8-2015

Germline Mutation Detection in Next Generation Sequencing Data and TP53 Mutation Carrier Probability Estimation for Li-Fraumeni Syndrome

Gang Peng

GERMLINE MUTATION DETECTION IN NEXT GENERATION SEQUENCING DATA AND *TP53* MUTATION CARRIER PROBABILITY ESTIMATION FOR LI-FRAUMENI SYNDROME

by

Gang Peng, B.S.

APPROVED:

_____

Wenyi Wang, Ph.D., Supervisory Professor

_____

Keith A. Baggerly, Ph.D.

_____

Han Liang, Ph.D.

_____

Paul A. Scheet, Ph.D.

_____

Louise C. Strong, M.D.

APPROVED:

_____

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

GERMLINE MUTATION DETECTION IN NEXT GENERATION SEQUENCING DATA AND *TP53* MUTATION CARRIER PROBABILITY ESTIMATION FOR LI-FRAUMENI SYNDROME

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

MD Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Gang Peng, B.S.

Houston, Texas

August, 2015

ACKNOWLEDGEMENTS

First and foremost, I want to give my utmost gratitude to my advisor, Dr. Wenyi Wang, for her mentoring with knowledge, wisdom, patience and encouragement for my research and personal life. Her optimism, enthusiasm and ingenuity of science have set a model for me. It is a great pleasure and honor to study and work with her in the last five years. I would also like to convey my dearest thanks to my advisory committee members Drs. Keith A. Baggerly, Han Liang, Paul A. Scheet, Louise C. Strong and Ying Yuan. Their encouragement and incisive advising help me a lot to improve my research. I also wish to express my thanks to my candidacy exam committee members and other faculties in the Department of Biostatistics and Department of Bioinformatics and Computational Biology including Drs. Veerabhadran Baladandayuthapani, Ken Chen, Mary E. Edgerton, Yuan Ji, Nicholas Navin and Jing Ning.

I thank Drs. Victoria P. Knutson, William W. Mattox, Ms. Brenda Gaughan and other GSBS staffs for their academic oversight and suggestions that make my Ph.D. life easier. I also want to thank other members in Dr. Wenyi Wang's lab: Yu Fan, Jialu Li, Zeya Wang, Amir Nikooienejad, Jaeil Ahn and Seung Jun Shin. I learned a lot from the insightful discussion with them.

Lastly, I would like to express my gratitude to my parents. Without their unconditional love and support, I cannot be here to pursuit what I want.

GERMLINE MUTATION DETECTION IN NEXT GENERATION SEQUENCING DATA AND *TP53* MUTATION CARRIER PROBABILITY ESTIMATION FOR LI-FRAUMENI SYNDROME

Gang Peng, B.S.

Advisory Professor: Wenyi Wang, Ph.D.

Next generation sequencing technology has been widely used in genomic analysis, but its application has been compromised by the missing true variants, especially when these variants are rare. We proposed a family-based variant calling method, FamSeq, integrating Mendelian transmission information with de novo mutation and sequencing data to improve the variant calling accuracy. We investigated the factors impacting the improvement of family-based variant calling in simulation data and validated it in real sequencing data. In both simulation and real data, FamSeq works better than the single individual based method.

In FamSeq, we implemented four different methods for the Mendelian genetic model to accommodate variations in data complexity. We parallelized the Bayesian network algorithm on an NVIDIA graphics processing unit to make the algorithm 10 times faster for relatively large families. Our simulation shows that Elston-Stewart algorithm performs the best when there is no loop in the pedigree. If there are loops, we recommend the Bayesian network method, which provides exact answers.

The next generation sequencing technology has been developed over ten years. Many different sequencing platforms have been created to generate the sequencing data. Although all these platforms have their own strengths and weaknesses, people usually focus on one latest platform. Here we propose a method based on Bayesian hierarchical model to combine the sequencing data from multiple platforms. Our method was applied to both the simulation and real

data. The result showed that our method reduced the variant calling error rate comparing to single platform method.

Besides the application of Mendelian transmission in sequencing data analysis, we also use it to estimate the TP53 mutation carrier probability for Li-Fraumeni syndrome (LFS). LFS is an autosomal dominant hereditary disorder. People with LFS have high risk of developing early onset cancers. We proposed LFSpro that is built on a Mendelian model and estimates *TP53* mutation probability, incorporating de novo mutation rates. With independent validation data from 765 families, we compared estimations using LFSpro versus classic LFS and Chompret clinical criteria. LFSpro outperformed Chompret and classic criteria in the pediatric sarcoma cohort and was comparable to Chompret criteria in the adult sarcoma cohort.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATONS

CPU	Central Processing Unit
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GPU	Graphic Processing Unit
LFS	Li-Fraumeni Syndrome
LLR	Log10 Likelihood Ratio
MAF	Minor Allele Frequency
MCMC	Markov chain Mote Carlo
NR	No Call Rate
TS	Target Sequencing
WGS	Whole Genome Sequencing
WT	Wilms Tumor

1. Introduction

1.1 Background

Next generation sequencing technology has been widely used in recent 10 years in different researching area[1-5]. Since the first next generation sequencing technology, Roche/454 system, was developed in 2005[6], many different kinds of sequencing technology have been developed, including ABI/SOLiD[7], Illumina/Solexa[8,9], Life Technology Ion Torrent[10,11] and Pacific Biosciences PacBio[12]. A lot of next generation sequencing data is generated from these platforms.

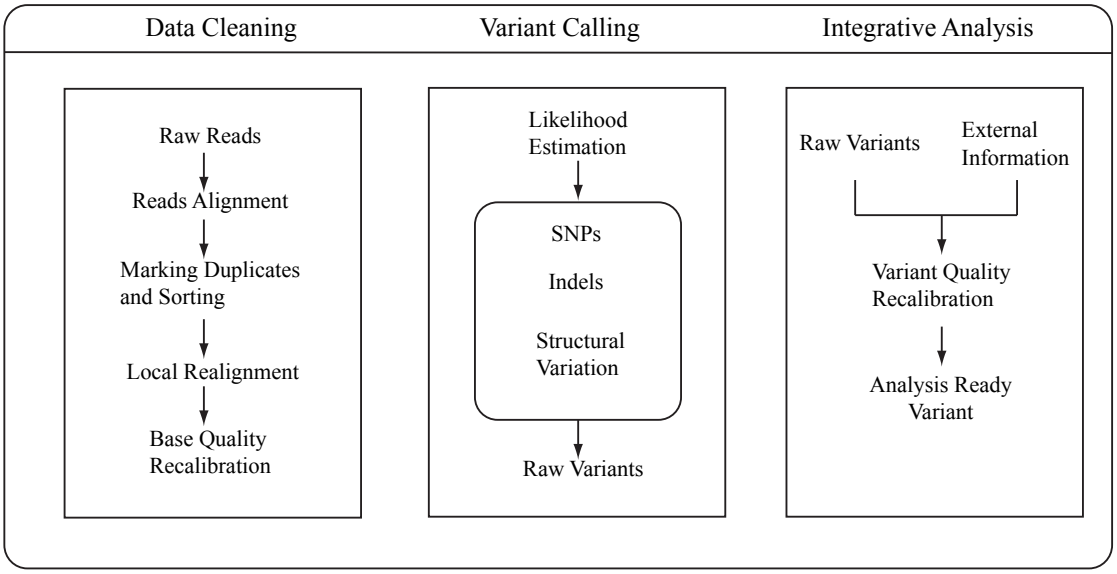


Figure 1.1: Illustration of variant calling process of DNA sequencing data.

