Dystrophy-associated fer-1-like protein)(Fer-1-like protein 1) O75923 |
| ***EFHB*** | EF-hand domain-containing family member B Q8N7U6 |
| ***EFHD2*** | EF-hand domain-containing protein D2 (Swiprosin-1) Q96C19 |
| ***EIF2C1*** | Eukaryotic translation initiation factor 2C 1 (eIF2C 1)(eIF-2C 1)(Argonaute-1)(Putative RNA-binding protein Q99) Q9UL18 |
| ***ELA3A*** | Elastase-3A Precursor (EC 3.4.21.70)(Elastase IIIA)(Protease E) P09093 |
| ***EPM2A*** | Laforin (EC 3.1.3.48)(EC 3.1.3.16)(Lafora PTPase)(LAFPTPase) O95278 |
| ***ERMAP*** | Erythroid membrane-associated protein Precursor (hERMAP)(Scianna blood group antigen)(Radin blood group antigen) Q96PL5 |
| ***EVC*** | Ellis-van Creveld syndrome protein (DWF-1) P57679 |
| ***EXOSC10*** | Exosome component 10 (Polymyositis/scleroderma autoantigen 2)(Autoantigen PM/Scl 2)(Polymyositis/scleroderma autoantigen 100 kDa)(PM/Scl-100)(P100 polymyositis-scleroderma overlap syndrome-associated autoantigen) Q01780 |
| ***EXPH5*** | Exophilin-5 (Synaptotagmin-like protein homolog lacking C2 domains b)(Slp homolog lacking C2 domains b)(SlaC2-b) Q8NEV8 |
| ***FABP6*** | Gastrotropin (GT)(Fatty acid-binding protein 6)(Ileal lipid-binding protein)(ILBP)(Intestinal 15 kDa protein)(I-15P)(Intestinal bile acid-binding protein)(I-BABP) P51161 |

**Table A-4.** Continued

| | |
|---|---|
| ***FAF2*** | FAS-associated factor 2 (UBX domain-containing protein 3B)(UBX domain-containing protein 8)(Protein ETEA) Q96CS3 |
| ***FAM182B*** | Uncharacterized protein DKFZp434B104 Q5T319 |
| ***FAM20B*** | Protein FAM20B Precursor O75063 |

| | |
|---|---|
| ***FAM5B*** | Protein FAM5B Precursor (BMP/retinoic acid-inducible neural-specific protein 2)(DBCCR1-like protein 2) Q9C0B6 |
| ***FAT2*** | Protocadherin Fat 2 Precursor (hFat2)(Multiple epidermal growth factor-like domains 1) Q9NYQ8 |
| ***FAT4*** | Protocadherin Fat 4 Precursor (hFat4)(FAT tumor suppressor homolog 4)(Fat-like cadherin protein FAT-J) Q6V0I7 |
| ***FBXO38*** | F-box only protein 38 (Modulator of KLF7 activity homolog)(MoKA) Q6PIJ6 |
| ***FCER1A*** | High affinity immunoglobulin epsilon receptor subunit alpha Precursor (IgE Fc receptor subunit alpha)(Fc-epsilon RI-alpha)(FcERI) P12319 |
| ***FGFR4*** | Fibroblast growth factor receptor 4 Precursor (FGFR-4)(EC 2.7.10.1)(CD334 antigen) P22455 |
| ***FLAD1*** | FAD synthetase (EC 2.7.7.2)(FMN adenylyltransferase)(FAD pyrophosphorylase)(Flavin adenine dinucleotide synthetase) Includes Molybdenum cofactor biosynthesis protein-like region;FAD synthetase region] Q8NFF5 |
| ***FLAD1*** | FAD synthetase (EC 2.7.7.2)(FMN adenylyltransferase)(FAD pyrophosphorylase)(Flavin adenine dinucleotide synthetase) Includes Molybdenum cofactor biosynthesis protein-like region;FAD synthetase region] Q8NFF5 |
| ***FLG*** | Filaggrin P20930 |
| ***FN1*** | Fibronectin Precursor (FN)(Cold-insoluble globulin)(CIG) Contains Ugl-Y1;Ugl-Y2;Ugl-Y3 P02751] |
| ***FOXO3*** | Forkhead box protein O3 (Forkhead in rhabdomyosarcoma-like 1)(AF6q21 protein) O43524 |
| ***FRG1*** | Protein FRG1 (FSHD region gene 1 protein) Q14331 |
| ***FTMT*** | Ferritin, mitochondrial Precursor (EC 1.16.3.1) Q8N4E7 |
| ***FTSJD2*** | FtsJ methyltransferase domain-containing protein 2 (EC 2.1.1.-) Q8N1G2 |
| ***GABRA6*** | Gamma-aminobutyric acid receptor subunit alpha-6 Precursor (GABA(A) receptor subunit alpha-6) Q16445 |
| ***GABRB1*** | Gamma-aminobutyric acid receptor subunit beta-1 Precursor (GABA(A) receptor subunit beta-1) P18505 |
| ***GANAB*** | Neutral alpha-glucosidase AB Precursor (EC 3.2.1.84)(Glucosidase II subunit alpha)(Alpha-glucosidase 2) Q14697 |
| ***GATAD2B*** | Transcriptional repressor p66-beta (p66/p68)(GATA zinc finger domain-containing protein 2B) Q8WXI9 |
| ***GBF1*** | Golgi-specific brefeldin A-resistance guanine nucleotide exchange factor 1 (BFA-resistant GEF 1) Q92538 |
| ***GBP4*** | Guanylate-binding protein 4 (Guanine nucleotide-binding protein 4)(GTP-binding protein 4)(GBP-4) Q96PP9 |
| ***GEMIN5*** | Gem-associated protein 5 (Gemin5) Q8TEQ6 |
| ***GGA3*** | ADP-ribosylation factor-binding protein GGA3 (Golgi-localized, gamma ear-containing, ARF-binding protein 3) Q9NZ52 |