Although the methods for each platform to generate the sequencing data are quite different, all of them can produce a file including the sequencing reads information, such as a BAM file[13]. The analysis of sequencing data usually starts from these sequencing reads information files. There are two kinds of sequencing data, DNA sequencing data and RNA sequencing data. The processing of these two kinds of data is different. In this dissertation, we focus on variant calling process of the DNA sequencing data (Figure 1.1). In processing DNA sequencing data[14-16], the reads are first mapped to a reference sequence. We search and mark the duplicated reads that are

usually removed from the following analysis to reduce the error rate. Indel realignment and base recalibration are followed to improve mapping accuracy and mapping quality calculation. After data cleanup, different methods are used to call the genotypes and indels. The called genotypes and indels are using in the evaluation step, such as phasing, association study and disease gene identification.

In the second step of DNA sequencing data analysis, many different method are developing to calling the variant position including short oligonucleotide analysis package 2(SOAP2)[17], sequencing alignment/map tools (Samtools)[13], and genome analysis toolkit (GATK)[14,15]. These methods only consider the individuals are independent or assume they are in the same population without close relationship. When the individuals are from the same family, simple filtering the variants that are not followed the Mendelian transmission are adopted to reduce the false positive rate[18]. However, this method does not reduce the false negative rate and loses all the de novo mutations. A few studies showed that it reduced both false positive and false negative rate of variant calling by incorporating the family information[19-21]. We proposed a family-based variant calling method, FamSeq, integrating Mendelian transmission information with de novo mutation and sequencing data to improve the variant calling accuracy. We investigated the factors impacting the improvement of family-based variant calling in simulation data, including family size, pedigree structure, minor allele frequency and coverage depth of sequencing to guide the design, analysis and interpretation of family-based sequencing analysis.

There are three kinds of methods to implement family information into variant calling: Bayesian network algorithm[22], Elston-Stewart algorithm[23] and Markov chain Monte Carlo algorithm[24,25]. The Markov chain Monte Carlo method allows for uncertainty in the minor allele frequency estimation. The accuracy of Markov chain Monte Carlo method depends on the iteration times. It takes a long time for Markov chain Monte Carlo method to converge and the result is an approximation not an exact value. The Elston-Stewart algorithm has high speed and

exact result when there are no loops in the pedigree. However, when there are loops in the pedigree, looping-cutting method is required to cut the loops in the pedigree, which will increase the computing time substantially and give an approximate result that is not very close to the real one in some situation[26], or adopting the method proposed by Cannings et al. that is very hard to implemented and requires much larger computing memory[27]. We can get the exact result from Bayesian network algorithm for pedigree with or without loops. However, the computing time for Bayesian network algorithm increases exponentially. It takes several days to processing the whole genome sequencing data for a family of over 10 individuals. We can hardly use this method when the family size is over 12. Most part of computing tasks in Bayesian network are homogeneous. It is easy to parallel these tasks into a GPU, which has hundreds of computing cores to calculate these tasks at the same time. After we implemented these algorithms into FamSeq, FamSeq were applied to pedigree sequencing data with different family structure and size that varied from 7 to 12 to show the different performance of the four methods.

The next generation sequencing technology has been developed about ten years[28]. Many different sequencing platforms have been created to generate the sequencing data[28-30]. All these platforms have their own strengths and weaknesses. The Illumina sequencing platform has high accuracy with short length of reads which cannot be used to get haplotype directly[8,9,31], while the reads in PacBio platform could be over 1000bp to show the haplotype with high error rate[12,31]. The error type for each platform is different[29,31]. For a single platform, we may always make the same error even we control the data quality carefully and increase the read coverage. People usually focus on one platform without considering combing multiple platforms together to give a more reliable result. In the past 10 years, some platforms have faded away, such as Roach/454 platform, there are many data generated by them remaining. Some individuals have been sequenced with many different platforms[32], only the latest data is considered in most situation. It will save a lot of money if we can combine the data from these old platforms to generate the

result with comparable accuracy rate as the current platforms. Here we propose a method based on Bayesian hierarchical model to combine the sequencing data from multiple platforms to give a more accurate variant calling result than from a single platform. In simulation, we investigated the factors that might influence the accuracy of the model, including sample size, error rate, coverage depth, correlation between features and likelihood to show the guidelines for incorporating the multiple platforms together. We also collected the sequencing data of one sample from the two most popular sequencing platforms: MiSeq platform and Ion Torrent platform, to validate our method.

Besides the application of Mendelian transmission in sequencing data analysis, we also use it in the risk prediction of Li-Fraumeni syndrome. Li-Fraumeni syndrome is an autosomal dominant hereditary disorder. People with LFS have high risk of developing early onset cancers[33,34]. For women the lifetime cancer risk is almost 100% because of breast cancer[35]. The main cause of Li-Fraumeni syndrome is germline mutation in *TP53* gene[36]. There are two commonly used clinical criteria: classic Li-Fraumeni syndrome criteria[37] and Chompret criteria[38,39]. However, these two criteria only focus on the family members with disease without considering the healthy members. Much of the family information is discounted. In addition, the two criteria can only applied to the individual with cancers. It cannot estimate the probability of *TP53* mutation for healthy individual. Given the cancer spectrum and onset in LFS and limitations of the clinical criteria, accurate identification of candidates for prospective *TP53* mutation testing has been difficult. A more efficient prediction tool is needed for LFS identification, management and screening, which should ultimately decrease mortality. We proposed LFSpro that is built on a Mendelian risk prediction model[40] and estimates *TP53* mutation probability through the Elston-Stewart algorithm[23], incorporating de novo mutation rates. We compared estimation using LFSpro ersus classic Li-Fraumeni syndrome criteria and Chompret criteria with independent validation data from 765 families (19,530 individuals in the United States [pediatric-onset sarcoma] and

Australia [adult-onset sarcoma]). LFSpro outperformed Chompret and classic criteria in the pediatric sarcoma cohort and was comparable to Chompret criteria in the adult sarcoma cohort. As the de novo mutation plays an important role in Li-Fraumeni syndrome[41], we change the de novo mutation rate in the model to process the sensitivity analysis for the two data set. We developed and validated a clinically accessible tool that incorporates de novo mutation rates to accurately estimate *TP53* mutation carriers. Family history of cancer evolves over time. LFSpro is sensitive to mutation carriers in families newly presenting in high-risk clinics, as well as those that we have followed for years. It is more broadly applicable than the clinical criteria.

1.2 Dissertation Organization

This dissertation focuses on two parts: germline mutation detection (variant calling) in next generation DNA sequencing data and *TP53* mutation carrier probability estimation in families with Li-Fraumeni syndrome. In the first part, we developed two different methods. In the first method, we incorporated the pedigree information with Mendelian model to improve the variant calling accuracy. The second method is very different from the first one. We built a Bayesian hierarchical model combining the data from multiple platforms to reduce the error rate of germline mutation detection. In the second part, the family cancer history data instead of the next generation sequencing data is used to estimate the *TP53* mutation carrier probability with Mendelian risk prediction model. Although the method used in the second part is similar to the first method in the first part, we decided to organize the dissertation according to the data type and purpose of the study for better fluency.