| | |
|---|---|
| ***GIMAP8*** | GTPase IMAP family member 8 (Immune-associated nucleotide-binding protein 9)(Protein IanT) Q8ND71 |
| ***GLI1*** | Zinc finger protein GLI1 (Glioma-associated oncogene)(Oncogene GLI) P08151 |
| ***GOLGA1*** | Golgin subfamily A member 1 (Golgin-97) Q92805 |
| ***GOLGA6B*** | Golgin subfamily A member 6B A6NDN3 |
| ***GOLGA6C*** | Golgin subfamily A member 6A (Golgin linked to PML)(Golgin-like protein) Q9NYA3 |
| ***GPRC6A*** | G-protein coupled receptor family C group 6 member A Precursor (hGPRC6A)(G-protein coupled receptor 33)(hGPCR33) Q5T6X5 |
| ***GPRIN1*** | G protein-regulated inducer of neurite outgrowth 1 (GRIN1) Q7Z2K8 |
| ***GRIA4*** | Glutamate receptor 4 Precursor (GluR-4)(GluR4)(GluR-D)(Glutamate receptor ionotropic, AMPA 4)(AMPA-selective glutamate receptor 4) P48058 |
| ***GRK4*** | G protein-coupled receptor kinase 4 |
| ***GRK7*** | G protein-coupled receptor kinase 7 Precursor (EC 2.7.11.16)(G protein-coupled receptor kinase GRK7) Q8WTQ7 |
| ***GSN*** | Gelsolin Precursor (Actin-depolymerizing factor)(ADF)(Brevin)(AGEL) P06396 |
| ***GTDC1*** | Glycosyltransferase-like domain-containing protein 1 (Mat-Xa) Q4AE62 |
| ***GTF2IRD2B*** | General transcription factor II-I repeat domain-containing protein 2B (GTF2I repeat domain-containing protein 2B)(Transcription factor GTF2IRD2-beta) Q6EKJ0 |
| ***HCK*** | Tyrosine-protein kinase HCK (EC 2.7.10.2)(Hemopoietic cell kinase)(p59-HCK/p60-HCK) P08631 |
| ***HERC2P2*** | Putative uncharacterized protein HERC2P2 :UniProtKB/TrEMBL;Acc:A6NLQ0 |
| ***HERC2P2*** | Putative uncharacterized protein HERC2P2 :UniProtKB/TrEMBL;Acc:A6NLQ0 |
| ***HLCS*** | Biotin--protein ligase (EC 6.3.4.-)(Biotin apo-protein ligase) Includes Biotin--[methylmalonyl-CoA-carboxytransferase ligase(EC 6.3.4.9);Biotin--[propionyl-CoA-carboxylase [ATP-hydrolyzing ligase(EC 6.3.4.10)(Holocarboxylase synthetase)(HCS);Biotin--[methylcrotonoyl-CoA-carboxylase ligase(EC 6.3.4.11)];Biotin--[acetyl-CoA-carboxylase ligase(EC 6.3.4.15) P50747] |
| ***HMCN1*** | Hemicentin-1 Precursor (Fibulin-6)(FIBL-6) Q96RW7 |

**Table A-4.** Continued

| | |
|---|---|
| ***HMGB2*** | High mobility group protein B2 (High mobility group protein 2)(HMG-2) P26583 |
| ***HOOK3*** | Protein Hook homolog 3 (hHK3) :UniProtKBSwiss-Prot;Acc:Q86VS8 ENSFM00250000001958 hook homolog 3 |
| ***HPSE2*** | Heparanase-2 (Hpa2)(EC 3.2.-.-) Q8WWQ2 |

| | |
|---|---|
| ***HSF2*** | Heat shock factor protein 2 (HSF 2)(Heat shock transcription factor 2)(HSTF 2) Q03933 |
| ***HSPA4L*** | Heat shock 70 kDa protein 4L (Osmotic stress protein 94)(Heat shock 70-related protein APG-1) O95757 |
| ***HYI*** | Putative hydroxypyruvate isomerase (EC 5.3.1.22)(Endothelial cell apoptosis protein E-CE1) Q5T013 |
| ***IDH3A*** | Isocitrate dehydrogenase [NAD subunit alpha, mitochondrial Precursor (EC 1.1.1.41)(Isocitric dehydrogenase)(NAD(+)-specific ICDH) P50213 |
| ***IGF2R*** | Cation-independent mannose-6-phosphate receptor Precursor (CI Man-6-P receptor)(CI-MPR)(M6PR)(Insulin-like growth factor 2 receptor)(Insulin-like growth factor II receptor)(IGF-II receptor)(M6P/IGF2 receptor)(M6P/IGF2R)(300 kDa mannose 6-phosphate receptor)(MPR 300)(CD222 antigen) P11717 |
| ***IGF2R*** | Cation-independent mannose-6-phosphate receptor Precursor (CI Man-6-P receptor)(CI-MPR)(M6PR)(Insulin-like growth factor 2 receptor)(Insulin-like growth factor II receptor)(IGF-II receptor)(M6P/IGF2 receptor)(M6P/IGF2R)(300 kDa mannose 6-phosphate receptor)(MPR 300)(CD222 antigen) P11717 |
| ***IGSF10*** | Immunoglobulin superfamily member 10 Precursor (Calvaria mechanical force protein 608)(CMF608) Q6WRI0 |
| ***IKZF2*** | Zinc finger protein Helios (Ikaros family zinc finger protein 2) Q9UKS7 |
| ***IL22RA1*** | Interleukin-22 receptor subunit alpha-1 Precursor (IL-22R-alpha-1)(Cytokine receptor family 2 member 9)(CRF2-9) Q8N6P7 |
| ***IL6ST*** | Interleukin-6 receptor subunit beta Precursor (IL-6R-beta)(Interleukin-6 signal transducer)(Membrane glycoprotein 130)(gp130)(CDw130)(Oncostatin-M receptor alpha subunit)(CD130 antigen) :UniProtKBSwiss-Prot;Acc:P40189 ENSFM00500000270708 interleukin 6 signal transducer (gp130 oncostatin M receptor) |
| ***IL8RA*** | High affinity interleukin-8 receptor A (IL-8R A)(IL-8 receptor type 1)(CXCR-1)(CDw128a)(CD181 antigen) P25024 |
| ***ILK*** | Integrin-linked protein kinase (EC 2.7.11.1)(ILK-1)(ILK-2)(59 kDa serine/threonine-protein kinase)(p59ILK) Q13418 |
| ***INTS4*** | integrator complex subunit 4 |
| ***KIAA0393*** | Probable E3 ubiquitin-protein ligase HERC2 |
| ***KIAA0556*** | Uncharacterized protein KIAA0556 O60303 |
| ***KIAA1407*** | Coiled-coil domain-containing protein KIAA1407 Q8NCU4 |

**Table A-4.** Continued

| | |
|---|---|
| ***KIAA1841*** | Uncharacterized protein |
| ***KIAA1922*** | CROCCL2; Ciliary rootlet coiled-coil protein-like 2 protein |
| ***KIF20A*** | Kinesin-like protein KIF20A (Rabkinesin-6)(Rab6-interacting kinesin-like protein)(GG10 2) O95235 |

| | |
|---|---|
| ***KIF23*** | Kinesin-like protein KIF23 (Mitotic kinesin-like protein 1)(Kinesin-like protein 5) Q02241 |
| ***KIF3B*** | Kinesin-like protein KIF3B (Microtubule plus end-directed kinesin motor 3B)(HH0048) O15066 |
| ***KIF5A*** | Kinesin heavy chain isoform 5A (Neuronal kinesin heavy chain)(NKHC)(Kinesin heavy chain neuron-specific 1) Q12840 |
| ***KIFC3*** | Kinesin-like protein KIFC3 Q9BVG8 |
| ***KIR2DL4*** | Killer cell immunoglobulin-like receptor 2DL4 Precursor (MHC class I NK cell receptor KIR103AS)(Killer cell inhibitory receptor 103AS)(KIR-103AS)(G9P)(CD158 antigen-like family member D)(CD158d antigen) Q99706 |
| ***KIR2DS4*** | Killer cell immunoglobulin-like receptor 2DS4 Precursor (MHC class I NK cell receptor)(Natural killer-associated transcript 8)(NKAT-8)(P58 natural killer cell receptor clone CL-39)(p58 NK receptor)(CL-17)(CD158 antigen-like family member I)(CD158i antigen) P43632 |
| ***LAMA2*** | Laminin subunit alpha-2 |
| ***LAMB2*** | Laminin subunit beta-2 Precursor (S-laminin)(Laminin B1s chain) P55268 |
| ***LDB2*** | LIM domain-binding protein 2 (Carboxyl-terminal LIM domain-binding protein 1)(CLIM-1)(LIM domain-binding factor CLIM1) O43679 |
| ***LRP8*** | Low-density lipoprotein receptor-related protein 8 Precursor (Apolipoprotein E receptor 2) Q14114 |
| ***LYPLAL1*** | Lysophospholipase-like protein 1 (EC 3.1.2.-) Q5VWZ2 |
| ***LYST*** | Lysosomal-trafficking regulator (Beige homolog) Q99698 |
| ***MAN2B2*** | Epididymis-specific alpha-mannosidase Precursor (EC 3.2.1.24)(Mannosidase alpha class 2B member 2) Q9Y2E5 |
| ***MAP3K5*** | Mitogen-activated protein kinase kinase kinase 5 (EC 2.7.11.25)(MAPK/ERK kinase kinase 5)(MEK kinase 5)(MEKK 5)(Apoptosis signal-regulating kinase 1)(ASK-1) Q99683 |
| ***MAP3K9*** | Mitogen-activated protein kinase kinase kinase 9 (EC 2.7.11.25)(Mixed lineage kinase 1) P80192 |
| ***MAPKBP1*** | Mitogen-activated protein kinase-binding protein 1 (JNK-binding protein 1)(JNKBP1) O60336 |
| ***MAT2B*** | Methionine adenosyltransferase 2 subunit beta (Methionine adenosyltransferase II beta)(MAT II beta)(Methionine adenosyltransferase 2 beta subunit)(DTDP-4-keto-6-deoxy-D-glucose 4-reductase) Q9NZL9 |
| ***MBOAT2*** | Membrane-bound O-acyltransferase domain-containing protein 2 (O-acyltransferase domain-containing protein 2)(EC 2.3.-.-) Q6ZWT7 |

**Table A-4.** Continued

| | |
|---|---|
| ***MC2R*** | Adrenocorticotropic hormone receptor (ACTH receptor)(ACTH-R)(Adrenocorticotropin receptor)(Melanocortin receptor 2)(MC2-R) Q01718 |

| | |
|---|---|
| ***MECR*** | Trans-2-enoyl-CoA reductase, mitochondrial Precursor (HsNrbf-1)(EC 1.3.1.38)(NRBF-1) Q9BV79 |
| ***MFAP2*** | Microfibrillar-associated protein 2 Precursor (MFAP-2)(Microfibril-associated glycoprotein 1)(MAGP-1)(MAGP) P55001 |
| ***MFSD2*** | Major facilitator superfamily domain-containing protein 2 Q8NA29 |
| ***MLL3*** | Histone-lysine N-methyltransferase MLL3 (EC 2.1.1.43)(Myeloid/lymphoid or mixed-lineage leukemia protein 3)(Homologous to ALR protein)(Lysine N-methyltransferase 2C) Q8NEZ4 |
| ***MLLT4*** | Afadin (Protein AF-6) :UniProtKB Swiss-Prot;Acc:P55196 ENSFM00250000001587 myeloid lymphoid or mixed-lineage leukemia (trithorax homolog Drosophila); translocated to 4 (AF6) ENSFM00250000001587 myeloid-lymphoid or mixed-lineage leukemia (trithorax homolog Drosophila); translocated to 4 (AF6) |
| ***MRC1L1*** | Macrophage mannose receptor 1-like protein 1 Precursor (C-type lectin domain family 13 member D-like) Q5VSK2 |
| ***MRPL2*** | 39S ribosomal protein L2, mitochondrial Precursor (L2mt)(MRP-L2) Q5T653 |
| ***MSH6*** | DNA mismatch repair protein Msh6 (MutS-alpha 160 kDa subunit)(GT mismatch-binding protein)(GTMBP)(GTBP)(p160) :UniProtKBSwiss-Prot;Acc:P52701 ENSFM00500000270310 MutS, E. COLI, HOMOLOG OF, 6 mutS homolog 6 (E. coli) |
| ***MSL2*** | Male-specific lethal 2 homolog (MSL-2)(Male-specific lethal 2-like 1)(MSL2-like 1)(Male-specific lethal-2 homolog 1)(RING finger protein 184) Q9HCI7 |
| ***MST1*** | STK4;Serine/threonine-protein kinase 4 |
| ***MUC17*** | Mucin-17 Precursor (MUC-17)(Small intestinal mucin-3)(MUC-3) |
| ***MUC4*** | Mucin-4 Precursor (MUC-4)(Pancreatic adenocarcinoma mucin)(Testis mucin)(Ascites sialoglycoprotein)(ASGP)(Tracheobronchial mucin) Contains Mucin-4 alpha chain(Ascites sialoglycoprotein 1)(ASGP-1);Mucin-4 beta chain(Ascites sialoglycoprotein 2)(ASGP-2) Q99102 |
| ***MYCBPAP*** | MYCBP-associated protein (AMY-1-binding protein 1)(AMAP-1)(AMAM-1) Q8TBZ2 |
| ***MYOM2*** | Myomesin-2 (Myomesin family member 2)(M-protein)(165 kDa titin-associated protein)(165 kDa connectin-associated protein) P54296 |
| ***MYT1*** | Myelin transcription factor 1 (MyT1)(MyTI)(Proteolipid protein-binding protein)(PLPB1) Q01538 |
| ***NACA*** | Nascent polypeptide-associated complex subunit alpha (NAC-alpha)(Alpha-NAC)(Allergen Hom s 2) :UniProtKBSwiss-Prot;Acc:Q13765 ENSFM00500000270490 nascent-polypeptide-associated complex alpha polypeptide |