Chapter 2 and 3 address the variant calling method with pedigree information. In Chapter 2, we introduced method of using family information to improve the variant calling accuracy. We investigated the factors that influence the method in simulation, including family size, pedigree

structure, minor allele frequency and coverage depth of sequencing to guide the design, analysis and interpretation of family-based sequencing analysis. We also applied FamSeq in the whole-genome and target sequencing data analysis from 28 families. Both false positive rate and false negative rate are decreased using FamSeq comparing to the single individual based method. In Chapter 3, we described the details of four different methods implemented in FamSeq and compared these methods with different family size and structure. We focused on paralleling Bayesian network algorithm for family with loops when the pedigree size is relatively large, about 10 to 12 individuals in the family.

In Chapter 4, we introduced the characteristics of different next generation sequencing platforms. Then we proposed a novel method with Bayesian hierarchical model to incorporate the data from multiple platforms. In simulation, we investigated the factors that might influence the accuracy of the model, including sample size, error rate, coverage depth, correlation between features and likelihood. Then we applied our method to next generation sequencing data of a 1000 genome project[42] sample NA12878 from two platforms: MiSeq platform and Ion Torrent platform. The result showed that our method reduced the variant calling error rate comparing to single platform method when the training sample size is not very small.

TP53 mutation carrier probability estimation in families with Li-Fraumeni syndrome was introduced in Chapter 5. We gave an introduction about Li-Fraumeni syndrome at first. Then we described the Mendelian risk prediction model and our adaption of de novo mutation in the model. Our method was illustrated with some hypothetical families. We also compared our method with two clinical criteria with two real data sets.

Finally, we concluded the dissertation discussed the future research in Chapter 6.

2. Germline Mutation Detection Using Family-based Sequencing Data

This chapter is based upon the journal paper: Peng, G., Fan, Y., Palculict, T. B., Shen, P., Ruteshouser, E. C., Chi, A. K., Davis, R. W., Huff, V., Scharfe, C. & Wang, W. "Rare variant detection using family-based sequencing analysis". Proceedings of the National Academy of Sciences of the United States of America 110, 3985-3990, doi:10.1073/pnas.1222158110 (2013). According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.

2.1 Introduction

Challenges in using whole-genome sequencing (WGS) data for identifying rare DNA variants responsible for heritable dis- ease include high false negative (FN) rates and the need to minimize the number of false positive (FP) variants to reduce the total number of variants for subsequent validation. Family-based sequencing designs have been applied to gene discovery for several diseases [2,43,44]. Methods for calling variant positions in DNA sequence data include short oligonucleotide analysis package 2 (SOAP2)[17], sequence alignment/map tools (Samtools)[13], and genome analysis toolkit (GATK)[14,15]. When assessing data from related individuals, simple filtering can remove variants that do not conform to Mendelian transmission expectations, thereby reducing FPs. However, this approach does not reduce the frequency of FNs, and it removes all de novo mutations[18]. There are approaches that borrow information across neighboring variants through family-based haplotype phasing[19,45]. As an orthogonal approach, integrating Mendelian inheritance and raw data of family members at a single position can reduce both FPs and FNs and has been implemented in variant calling tools for family trios[19,20]. In recent simulation studies, Li et al.[21] showed that joint variant calling in data from extended families will further improve detection of Mendelian variants and reduce FP de novo mutations. However, a limitation of their study is that simulations cannot incorporate many sources of variations that are

observed across millions of positions within a sample and across samples and families. Their study did not evaluate variant positions with base coverage greater than 40×, nor compare data generated by targeted versus WGS; they did not compare the performance of family-based calling for founder versus non- founder or for common versus rare variants. Thus, evaluation of a family-integrated method under real settings across many individuals is required to prevent underestimation of its actual contributions to identifying rare variants in families.

In addition, accurate variant calling and decreased FN rates (FNRs) enable the development of more efficacious and efficient studies that incorporate decisions about study design (who should be sequenced first in a large family and at what sequencing coverage), data analysis (setting up unknown parameters), and results interpretation (distinguishing true variants from FPs for functional association). Knowledge of factors contributing to accurate variant calling in families facilitates these decisions.

We have developed a family-based variant calling program (Family-Based Sequencing Program, FamSeq) that provides a confidence measure for variant calls using data from all family members and builds on Bayesian networks and the Markov chain Monte Carlo (MCMC) algorithm[25]. We used this method to perform simulation studies and analyze sequencing data from 28 families [one from the HapMap (Haplotype Map) project, one with Wilms tumor (WT), and 26 with mitochondrial neuro-developmental disorders] presenting various pedigree structures. Compared with variant calling using a single-individual-based method or using only a family trio (14%), FamSeq reduced FN variant calls by 33% in the extended HapMap pedigree. In the analysis of actual data from one family, FamSeq resulted in the identification of an additional ~300 to ~1,200 new variant positions in WGS that otherwise would have been undetected using the Single method.

Our goal is to provide a method for rare variant detection and to guide the design and analysis of a family-based sequencing study. We describe and validate our method, then describe

simulations and analyses of 92 samples from 28 families. We present a comprehensive investigation of factors that may determine the improvements achievable by our family-integrated method on a per-person and per-position basis. We also illustrate the effect of annotating Mendelian variants in studying either dominant or recessive traits.

2.2 Method

2.2.1 Individual-Based (Single) Method. Let D_i denote the raw sequencing measurements—that is, read counts, read quality, and mapping quality— and G_i denote the genotype for sample i. For a family of n members, we use **D** to denote a vector $\{D_1;\ D_2;\ \ldots;\ D_n\}$ and **G** a vector $\{G_1;\ G_2;\ \ldots;\ G_n\}$. GATK provides likelihood estimates $\Pr(D_i|G_i)$ in VCF files. By following Bayes' rule, the genotype posterior probabilities are calculated as $\Pr(D_i|G_i) \propto \Pr(D_i|G_i)\Pr(G_i)$, where the prior $\Pr(G_i)$ is the expected genotype frequency in the population, and is calculated based on the MAF and Hardy-Weinberg equilibrium.

2.2.2 Family-Based Method (FamSeq). Let **P** denote the pedigree structure. We calculate a genotype posterior probability $\Pr(G_i|\mathbf{P};\ \mathbf{D})$, which incorporates the actual pedigree structure and raw sequencing data and accommodates de novo mutations. We use three methods to compute $\Pr(G_i|\mathbf{P};\mathbf{D})$: a Bayesian network[46], and Elston-Stewart and MCMC algorithms. FamSeq provides an updated VCF file that includes the family-based variant calling results and posterior probabilities.

2.2.3 Full Simulation. To evaluate the contribution of family-based analysis to improving variant calling accuracy, we simulated all genotype configurations for a family trio and a family quartet.

First, we simulated the two parents' genotypes at known MAFs. Then based on the parents' genotypes, we simulated the children's genotypes according to Mendelian transmission and allowing for de novo mutations. We then simulated likelihoods from the density functions of bivariate normal distributions for each genotype [where μ is equal to (1.7,1.7), (0,0), and (−1.7, −1.7) for the three genotypes, and Σ is equal to (0.3,0.15; 0.15,0.3) for homozygous and (0.45,0.225; 0.225,0.45) for heterozygous genotypes]. We performed this simulation using two settings: (i) MAF = 0.2, m = 1e-3, 10 million iterations, and (ii) MAF = 0.01, m = 1e−5, 100 million iterations. We then calculated likelihoods $\Pr(D_i|G_i)$ at the true μ's and Σ* at (0.1,0.05; 0.05,0.1). We used different variance matrices to account for additional technical effects that cannot be observed and estimated.