**Table A-4.** Continued

| | |
|---|---|
| ***NADK*** | NAD kinase (EC 2.7.1.23)(Poly(P)/ATP NAD kinase) O95544 |
| ***NAV1*** | Neuron navigator 1 (Steerin-1)(Pore membrane and/or filament-interacting-like protein 3)(Unc-53 homolog 1)(unc53H1) Q8NEY1 |

| | |
|---|---|
| ***NDUFA13*** | NADH dehydrogenase [ubiquinone 1 alpha subcomplex subunit 13 (NADH-ubiquinone oxidoreductase B16.6 subunit)(Complex I-B16.6)(CI-B16.6)(Gene associated with retinoic-interferon-induced mortality 19 protein)(GRIM-19)(Cell death regulatory protein GRIM-19) Q9P0J0 |
| ***NDUFS5*** | NADH dehydrogenase [ubiquinone iron-sulfur protein 5 (NADH-ubiquinone oxidoreductase 15 kDa subunit)(Complex I-15 kDa)(CI-15 kDa) O43920 |
| ***NFATC3*** | Nuclear factor of activated T-cells, cytoplasmic 3 (NF-ATc3)(NFATc3)(T-cell transcription factor NFAT4)(NF-AT4)(NFATx) Q12968 |
| ***NFRKB*** | Nuclear factor related to kappa-B-binding protein (DNA-binding protein R kappa-B)(INO80 complex subunit G) Q6P4R8 |
| ***NGF*** | Beta-nerve growth factor Precursor (Beta-NGF) P01138 |
| ***NIN*** | Ninein (hNinein)(Glycogen synthase kinase 3 beta-interacting protein)(GSK3B-interacting protein) :UniProtKBSwiss-Prot;Acc:Q8N4C6 ENSFM00250000001557 ninein (GSK3B interacting protein) |
| ***NLGN1*** | Neuroligin-1 Precursor Q8N2Q7 |
| ***NOMO3*** | Nodal modulator 3 Precursor (pM5 protein 3) P69849 |
| ***NOS3*** | Nitric oxide synthase, endothelial (EC 1.14.13.39)(Endothelial NOS)(eNOS)(EC-NOS)(NOS type III)(NOSIII)(Constitutive NOS)(cNOS) P29474 |
| ***NOTCH2*** | Neurogenic locus notch homolog protein 2 Precursor (Notch 2)(hN2) Contains Notch 2 extracellular truncation;Notch 2 intracellular domain :UniProtKBSwiss-Prot;Acc:Q04721 ENSFM00500000269589 ALAGILLE SYNDROME 2 Notch homolog 2 |
| ***NPR1*** | Atrial natriuretic peptide receptor A Precursor (EC 4.6.1.2)(Atrial natriuretic peptide A-type receptor)(ANPRA)(ANP-A)(NPR-A)(Guanylate cyclase)(GC-A) P16066 |
| ***NR1D1*** | Nuclear receptor subfamily 1 group D member 1 (V-erbA-related protein EAR-1)(Rev-erbA-alpha) P20393 |
| ***NR4A3*** | Nuclear receptor subfamily 4 group A member 3 (Nuclear hormone receptor NOR-1)(Neuron-derived orphan receptor 1)(Mitogen-induced nuclear orphan receptor) Q92570 |
| ***NR4A3*** | Nuclear receptor subfamily 4 group A member 3 (Nuclear hormone receptor NOR-1)(Neuron-derived orphan receptor 1)(Mitogen-induced nuclear orphan receptor) :UniProtKBSwiss-Prot;Acc:Q92570 ENSFM00250000001653 nuclear receptor subfamily 4 group A member 3 (NOR1) |
| ***NTRK2*** | BDNF/NT-3 growth factors receptor Precursor (EC 2.7.10.1)(Neurotrophic tyrosine kinase receptor type 2)(TrkB tyrosine kinase)(GP145-TrkB)(Trk-B) Q16620 |

**Table A-4.** Continued

| | |
|---|---|
| ***NUP133*** | Nuclear pore complex protein Nup133 (Nucleoporin Nup133)(133 kDa nucleoporin) Q8WUM0 |

| | |
|---|---|
| ***NUP214*** | Nuclear pore complex protein Nup214 (Nucleoporin Nup214)(214 kDa nucleoporin)(Protein CAN) P35658 |
| ***OLFML1*** | Olfactomedin-like protein 1 Precursor Q6UWY5 |
| ***OR10A2*** | Olfactory receptor 10A2 (Olfactory receptor OR11-86)(HP4) Q9H208 |
| ***OR10H5*** | Olfactory receptor 10H5 (Olfactory receptor OR19-25)(Olfactory receptor OR19-26) Q8NGA6 |
| ***OR13C5*** | Olfactory receptor 13C5 (Olfactory receptor OR9-11) Q8NGS8 |
| ***OR14A16*** | Olfactory receptor 14A16 (Olfactory receptor 5AT1)(Olfactory receptor OR1-45) Q8NHC5 |
| ***OR2L8*** | Olfactory receptor 2L8 (Olfactory receptor OR1-46) Q8NGY9 |
| ***OR2T3*** | Olfactory receptor 2T3 Q8NH03 |
| ***OR2T34*** | Olfactory receptor 2T34 (Olfactory receptor OR1-63) Q8NGX1 |
| ***OR2T4*** | Olfactory receptor 2T4 (Olfactory receptor OR1-60) Q8NH00 |
| ***OR4F17*** | Olfactory receptor 4F17 Q8NGA8 |
| ***OR6K2*** | Olfactory receptor 6K2 (Olfactory receptor OR1-17) Q8NGY2 |
| ***OR8H1*** | Olfactory receptor 8H1 (Olfactory receptor OR11-180) Q8NGG4 |
| ***OR9G9*** | Olfactory receptor 9G1 (Olfactory receptor 9G5)(Olfactory receptor OR11-114) Q8NH87 |
| ***OR9Q1*** | Olfactory receptor 9Q1 Q8NGQ5 |
| ***ORC1L*** | Origin recognition complex subunit 1 (Replication control protein 1) Q13415 |
| ***OTUD4*** | OTU domain-containing protein 4 (HIV-1-induced protein HIN-1) Q01804 |
| ***OVGP1*** | Oviduct-specific glycoprotein Precursor (Oviductal glycoprotein)(Oviductin)(Estrogen-dependent oviduct protein)(Mucin-9) Q12889 |
| ***PABPC4*** | Polyadenylate-binding protein 4 (Poly(A)-binding protein 4)(PABP 4)(Inducible poly(A)-binding protein)(iPABP)(Activated-platelet protein 1)(APP-1) Q13310 |
| ***PAF1*** | RNA polymerase II-associated factor 1 homolog (hPAF1)(Pancreatic differentiation protein 2) Q8N7H5 |
| ***PBX4*** | Pre-B-cell leukemia transcription factor 4 (Homeobox protein PBX4) Q9BYU1 |
| ***PCDHB2*** | Protocadherin beta-2 Precursor (PCDH-beta-2) Q9Y5E7 |
| ***PCDHGA12*** | Protocadherin gamma-A12 Precursor (PCDH-gamma-A12)(Cadherin-21)(Fibroblast cadherin-3) O60330 |
| ***PCGF2*** | Polycomb group RING finger protein 2 |
| ***PDE3A*** | cGMP-inhibited 3',5'-cyclic phosphodiesterase A (EC 3.1.4.17)(Cyclic GMP-inhibited phosphodiesterase A)(CGI-PDE A) Q14432 |
| ***PDE4DIP*** | Myomegalin (Phosphodiesterase 4D-interacting protein)(Cardiomyopathy-associated protein 2) Q5VU43 |