2.2.4 Targeted Simulation. To evaluate the impact of pedigree size, structure, and MAF, we fixed the genotype configurations to extensions of (father = 0, mother = 0, child = 0) and (father = 0, mother = 1, child = 1), and simulated the raw intensity data based on bivariate normal distributions for each genotype. For pedigree size and structure, we considered the following scenarios: size = 2, parent–child pair; size = 3, family trio; size = 4, nuclear family with two children; size = 5, nuclear family with three children; size = 6, nuclear family with four children, or nuclear family with three children and one grandparent; and size = 7, two grandparents, two parents, and three children. For MAF, we considered 0.5, 0.2, 0.1, 0.05, 0.01, 1e−3, 1e−4, and 1e−5. For each scenario, we repeated the simulation for 1 million times. There are two different groups, one for computing the FPRs and one for computing the FNRs. We computed the FPRs in individuals carrying homozygous references and computed the FNRs in individuals carrying heterozygous variants. When there is only one parent/grandparent, the parent/ grandparent is 1. When there are two parents/grandparents, one parent/ grandparent is 1 and the other is 0.

2.2.5 Whole Genome Sequencing Data. We downloaded the WGS data of NA12891, NA12892, and NA12878 from the 1000 Genomes Project (www.1000genomes.org), and NA12877 and NA12882 sequence data from the Sequence Read Archive (www.ncbi. nlm.nih.gov/Traces/sra). The data for the first three samples are generated using GAII and those for the other two are generated using HiSeq2000. We downloaded the consistent calls from the HapMap Phase II and III merged genotype data for NA12891, NA12892, and NA12878 (http://hapmap.ncbi. nlm.nih.gov). Using HiSeq2000, we also conducted WGS of five samples (one trio plus two distant relatives) from a large pedigree that presented with Wilms tumor. Our analysis focused on a 5.6 MB linkage region in Chr19q. In WGS analysis, we filtered out bases that are noted as simple repeats or segmental duplications by the University of California, Santa Cruz human genome assembly hg19, and those with total allelic counts less than 10.

Table 2.1 Summary of target sequencing data from 26 families. (*Table reprint from Peng, G. et al PNAS, 2013*)

Family	Member	Sequncing Technology	Sequecing Platform	Average Coverage
Leigh	Child	Targeted Sequencing	HiSeq2000	500
Leigh	Father	Targeted Sequencing	HiSeq2000	385
Leigh	Mother	Targeted Sequencing	HiSeq2000	426
Msy	Child	Targeted Sequencing	HiSeq2000	541
Msy	Mother	Targeted Sequencing	HiSeq2000	445
MTF01	Child 1	Targeted Sequencing	HiSeq2000	400
MTF01	Child 2	Targeted Sequencing	HiSeq2000	345
MTF01	Father	Targeted Sequencing	HiSeq2000	355
MTF01	Paternal grandmother	Targeted Sequencing	HiSeq2000	445
MTF01	Mother	Targeted Sequencing	HiSeq2000	349
MTF02	Child 1	Targeted Sequencing	HiSeq2000	392
MTF02	Child 2	Targeted Sequencing	HiSeq2000	332
MTF02	Father	Targeted Sequencing	HiSeq2000	327
MTF02	Mother	Targeted Sequencing	HiSeq2000	203
MTF03	Child	Targeted Sequencing	HiSeq2000	289
MTF03	Father	Targeted Sequencing	HiSeq2000	334
MTF03	Mother	Targeted Sequencing	HiSeq2000	281
MTF04-a	Child	Targeted Sequencing	HiSeq2000	286
MTF04-a	Father	Targeted Sequencing	HiSeq2000	358
MTF04-a	Mother	Targeted Sequencing	HiSeq2000	262
MTF04-b	Child	Targeted Sequencing	HiSeq2000	286
MTF04-b	Father	Targeted Sequencing	HiSeq2000	358
MTF04-b	Mother	Targeted Sequencing	HiSeq2000	262
MTF04-b	Maternal grandfather	Targeted Sequencing	HiSeq2000	243
MTF04-b	Maternal grandmother	Targeted Sequencing	HiSeq2000	302
MTF04-c	Child	Targeted Sequencing	HiSeq2000	286
MTF04-c	Father	Targeted Sequencing	HiSeq2000	358
MTF04-c	Mother	Targeted Sequencing	HiSeq2000	262
MTF04-c	Paternal grandfather	Targeted Sequencing	HiSeq2000	306
MTF04-c	Paternal grandmother	Targeted Sequencing	HiSeq2000	225
MTF04-d	Child	Targeted Sequencing	HiSeq2000	286
MTF04-d	Father	Targeted Sequencing	HiSeq2000	358
MTF04-d	Mother	Targeted Sequencing	HiSeq2000	262
MTF04-d	Maternal grandfather	Targeted Sequencing	HiSeq2000	243
MTF04-d	Maternal grandmother	Targeted Sequencing	HiSeq2000	302
MTF04-d	Paternal grandfather	Targeted Sequencing	HiSeq2000	306
MTF04-d	Paternal grandmother	Targeted Sequencing	HiSeq2000	225
MTF05	Child	Targeted Sequencing	HiSeq2000	418
MTF05	Mother	Targeted Sequencing	HiSeq2000	242
MTF06	Child 1	Targeted Sequencing	HiSeq2000	492
MTF06	Child 2	Targeted Sequencing	HiSeq2000	686
MTF06	Father	Targeted Sequencing	HiSeq2000	643
MTF06	Mother	Targeted Sequencing	HiSeq2000	559
MTF07	Child 1	Targeted Sequencing	HiSeq2000	643
MTF07	Child 2	Targeted Sequencing	HiSeq2000	563
MTF07	Child 3	Targeted Sequencing	HiSeq2000	682
MTF07	Father	Targeted Sequencing	HiSeq2000	461
MTF07	Mother	Targeted Sequencing	HiSeq2000	582
MTF08	Child 1	Targeted Sequencing	HiSeq2000	398

MTF08	Child 2	Targeted Sequencing	HiSeq2000	853
MTF08	Father	Targeted Sequencing	HiSeq2000	692
MTF08	Mother	Targeted Sequencing	HiSeq2000	530
MTF09	Child 1	Targeted Sequencing	HiSeq2000	899
MTF09	Child 2	Targeted Sequencing	HiSeq2000	643
MTF09	Father	Targeted Sequencing	HiSeq2000	725
MTF09	Mother	Targeted Sequencing	HiSeq2000	612
MTF10	Child 1	Targeted Sequencing	HiSeq2000	712
MTF10	Child 2	Targeted Sequencing	HiSeq2000	635
MTF10	Father	Targeted Sequencing	HiSeq2000	575
MTF10	Mother	Targeted Sequencing	HiSeq2000	428
MTF11	Child 1	Targeted Sequencing	HiSeq2000	354
MTF11	Child 2	Targeted Sequencing	HiSeq2000	221
MTF11	Father	Targeted Sequencing	HiSeq2000	443
MTF11	Mother	Targeted Sequencing	HiSeq2000	583
MTF12	Child	Targeted Sequencing	HiSeq2000	1086
MTF12	Father	Targeted Sequencing	HiSeq2000	364
MTF12	Mother	Targeted Sequencing	HiSeq2000	790
MTF13	Child 1	Targeted Sequencing	HiSeq2000	637
MTF13	Child 2	Targeted Sequencing	HiSeq2000	646
MTF13	Mother	Targeted Sequencing	HiSeq2000	1232
MTF14	Child	Targeted Sequencing	HiSeq2000	542
MTF14	Father	Targeted Sequencing	HiSeq2000	708
MTF14	Mother	Targeted Sequencing	HiSeq2000	543
MTF15	Child	Targeted Sequencing	HiSeq2000	649
MTF15	Father	Targeted Sequencing	HiSeq2000	519
MTF15	Mother	Targeted Sequencing	HiSeq2000	592
MTF16	Child	Targeted Sequencing	HiSeq2000	590
MTF16	Father	Targeted Sequencing	HiSeq2000	429
MTF16	Mother	Targeted Sequencing	HiSeq2000	621
MTF17	Child	Targeted Sequencing	HiSeq2000	350
MTF17	Father	Targeted Sequencing	HiSeq2000	412
MTF17	Mother	Targeted Sequencing	HiSeq2000	895
Myo	Child	Targeted Sequencing	HiSeq2000	1020
Myo	Father	Targeted Sequencing	HiSeq2000	702
OTC	Mother	Targeted Sequencing	HiSeq2000	1072
OTC	Child	Targeted Sequencing	HiSeq2000	738
PDH	Child	Targeted Sequencing	HiSeq2000	297
PDH	Father	Targeted Sequencing	HiSeq2000	468
PDH	Mother	Targeted Sequencing	HiSeq2000	478
RCC4	Child 1	Targeted Sequencing	HiSeq2000	886
RCC4	Child 2	Targeted Sequencing	HiSeq2000	505
RCC4	Child 3	Targeted Sequencing	HiSeq2000	538
HapMap	NA12882 (Child)	Whole Genome Sequecing	HiSeq2000	56
HapMap	NA12877 (Father)	Whole Genome Sequecing	HiSeq2000	56
HapMap	NA12878 (Mother)	Whole Genome Sequecing	GAII	41
HapMap	NA12891 (Maternal grandfather)	Whole Genome Sequecing	GAII	30
HapMap	NA12892 (Maternal grandmother)	Whole Genome Sequecing	GAII	25
Wilms Tumor	Mother	Whole Genome Sequecing	HiSeq2000	35
Wilms Tumor	Father	Whole Genome Sequecing	HiSeq2000	36
Wilms Tumor	Child	Whole Genome Sequecing	HiSeq2000	31

2.2.6 Target Sequencing Data. We performed DNA sequence capture of 524 nuclear-encoded mitochondrial genes[47] from 92 samples in 26 families (Table 2.1) and multiplex-sequenced all capture libraries using Illumina HiSeq2000, except for MTF04-b, c, and d, which were sequenced using MiSeq. Our analysis focused on a 762 KB region of autosomes.

2.2.7 Methods Evaluation. We calculated the FNR as the rate of reference or no calls for a true variant genotype, and the FPR as the rate of variant calls for a reference genotype. We defined concordance as having genotype calls that are identical to the HapMap truth, and discordance as having genotype calls that are different from the HapMap truth.

2.2.8 Unique Variant Calls of FamSeq and the Single Method. We combined the posterior probability of the heterozygous variant with that of the homozygous variant to evaluate the number of unique variant calls added by either FamSeq or the Single method. We designated a call as a unique variant call if the method of interest changes the calls from those of the alternative method in the following ways: reference to variant, no call to variant, or reference to no call.

2.3 Results

2.3.1 FamSeq. Figure 2.1 describes the FamSeq framework. This method provides a confidence measure for genotype calls, which is a posterior probability $\Pr(G_i|\mathbf{D}, \mathbf{P})$. Here G denotes genotype, i denotes an individual, **P** denotes pedigree structure, and **D** is a vector that denotes sequencing data, including read counts, base quality, and mapping quality, for all n family members (individual i and relatives). Incorporating data from family members, $\Pr(G_i|\mathbf{D}, \mathbf{P})$ allows for accurate variant calling when the data from person i are not informative, perhaps due to a weak signal-to- noise-ratio, by borrowing strength from all relatives (Figure. 1B). Here we measure the signal-to-noise-ratio using the ratios of the likelihood estimates ($\Pr(D_i|G_i)$) for the two most likely genotypes. FamSeq has included probabilities of de novo mutations. It allows for variable pedigree size ($n > 3$) and structure. In addition to using the Elston-Stewart algorithm as in Li et al.[21] for pedigree analysis, we implemented two unique approaches, Bayesian network

and MCMC. The Bayesian network approach directly calculates joint probabilities for each combination of genotypes of all family members and allows for analytic calculation in pedigrees with marriage loops and/or consanguinity, as long as they form directed acyclic graphs. This method gives faster computation than the Elston-Stewart algorithm with or without loops in pedigrees of size less than 7. The MCMC method allows for the use of continuous probability density functions as priors on the genotype probability Pr(G_i) and likelihood Pr($D_i|G_i$), instead of designating the point mass a priori.

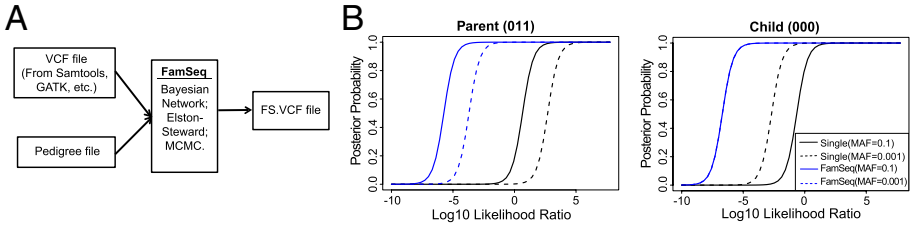


Figure 2.1 Illustration of variant calling using FamSeq. (A) FamSeq variant calling framework. (B) Two examples in a family trio. We use 0 to denote reference and 1 to denote heterozygous variant. The order of genotypes presented in the parentheses is father, mother, and child. In both cases, FamSeq gives the child a high posterior probability (>0.9) for the true genotype even when the child has a relatively low log10 LLR. This is done in FamSeq by borrowing strength from data of the parents. (*Figure reprint from Peng, G. et al PNAS, 2013*)

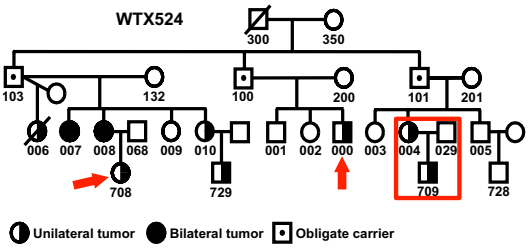


Figure 2.2 A family with Wilms tumor for genomic sequencing of 19q13-linked region. The family trio is affected mother (WTX524-004), unaffected father (WTX524-029), and affected child (WTX524-709). Two affected distant rela- tives (WTX524-708, WTX524-000) are also sequenced. (*Figure reprint from Peng, G. et al PNAS, 2013*)

2.3.2 Motivating Example: Family with Inherited Wilms Tumor. Familial transmission of predisposition to WT, a childhood kidney tumor, is consistent with an autosomal dominant mutation with incomplete penetrance. Two predisposition genes have been localized by genetic linkage studies, but neither gene has been identified[48]. We generated WGS data for five members of a large WT family and focused on a 5.6 MB linkage region on chr19q. Because genetic linkage has been previously demonstrated, the two distantly related individuals WTX524-708 and WTX524-000 are expected to share the same Mendelian variants as individuals WTX524-709 and WTX524-004 in the trio (Figure. 2.2). Comparing FamSeq with GATK (with variant recalibration), we found that both methods identified 4,920 positions with variant calls in all four affected family members. FamSeq identified an additional 132 positions and GATK uniquely identified one position.