**Table A-4.** Continued

| | |
|---|---|
| ***PDIA6*** | Protein disulfide-isomerase A6 Precursor (EC 5.3.4.1)(Protein disulfide isomerase P5)(Thioredoxin domain-containing protein 7) Q15084 |

| | |
|---|---|
| ***PEX5L*** | PEX5-related protein (Peroxisome biogenesis factor 5-like)(Peroxin-5-related protein)(Pex5Rp)(PEX5-like protein)(PEX2-related protein) Q8IYB4 |
| ***PFKFB4*** | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 (6PF-2-K/Fru-2,6-P2ASE testis-type isozyme) Includes 6-phosphofructo-2-kinase(EC 2.7.1.105);Fructose-2,6-bisphosphatase(EC 3.1.3.46) Q16877 |
| ***PFKL*** | 6-phosphofructokinase, liver type (EC 2.7.1.11)(Phosphofructokinase 1)(Phosphohexokinase)(Phosphofructo-1-kinase isozyme B)(PFK-B) P17858 |
| ***PHGDH*** | D-3-phosphoglycerate dehydrogenase (3-PGDH)(EC 1.1.1.95) O43175 |
| ***PIK3C2B*** | Phosphatidylinositol-4-phosphate 3-kinase C2 domain-containing beta polypeptide (EC 2.7.1.154)(Phosphoinositide 3-Kinase-C2-beta)(PtdIns-3-kinase C2 beta)(PI3K-C2beta)(C2-PI3K) O00750 |
| ***PIK3CD*** | Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta isoform (EC 2.7.1.153)(PI3-kinase p110 subunit delta)(PtdIns-3-kinase p110)(PI3K)(p110delta) O00329 |
| ***PIP5K2B*** | Phosphatidylinositol-5-phosphate 4-kinase type-2 beta (EC 2.7.1.149)(Phosphatidylinositol-5-phosphate 4-kinase type II beta)(1-phosphatidylinositol-5-phosphate 4-kinase 2-beta)(PtdIns(5)P-4-kinase isoform 2-beta)(PIP4KII-beta)(Diphosphoinositide kinase 2-beta) P78356 |
| ***PIPSL*** | Putative PIP5K1A and PSMD4-like protein (PIP5K1A-PSMD4) A2A3N6 |
| ***PKP1*** | Plakophilin-1 (Band-6 protein)(B6P) Q13835 |
| ***PLA2R1*** | Secretory phospholipase A2 receptor Precursor (PLA2-R)(PLA2R)(180 kDa secretory phospholipase A2 receptor)(M-type receptor) Contains Soluble secretory phospholipase A2 receptor(Soluble PLA2-R)(Soluble PLA2R) Q13018 |
| ***PLCL1*** | Inactive phospholipase C-like protein 1 (PLC-L1)(Phospholipase C-deleted in lung carcinoma)(Phospholipase C-related but catalytically inactive protein)(PRIP) Q15111 |
| ***PLGLA1*** | Plasminogen-related protein A Precursor (Plasminogen-like protein A)(Plasminogen-like protein A1) Q15195 |
| ***PLSCR4*** | Phospholipid scramblase 4 (PL scramblase 4)(Ca(2+)-dependent phospholipid scramblase 4)(TRA1)(Cell growth-inhibiting gene 43 protein) Q9NRQ2 |
| ***POLDIP3*** | Polymerase delta-interacting protein 3 (46 kDa DNA polymerase delta interaction protein)(p46) Q9BY77 |
| ***POTEC*** | ANKRD26-like family B member 1 (Prostate, ovary, testis-expressed protein on chromosome 15)(POTE-15) Q6S5H4 |

**Table A-4.** Continued

| | |
|---|---|
| ***PPARGC1B*** | Peroxisome proliferator-activated receptor gamma coactivator 1-beta (PPAR-gamma coactivator 1-beta)(PPARGC-1-beta)(PGC-1-beta)(PGC-1-related estrogen receptor alpha coactivator) Q86YN6 |

| | |
|---|---|
| ***PPP1R10*** | Serine/threonine-protein phosphatase 1 regulatory subunit 10 (Phosphatase 1 nuclear targeting subunit)(MHC class I region proline-rich protein CAT53)(FB19 protein)(PP1-binding protein of 114 kDa)(p99) Q96QC0 |
| ***PRAMEF23*** | PRAME family member 23 A6NMV5 |
| ***PRAMEF5*** | PRAME family member 5 Q5TYX0 |
| ***PREP*** | Prolyl endopeptidase (PE)(EC 3.4.21.26)(Post-proline cleaving enzyme) P48147 |
| ***PRKRIR*** | 52 kDa repressor of the inhibitor of the protein kinase (p58IPK-interacting protein)(58 kDa interferon-induced protein kinase-interacting protein)(P52rIPK)(Death-associated protein 4)(THAP domain-containing protein 0) O43422 |
| ***PRPF4B*** | Serine/threonine-protein kinase PRP4 homolog (EC 2.7.11.1)(PRP4 pre-mRNA-processing factor 4 homolog)(PRP4 kinase) Q13523 |
| ***PRSS12*** | Neurotrypsin Precursor (EC 3.4.21.-)(Serine protease 12)(Motopsin)(Leydin) P56730 |
| ***PSG8*** | Pregnancy-specific beta-1-glycoprotein 8 Precursor (PSBG-8) Q9UQ74 |
| ***PSMD2*** | 26S proteasome non-ATPase regulatory subunit 2 (26S proteasome regulatory subunit RPN1)(26S proteasome regulatory subunit S2)(26S proteasome subunit p97)(Tumor necrosis factor type 1 receptor-associated protein 2)(55.11 protein) Q13200 |
| ***PTAFR*** | Platelet-activating factor receptor (PAF-R) P25105 |
| ***PTPN7*** | Tyrosine-protein phosphatase non-receptor type 7 (EC 3.1.3.48)(Protein-tyrosine phosphatase LC-PTP)(Hematopoietic protein-tyrosine phosphatase)(HEPTP) P35236 |
| ***PUM1*** | Pumilio homolog 1 (Pumilio-1)(HsPUM) Q14671 |
| ***QSOX1*** | Sulfhydryl oxidase 1 Precursor (hQSOX)(EC 1.8.3.2)(Quiescin Q6) O00391 |
| ***RAD21*** | Double-strand-break repair protein rad21 homolog (hHR21)(Nuclear matrix protein 1)(NXP-1)(SCC1 homolog) O60216 |
| ***RAG2*** | V(D)J recombination-activating protein 2 (RAG-2) P55895 |
| ***RANBP2*** | E3 SUMO-protein ligase RanBP2 (Ran-binding protein 2)(Nuclear pore complex protein Nup358)(Nucleoporin Nup358)(358 kDa nucleoporin)(p270) P49792 |
| ***RBM15*** | Putative RNA-binding protein 15 (RNA-binding motif protein 15)(One-twenty two protein) Q96T37 |
| ***RBM16*** | Putative RNA-binding protein 16 (RNA-binding motif protein 16) Q9UPN6 |
| ***RBM19*** | Probable RNA-binding protein 19 (RNA-binding motif protein 19) Q9Y4C8 |

**Table A-4.** Continued

| | |
|---|---|
| ***RBMXL1*** | kynurenine aminotransferase III isoform 3 :RefSeq peptide;Acc:NP 062556 |
| ***RERE*** | Arginine-glutamic acid dipeptide repeats protein (Atrophin-1-related protein)(Atrophin-1-like protein) Q9P2R6 |

| | |
|---|---|
| ***RET*** | Proto-oncogene tyrosine-protein kinase receptor ret Precursor (C-ret)(EC 2.7.10.1) P07949 |
| ***RNF145*** | RING finger protein 145 Q96MT1 |
| ***RNF19B*** | E3 ubiquitin-protein ligase RNF19B (EC 6.3.2.-)(RING finger protein 19B)(IBR domain-containing protein 3)(Natural killer lytic-associated molecule) Q6ZMZ0 |
| ***SCTR*** | Secretin receptor Precursor (SCT-R) P47872 |
| ***SEC31A*** | Protein transport protein Sec31A (SEC31-related protein A)(SEC31-like 1)(ABP125)(ABP130)(Web1-like protein) O94979 |
| ***SEMA5B*** | Semaphorin-5B Q9P283 |
| ***SERPINB2*** | Plasminogen activator inhibitor 2 Precursor (PAI-2)(Placental plasminogen activator inhibitor)(Monocyte Arg-serpin)(Urokinase inhibitor) P05120 |
| ***SETD2*** | Histone-lysine N-methyltransferase SETD2 (EC 2.1.1.43)(SET domain-containing protein 2)(hSET2)(Huntingtin-interacting protein B)(Huntingtin yeast partner B)(Huntingtin-interacting protein 1)(HIF-1)(p231HBP)(Lysine N-methyltransferase 3A) :UniProtKBSwiss-Prot;Acc:Q9BYW2 ENSFM00250000002977 SET domain containing 2 |
| ***SETD6*** | SET domain-containing protein 6 Q8TBK2 |
| ***SFRS15*** | Splicing factor, arginine/serine-rich 15 (CTD-binding SR-like protein RA4) O95104 |
| ***SGK1*** | Serine/threonine-protein kinase Sgk1 (EC 2.7.11.1)(Serum/glucocorticoid-regulated kinase 1) O00141 |
| ***SH3TC1*** | SH3 domain and tetratricopeptide repeats-containing protein 1 |
| ***SHISA3*** | Protein shisa-3 homolog Precursor A0PJX4 |
| ***SI*** | Sucrase-isomaltase, intestinal Contains Sucrase(EC 3.2.1.48);Isomaltase(EC 3.2.1.10) P14410 |
| ***SIRPA*** | Tyrosine-protein phosphatase non-receptor type substrate 1 Precursor (SHP substrate 1)(SHPS-1)(Inhibitory receptor SHPS-1)(Signal-regulatory protein alpha-1)(Sirp-alpha-1)(Sirp-alpha-2)(Sirp-alpha-3)(MyD-1 antigen)(Brain Ig-like molecule with tyrosine-based activation motifs)(Bit)(Macrophage fusion receptor)(p84)(CD172 antigen-like family member A)(CD172a antigen) P78324 |
| ***SKIV2L*** | superkiller viralicidic activity 2-like homolog :RefSeq peptide;Acc:NP 008860 |
| ***SLC12A2*** | Solute carrier family 12 member 2 (Bumetanide-sensitive sodium-(potassium)-chloride cotransporter 1)(Basolateral Na-K-Cl symporter) P55011 |
| ***SLC16A14*** | Monocarboxylate transporter 14 (MCT 14)(Solute carrier family 16 member 14) Q7RTX9 |

**Table A-4.** Continued

| | |
|---|---|
| ***SLC1A1*** | Excitatory amino acid transporter 3 (Sodium-dependent glutamate/aspartate transporter 3)(Excitatory amino-acid carrier 1)(Neuronal and epithelial glutamate transporter)(Solute carrier family 1 member 1) P43005 |

| | |
|---|---|
| ***SLC25A12*** | Calcium-binding mitochondrial carrier protein Aralar1 (Mitochondrial aspartate glutamate carrier 1)(Solute carrier family 25 member 12) O75746 |
| ***SLC25A36*** | Solute carrier family 25 member 36 Q96CQ1 |
| ***SLC26A2*** | Sulfate transporter (Diastrophic dysplasia protein)(Solute carrier family 26 member 2) P50443 |
| ***SLC35A1*** | CMP-sialic acid transporter (CMP-Sia-Tr)(CMP-SA-Tr)(Solute carrier family 35 member A1) P78382 |
| ***SLC35F1*** | Solute carrier family 35 member F1 Q5T1Q4 |
| ***SLC44A1*** | Choline transporter-like protein 1 (Solute carrier family 44 member 1)(CDw92)(CD92 antigen) Q8WWI5 |
| ***SLC44A3*** | Choline transporter-like protein 3 (Solute carrier family 44 member 3) Q8N4M1 |
| ***SLC6A9*** | Sodium- and chloride-dependent glycine transporter 1 (GlyT1)(GlyT-1)(Solute carrier family 6 member 9) P48067 |
| ***SLC8A1*** | Sodium/calcium exchanger 1 Precursor (Na(+)/Ca(2+)-exchange protein 1) P32418 |
| ***SLC9A2*** | Sodium/hydrogen exchanger 2 (Na(+)/H(+) exchanger 2)(NHE-2)(Solute carrier family 9 member 2) Q9UBY0 |
| ***SLIT3*** | Slit homolog 3 protein Precursor (Slit-3)(Multiple epidermal growth factor-like domains 5) O75094 |
| ***SMEK1*** | Serine/threonine-protein phosphatase 4 regulatory subunit 3A (SMEK homolog 1) Q6IN85 |
| ***SNCA*** | Alpha-synuclein (Non-A beta component of AD amyloid)(Non-A4 component of amyloid precursor)(NACP) P37840 |
| ***SNRPD1*** | Small nuclear ribonucleoprotein Sm D1 (Sm-D1)(snRNP core protein D1)(Sm-D autoantigen) P62314 |
| ***SNRPG*** | Small nuclear ribonucleoprotein G (snRNP-G)(Sm protein G)(Sm-G)(SmG) P62308 |
| ***SP140*** | Nuclear body protein SP140 (Nuclear autoantigen Sp-140)(Speckled 140 kDa)(Lymphoid-restricted homolog of Sp100)(LYSp100 protein) Q13342 |
| ***SPAG11A*** | sperm associated antigen 11B isoform H preproprotein :RefSeq peptide;Acc:NP 478109 |
| ***SPATA21*** | Spermatogenesis-associated protein 21 Q7Z572 |
| ***SPATA5*** | Spermatogenesis-associated protein 5 (Spermatogenesis-associated factor protein)(ATPase family protein 2 homolog) Q8NB90 |
| ***SPNS3*** | Protein spinster homolog 3 Q6ZMD2 |