Table 2.2 Sanger verification on FamSeq-unique positions. These positions were found to have variant genotypes in all four affected relatives and a reference genotype in the unaffected spouse. There are 38 positions in total and 32 positions were confirmed by Sanger sequencing (84%, p-value<0.0001). The false positive positions are listed at the bottom of the table. (*Table reprint from Peng, G. et al PNAS, 2013*)

CHROM	POS	ID	REF	ALT	VQSR_FILTER	Sanger	SampleID	SNPLocation	Reported	ReportedAltFreq
19	52077035	.	T	A	-	T/A	524-709	UTR3	-	-
19	52095033	rs4802819	C	T	-	T, C/T	524-004, 524-709	downstream	-	-
19	**52190539**	**.**	**C**	**A**	**-**	**C/A**	**524-004**	**ncR-_exonic**	**-**	-
19	53748959	rs4803072	C	T	-	T	524-004, 524-709	intronic	-	-
19	52079404	rs74326269	T	G	TruthSensitivityTranche99.90to100.00	G, G/T	524-004, 524-709	intronic	-	-
19	**52910901**	**.**	**T**	**G**	**TruthSensitivityTranche99.90to100.00**	**T**	**524-029**	**intronic**	**-**	-
19	52924044	rs73934458	A	G	TruthSensitivityTranche99.90to100.00	A/G	524-004, 524-709	intergenic	-	-
19	53350680	rs75724618	A	G	TruthSensitivityTranche99.90to100.00	G/A	524-004, 524-709	intronic	-	-
19	53353962	rs111436395	C	A	TruthSensitivityTranche99.90to100.00	A/C	524-004, 524-709	intronic	-	-
19	53518336	rs139323953	G	A	TruthSensitivityTranche99.90to100.00	A/G	524-004, 524-709	intergenic	-	-
19	52079498	rs147557699	G	A	TruthSensitivityTranche99.90to100.00	G/A	524-004, 524-709	intronic	1000g2010nov_all	0.003
19	52068087	rs78071293	T	C	TruthSensitivityTranche99.90to100.00	T/C	524-004, 524-709	intergenic	1000g2010nov_all	0.01
19	52179651	rs113567777	C	G	TruthSensitivityTranche99.90to100.00	C/G	524-004, 524-709	intergenic	1000g2010nov_all	0.01
19	51979249	rs10414851	T	C	TruthSensitivityTranche99.90to100.00	T/C	524-004, 524-709	upstream	1000g2010nov_all	0.014
19	53688038	rs117654539	C	T	TruthSensitivityTranche99.90to100.00	T/C	524-004, 524-709	intronic	1000g2010nov_all	0.016
19	52924984	rs28528760	A	C	TruthSensitivityTranche99.90to100.00	A/C	524-004, 524-709	intergenic	1000g2010nov_all	0.025
19	53692832	rs12232823	C	A	TruthSensitivityTranche99.90to100.00	A	524-004, 524-709	intronic	1000g2010nov_all	0.047
19	53688095	rs7259002	C	T	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intronic	1000g2010nov_all	0.052
19	53629915	rs78806018	T	C	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intronic	1000g2010nov_all	0.06
19	53418471	rs4803021	A	G	TruthSensitivityTranche99.90to100.00	A/G	524-004, 524-709	exonic (syn)	1000g2010nov_all	0.078
19	53933358	rs67470095	G	A	-	A/G	524-004, 524-709	intergenic	1000g2010nov_all	0.094
19	53693049	rs9630864	A	G	TruthSensitivityTranche99.90to100.00	G	524-004, 524-709	intronic	1000g2010nov_all	0.1
19	53265582	rs58224318	C	T	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intergenic	1000g2010nov_all	0.128
19	54016382	rs74178334	G	A	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intergenic	1000g2010nov_all	0.144
19	52925067	rs28375755	C	A	TruthSensitivityTranche99.90to100.00	C/A	524-004, 524-709	intergenic	1000g2010nov_all	0.15
19	53022362	rs2868688	T	G	TruthSensitivityTranche99.90to100.00	T/G	524-004, 524-709	intergenic	1000g2010nov_all	0.179
19	53930220	rs28588000	C	T	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intronic	1000g2010nov_all	0.19
19	53350708	rs8106315	C	T	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intronic	1000g2010nov_all	0.221
19	53350604	rs6509697	C	T	TruthSensitivityTranche99.90to100.00	T/C	524-004, 524-709	intronic	1000g2010nov_all	0.43
19	53350597	rs6509696	G	A	TruthSensitivityTranche99.90to100.00	G/A	524-004, 524-709	intronic	1000g2010nov_all	0.44
19	53344118	rs10419826	C	G	TruthSensitivityTranche99.90to100.00	G/C	524-004, 524-709	exonic (non-syn)	1000g2010nov_all	0.481
19	53863983	rs4296358	C	T	TruthSensitivityTranche99.90to100.00	C/T	524-004, 524-709	intergenic	1000g2010nov_all	0.54
19	53676105	.	T	A	TruthSensitivityTranche99.90to100.00	T	524-004, 524-709	intronic	-	-
19	53773647	rs192983166	G	A	TruthSensitivityTranche99.90to100.00	G	524-004, 524-709	intergenic	-	-
19	51990997	rs12327792	A	G	TruthSensitivityTranche99.90to100.00	A	524-004, 524-709	intergenic	1000g2010nov_all	0.043
19	52924334	rs111240956	T	C	TruthSensitivityTranche99.90to100.00	T	524-004, 524-709	intergenic	1000g2010nov_all	0.005
19	52924340	rs113590683	C	G	TruthSensitivityTranche99.90to100.00	C	524-004, 524-709	intergenic	1000g2010nov_all	0.01
19	53936283	rs58354543	T	C	TruthSensitivityTranche99.90to100.00	T	524-004, 524-709	ncR-_intronic	1000g2010nov_all	0.43

2.3.3 Sanger Validation. To assess the validity of the FamSeq uniquely called variants, we performed Sanger sequencing on 57 of the 132 positions, which exist in a subregion and meet an additional requirement of presenting reference calls in the unaffected father. This four-variants-plus-one-reference filtering procedure is designed to prioritize variants potentially important for WT and was performed on both FamSeq and GATK-based calls. We obtained reliable Sanger results on 38 FamSeq-unique positions and confirmed that 32 (61 variant calls) are true (Table 2.2). Our validation rate is 61/73 = 84% (95% confidence interval: 75–92%). Among the confirmed FamSeq-unique variants, 17 (53%) are rare (not reported or at a minor allele frequency of less than 5%). Other than one position where FamSeq corrected a call from the variant by GATK to reference in the unaffected father, the FamSeq-unique positions were missed by GATK because they were (i) called as reference in one affected individual, (ii) removed during variant quality score recalibration, or (iii) had variant calls at a tranche level of 99.9– 100 or lower.

Using simulated and actual data, we identified variables that determine the possible improvements from using our family-based analysis. From here on, we compare FamSeq with the Single method based on their posterior probabilities. First, we describe the results based on simulations.

2.3.4 Genotype Configurations. FamSeq improved the accuracy in all Mendelian genotypes (15 scenarios for a family trio, Figure 2.3A) and made substantial improvements in two scenarios: (i) at positions where all family members have reference genotypes, FamSeq corrected FP calls (~30%; Figure 2.4), and (ii) at positions where a single parent and child carry heterozygous variants, FamSeq corrected FN calls (20–40%; Figure 2.4). FamSeq identified true Mendelian positions that were erroneously called as variants by the Single method, as shown by the red cells in the heatmap of Fig. 2.3A. For example, at truth = 000, FamSeq reduced discordant calls of 001; at truth = 101, again FamSeq reduced discordant calls of 001 and 102, made by the Single

method. When the de novo mutation rate is high [1 × 10−5, compared with variants with minor allele frequency (MAF) of 0.01; Figure 2.4B], FamSeq missed 34% of true de novo mutations correctly called by the Single method, suggesting possible underestimations. We made similar observations with a family quartet.

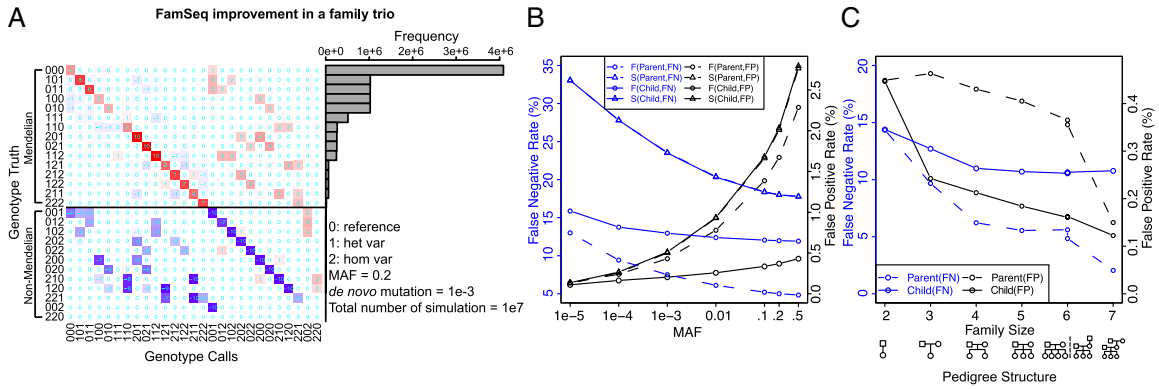
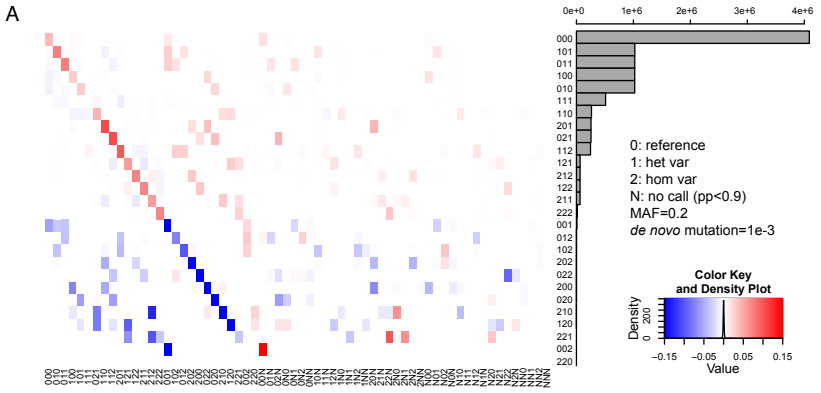
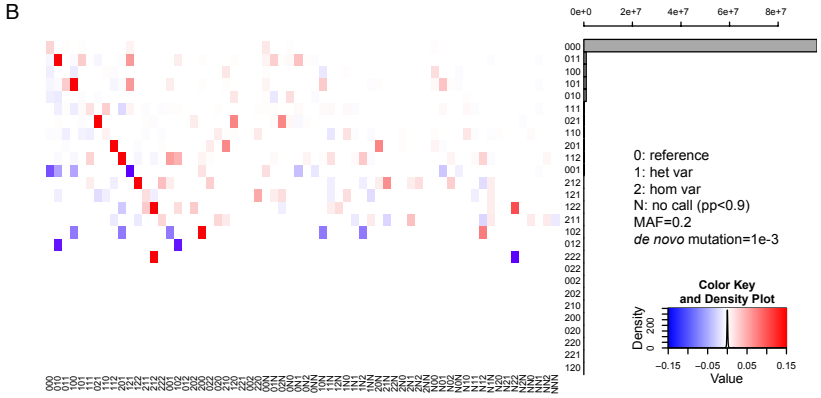


Figure 2.3 Simulation results. (A) Highlighted results from a full simulation of all possible genotype configurations of a family. Each row is the simulated genotype for the family trio (father, mother, child). Here, 0 is homozygous reference, 1 is heterozygous variant, and 2 is homozygous variant. Each heatmap entry is the percent reduction in discordance from using the Single method to using FamSeq. The values on the diagonal are equal to the sum of all other 63 values in the same row. Only 27 columns are shown. Additionally, there are 37 columns with genotypes containing "no calls." The corresponding complete results can be found in Figure 2.4. The barplot on the right presents the frequency for observing each configuration. (B) Targeted simulation to evaluate effect of MAF. F stands for FamSeq and S stands for single method. (C) Targeted simulation to evaluate effect of pedigree size and structure. (*Figure reprint from Peng, G. et al PNAS, 2013*)



Net improvement= 22%							
Configuration	000	101	011	100	010	111	Others
% improvement	29	21	21	12	12	0.4	13 to 38
% frequency	41	10	10	10	10	5	~14



Net improvement = 33%					
Configuration	000	011	101	100	010
% improvement	34	42	42	0	0
% frequency	96	1	1	1	1

Figure 2.4 Simulations for all possible genotype configurations in a family trio. In the heatmap each row is the simulated genotype for the family trio (father, mother, child). Here, 0 is homozygous reference, 1 is heterozygous variant; 2 is homozygous variant, and N is no call. Each heatmap entry is the percent reduction in discordance from using the Single method to using FamSeq. Red suggests improvement and blue suggests otherwise. The barplot on the right presents the frequency for observing each configuration. The percentage improvement out of all counts of truth for FamSeq within each configuration is shown in the following table. A) We simulated 10 million iterations for family trios at MAF=0.2 and de novo mutation rate=1e-3. B) We simulated 100 million iterations for family trios at MAF=0.01 and de novo mutation rate=1e-5. We could not observe some rare configurations. The table gives percentage improvement for FamSeq within each configuration. (*Figure reprint from Peng, G. et al PNAS, 2013*)

2.3.5 Minor Allele Frequency. The MAF parameter is used for computing prior probabilities of genotypes, Pr(G), in FamSeq and the Single method and is mostly unknown (Fig. 2.3B). Setting different values of MAF (from 10−5 to 0.5) switches the balance between the FNR and FPR in the Single method. As MAF increases, FNRs decrease and FPRs increase. With FamSeq, not only are both error rates lower at all values, but as the MAF varies, the changes in FNRs and in FPRs in the children, and changes in FNRs in the parents, are much attenuated; that is, error rates are less dependent on MAF values. Therefore, by jointly calling variants in all family members, we can set the same MAF at all base positions, for example 0.001, without compromising the detection of true variants.