**Table A-4.** Continued

| | |
|---|---|
| ***SPTBN4*** | Spectrin beta chain, brain 3 (Spectrin, non-erythroid beta chain 3)(Beta-IV spectrin) Q9H254 |

| | |
|---|---|
| ***SRRM1*** | Serine/arginine repetitive matrix protein 1 (Ser/Arg-related nuclear matrix protein)(SR-related nuclear matrix protein of 160 kDa)(SRm160) Q8IYB3 |
| ***ST6GAL1*** | Beta-galactoside alpha-2,6-sialyltransferase 1 (EC 2.4.99.1)(CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,6-sialyltransferase 1)(Alpha 2,6-ST)(Sialyltransferase 1)(ST6Gal I)(B-cell antigen CD75) P15907 |
| ***STAB1*** | Stabilin-1 Precursor (Fasciclin, EGF-like, laminin-type EGF-like and link domain-containing scavenger receptor 1)(FEEL-1)(MS-1 antigen) Q9NY15 |
| ***STARD8*** | Accelerates GTPase activity of RHOA and CDC42, but not RAC1. Stimulates the hydrolysis of phosphatidylinositol 4,5-bisphosphate by PLCD1 |
| ***STRC*** | Stereocilin Precursor Q7RTU9 |
| ***STX3*** | Syntaxin-3 Q13277 |
| ***SUGT1*** | Suppressor of G2 allele of SKP1 homolog (Sgt1)(Putative 40-6-3 protein) Q9Y2Z0 |
| ***SULT1A4*** | Sulfotransferase 1A3/1A4 (EC 2.8.2.1)(Monoamine-sulfating phenol sulfotransferase)(Aryl sulfotransferase 1A3/1A4)(Sulfotransferase, monoamine-preferring)(M-PST)(Thermolabile phenol sulfotransferase)(TL-PST)(Placental estrogen sulfotransferase)(Catecholamine-sulfating phenol sulfotransferase)(HAST3) P50224 |
| ***SYCP1*** | Synaptonemal complex protein 1 (SCP-1)(Cancer/testis antigen 8)(CT8) Q15431 |
| ***SYCP2*** | Synaptonemal complex protein 2 (SCP-2)(Synaptonemal complex lateral element protein)(hsSCP2) Q9BX26 |
| ***SYNE1*** | Nesprin-1 (Nuclear envelope spectrin repeat protein 1)(Synaptic nuclear envelope protein 1)(Syne-1)(Myocyte nuclear envelope protein 1)(Myne-1)(Enaptin) Q8NF91 |
| ***SYNGAP1*** | Ras GTPase-activating protein SynGAP (Synaptic Ras GTPase-activating protein 1)(Synaptic Ras-GAP 1)(Neuronal RasGAP) Q96PV0 |
| ***SYT2*** | Synaptotagmin-2 (Synaptotagmin II)(SytII) Q8N9I0 |
| ***TAF9*** | Transcription initiation factor TFIID subunit 9 (Transcription initiation factor TFIID 31 kDa subunit)(TAFII-31)(TAFII-32)(TAFII32)(RNA polymerase II TBP-associated factor subunit G)(STAF31/32) Q16594 |
| ***TAGAP*** | T-cell activation Rho GTPase-activating protein (T-cell activation GTPase-activating protein) Q8N103 |
| ***TBC1D29*** | Putative TBC1 domain family member 29 Q9UFV1 |
| ***TBC1D9B*** | TBC1 domain family member 9B Q66K14 |
| ***TBX15*** | T-box transcription factor TBX15 (T-box protein 15) Q96SF7 |
| ***TDG*** | G/T mismatch-specific thymine DNA glycosylase (EC 3.2.2.-) Q13569 |

**Table A-4.** Continued

| | |
|---|---|
| ***TGIF*** | Homeobox protein TGIF1 (5'-TG-3'-interacting factor 1) Q15583 |

| | |
|---|---|
| ***THRAP3*** | Thyroid hormone receptor-associated protein 3 (Thyroid hormone receptor-associated protein complex 150 kDa component)(Trap150) :UniProtKBSwiss-Prot;Acc:Q9Y2W1 ENSFM00250000003730 thyroid hormone receptor associated protein 3 (TRAP150) |
| ***TIAM2*** | T-lymphoma invasion and metastasis-inducing protein 2 (TIAM-2)(SIF and TIAM1-like exchange factor) Q8IVF5 |
| ***TICAM2*** | TIR domain-containing adapter molecule 2 (TICAM-2)(TRIF-related adapter molecule)(Toll/interleukin-1 receptor domain-containing protein)(Toll-like receptor adaptor protein 3)(Putative NF-kappa-B-activating protein 502) Q86XR7 |
| ***TIMELESS*** | Protein timeless homolog (hTIM) Q9UNS1 |
| ***TMC2*** | Transmembrane channel-like protein 2 (Transmembrane cochlear-expressed protein 2) Q8TDI7 |
| ***TMEM127*** | Transmembrane protein 127 O75204 |
| ***TMEM160*** | Transmembrane protein 160 Q9NX00 |
| ***TP53BP2*** | Apoptosis-stimulating of p53 protein 2 (Tumor suppressor p53-binding protein 2)(p53-binding protein 2)(p53BP2)(53BP2)(Bcl2-binding protein)(Bbp)(Renal carcinoma antigen NY-REN-51) Q13625 |
| ***TRHR*** | Thyrotropin-releasing hormone receptor (TRH-R)(Thyroliberin receptor) P34981 |
| ***TRIM2*** | Tripartite motif-containing protein 2 (RING finger protein 86) Q9C040 |
| ***TRIM45*** | Tripartite motif-containing protein 45 (RING finger protein 99) Q9H8W5 |
| ***TSPYL2*** | Testis-specific Y-encoded-like protein 2 |
| ***TTBK1*** | Tau-tubulin kinase 1 (EC 2.7.11.1)(Brain-derived tau kinase) Q5TCY1 |
| ***TTLL6*** | Tubulin polyglutamylase TTLL6 (EC 6.-.-.-)(Tubulin--tyrosine ligase-like protein 6) Q8N841 |
| ***UBA5*** | Ubiquitin-like modifier-activating enzyme 5 (Ubiquitin-activating enzyme 5)(Ubiquitin-activating enzyme E1 domain-containing protein 1)(UFM1-activating enzyme)(ThiFP1) Q9GZZ9 |
| ***UBE2T*** | Ubiquitin-conjugating enzyme E2 T (EC 6.3.2.19)(Ubiquitin-protein ligase T)(Ubiquitin carrier protein T) Q9NPD8 |
| ***UBR4*** | E3 ubiquitin-protein ligase UBR4 (EC 6.3.2.-)(N-recognin-4)(Zinc finger UBR1-type protein 1)(Retinoblastoma-associated factor of 600 kDa)(600 kDa retinoblastoma protein-associated factor)(RBAF600)(p600) Q5T4S7 |
| ***UGT1A8*** | UDP-glucuronosyltransferase 1-8 Precursor (EC 2.4.1.17)(UDP-glucuronosyltransferase 1A8)(UDPGT)(UGT1*8)(UGT1-08)(UGT1.8)(UGT-1H)(UGT1H) Q9HAW9 |

**Table A-4.** Continued

| | |
|---|---|
| ***UGT1A8*** | UDP-glucuronosyltransferase 1-8 Precursor (EC 2.4.1.17)(UDP-glucuronosyltransferase 1A8)(UDPGT)(UGT1*8)(UGT1-08)(UGT1.8)(UGT-1H)(UGT1H) Q9HAW9 |

| | |
|---|---|
| ***UPF1*** | Regulator of nonsense transcripts 1 (EC 3.6.1.-)(ATP-dependent helicase RENT1)(Nonsense mRNA reducing factor 1)(NORF1)(Up-frameshift suppressor 1 homolog)(hUpf1) Q92900 |
| ***USF1*** | Upstream stimulatory factor 1 (Major late transcription factor 1) P22415 |
| ***USP49*** | Ubiquitin carboxyl-terminal hydrolase 49 (EC 3.1.2.15)(Ubiquitin thioesterase 49)(Ubiquitin-specific-processing protease 49)(Deubiquitinating enzyme 49) Q70CQ1 |
| ***USP8*** | Ubiquitin carboxyl-terminal hydrolase 8 (EC 3.1.2.15)(Ubiquitin thioesterase 8)(Ubiquitin-specific-processing protease 8)(Deubiquitinating enzyme 8)(hUBPy) P40818 |
| ***UTP15*** | U3 small nucleolar RNA-associated protein 15 homolog Q8TED0 |
| ***UTRN*** | Utrophin (Dystrophin-related protein 1)(DRP1)(DRP) P46939 |
| ***VARSL*** | valyl-tRNA synthetase 2, mitochondrial (putative) (VARS2), nuclear gene encoding mitochondrial protein, mRNA :RefSeq DNA;Acc:NM 020442 |
| ***VPS13B*** | Vacuolar protein sorting-associated protein 13B (Cohen syndrome protein 1) Q7Z7G8 |
| ***VPS13D*** | Vacuolar protein sorting-associated protein 13D Q5THJ4 |
| ***VPS13D*** | Vacuolar protein sorting-associated protein 13D Q5THJ4 |
| ***VPS52*** | Vacuolar protein sorting-associated protein 52 homolog (SAC2 suppressor of actin mutations 2-like protein) Q8N1B4 |
| ***VPS54*** | Vacuolar protein sorting-associated protein 54 (Hepatocellular carcinoma protein 8)(HOM-HCC-8)(Tumor antigen SLP-8p) Q9P1Q0 |
| ***VRK1*** | Serine/threonine-protein kinase VRK1 (EC 2.7.11.1)(Vaccinia-related kinase 1) Q99986 |
| ***WDR59*** | WD repeat-containing protein 59 Q6PJI9 |
| ***WFDC12*** | WAP four-disulfide core domain protein 12 Precursor (Putative protease inhibitor WAP12)(Whey acidic protein 2) Q8WWY7 |
| ***WNT4*** | Protein Wnt-4 Precursor P56705 |
| ***XAGE5*** | G antigen family D member 5 |
| ***YY1AP1*** | YY1-associated protein 1 (Hepatocellular carcinoma susceptibility protein)(Hepatocellular carcinoma-associated protein 2) Q9H869 |
| ***ZBTB11*** | Zinc finger and BTB domain-containing protein 11 O95625 |
| ***ZBTB33*** | Transcriptional regulator Kaiso |
| ***ZBTB40*** | Zinc finger and BTB domain-containing protein 40 Q9NUA8 |
| ***ZC3H11A*** | Zinc finger CCCH domain-containing protein 11A O75152 |

**Table A-4.** Continued

| | |
|---|---|
| ***ZEB1*** | Zinc finger E-box-binding homeobox 1 (Transcription factor 8)(NIL-2-A zinc finger protein)(Negative regulator of IL2) P37275 |

| | |
|---|---|
| ***ZFYVE20*** | Rabenosyn-5 (FYVE finger-containing Rab5 effector protein rabenosyn-5)(Zinc finger FYVE domain-containing protein 20)(110 kDa protein) Q9H1K0 |
| ***ZFYVE26*** | Zinc finger FYVE domain-containing protein 26 (Spastizin) Q68DK2 |
| ***ZKSCAN2*** | Zinc finger protein with KRAB and SCAN domains 2 (Zinc finger protein 694) Q63HK3 |
| ***ZMYND11*** | Zinc finger MYND domain-containing protein 11 (Adenovirus 5 E1A-binding protein)(Protein BS69) Q15326 |
| ***ZNF141*** | Zinc finger protein 141 Q15928 |
| ***ZNF276*** | Zinc finger protein 276 (Zfp-276) Q8N554 |
| ***ZNF350*** | Zinc finger protein 350 (Zinc finger protein ZBRK1)(Zinc finger and BRCA1-interacting protein with a KRAB domain 1)(KRAB zinc finger protein ZFQR) Q9GZX5 |
| ***ZNF410*** | Zinc finger protein 410 (Zinc finger protein APA-1)(Another partner for ARF 1) Q86VK4 |
| ***ZNF416*** | Zinc finger protein 416 Q9BWM5 |
| ***ZNF509*** | Zinc finger protein 509 Q6ZSB9 |
| ***ZNF588*** | Zinc finger protein 107 (Zinc finger protein 588)(Zinc finger protein ZFD25) Q9UII5 |
| ***ZNF662*** | Zinc finger protein 662 Q6ZS27 |
| ***ZNF664*** | Zinc finger protein 664 (Zinc finger protein from organ of Corti) Q8N3J9 |
| ***ZNF687*** | Zinc finger protein 687 Q8N1G0 |
| ***ZNF806*** | Zinc finger protein 285A Q96NJ3 |