2.3.6 Family Size and Pedigree Structure. Starting from a parent–child pair, FamSeq reduced both FNR and FPR when we included the second parent (family size = 2 to size = 3), and then added another sibling (size = 3, 4) (Figure. 2.3C). Interestingly, adding more children (size = 4, 5, 6) did not further reduce error rates, whereas adding the grandparents (size = 5–7) made additional reductions in both FNR and FPR. When the parental data are not available, we also observed improvements made by FamSeq in analyzing all siblings together (size = 3, FNR 23.5% vs. 13.3%, FPR 0.5% vs. 0.4%). This has important implications when prioritizing individuals from a larger pedigree to accurately and comprehensively detect rare DNA variants.

2.3.7 Contribution to Family Members. The reduction in error rates using FamSeq is membership-dependent (Fig. 2.3 B and C). FNRs are better controlled in parents than in children. FPRs are better controlled in children than in parents (founders), which reduces the cost of subsequent sequence verifications. Both reduce the FPs in calling de novo mutations in children. Accordingly, when grandparents' data are available, the FPRs in the corresponding parent

(nonfounder) decrease substantially, which improves the detection of de novo mutations in children.

2.3.8 Whole Genome Sequencing Data Analysis. We analyzed a three-generation HapMap WGS dataset of five samples. In the whole genomes of HapMap samples, FamSeq found 1,179,317, and 494 new variant positions across all samples when analyzing pedigrees g3 (grandparent trio), c3 (child trio), and a5 (all five). Within each sample, FamSeq called ~7,000 to ~32,000 more variants than the Single method. Samples with lower coverage (NA12892 at ~25×; Table 2.3) benefited most from FamSeq analysis, exhibiting a greater percentage of increased variant calls.

Table 2.3 Summary of validation results using HapMap SNPs. The results are based on three HapMap samples NA12891, NA12892 and NA12878. A) Known SNPs called consistently from HapMap Phase II and III. B) Concordance and error rates for FamSeq as compared to the Single (individual-based) method. The SNP calls for these three individuals are provided by the HapMap project and used as truth. FamSeq is run on subsets of the family, g3 and c3, and on the entire family, a5. NR = no call rate; FPR = false positive rate; FDR = false discovery rate. (*Table reprint from Peng, G. et al PNAS, 2013*)

Known HapMap Calls								
	NA12891		NA12892		NA12878			
Homozygous reference	311,453		316,888		310,907			
Homozygous alternative	276,793		280,702		276,372			
Homozygous all	588,246		597,590		587,279			
Heterozygous all	427,578		418,233		428,544			
Total	1,015,824		1,015,823		1,015,823			
Validation								
	NA12891		NA12892		NA12878			
	Single	FamSeq(g3)	Single	FamSeq(g3)	Single	FamSeq(c3)	FamSeq(g3)	FamSeq(a5)
Concordance(%)	99.79	99.81	99.70	99.76	99.90	99.91	99.90	99.91
False Positives	597	604	599	591	496	509	539	546
(FPR, %)	(0.10)	(0.10)	(0.10)	(0.10)	(0.084)	(0.087)	(0.092)	(0.093)
(FDR, %)	(0.14)	(0.14)	(0.14)	(0.14)	(0.12)	(0.12)	(0.13)	(0.13)
False Negatives	1259	1085	1887	1367	387	328	297	260
(FNR, %)	(0.30)	(0.26)	(0.44)	(0.32)	(0.091)	(0.077)	(0.069)	(0.061)
No Calls	647	497	1616	1138	232	187	207	175
(NR, %)	(0.064)	(0.049)	(0.16)	(0.11)	(0.023)	(0.018)	(0.020)	(0.017)

2.3.9 HapMap Sample Validation. In three samples (mean coverage ~25– 30×), we compared FamSeq calls with HapMap calls at ~1 million single-nucleotide polymorphism (SNP) positions[49] (Table 2.3). Homozygous genotypes are more easily identified than heterozygous variants[50]. Using known SNP data, we combined all homozygous SNP positions as true negatives and used all heterozygous SNP positions as true positives, from NA12878, NA12891, and NA12892 (~400,000 true positives for each sample). As expected, FamSeq called more positions at high confidence (7–29% fewer no call positions) and identified more true variants with percent reduction in FNs of 14–33%, and without substantially increasing the number of false discoveries (1–3%; Figure 2.5A and Table 2.3). In particular, comparing pedigrees c3 and a5, we observed a statistically significant difference in the percent reduction of FNs (15% vs. 33% in NA12878, $P < 0.0001$). This result is consistent with simulations comparing sizes of 5 and 7 in the parent (Figure 2.3C). We also observed low sensitivity to varying MAF values in variant calling when using FamSeq (Figure 2.6). In contrast to the simulations, we did not observe a decrease in FPs in the child (NA12878 in g3). One explanation is we derived the input likelihood estimates from GATK, which may aggressively filter out FPs, but at a price of missing some true positives.

This validation was performed at HapMap SNP positions, including all common SNPs whose known genotypes may have been used for calibration by GATK. Additionally, most of these SNPs (98%) are located in the noncoding region. Therefore, we look for larger improvements from using FamSeq for finding rare DNA variants at sequence sites where variant calling in the Single method has not been optimized.

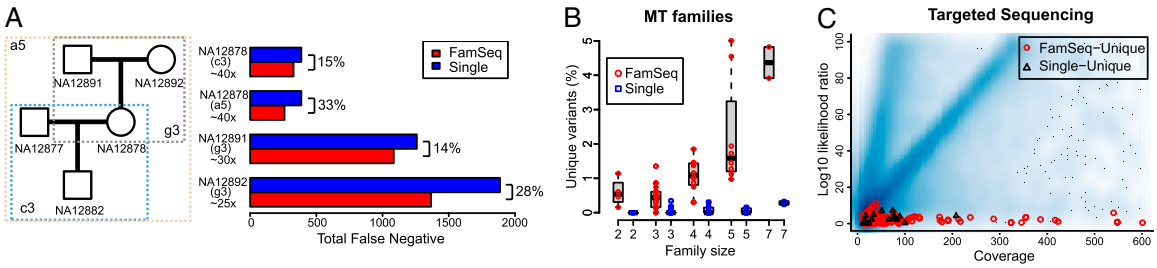


Figure 2.5 Analysis of sequencing data in extended pedigrees. (A) HapMap SNP validation (Table 2.3). (B) FamSeq-unique variants found in 45 people (parents) in 25 families affected with mitochondrial disorders. (C) Coverage versus LLR in TS samples. All positions called concordantly by the Single method and FamSeq are shown in the background as a smoothed scatterplot. Red circles represent FamSeq-unique variants; black triangles represent Single-unique variants. (*Figure reprint from Peng, G. et al PNAS, 2013*)

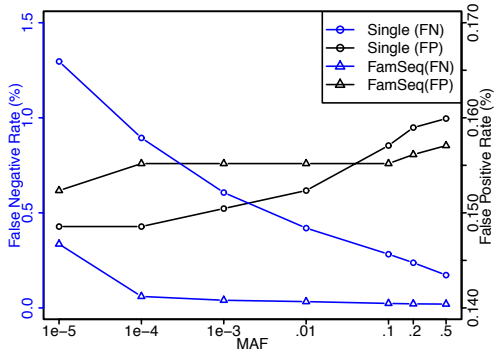


Figure 2.6 Effect of MAF values on variant calling on HapMap SNPs. The x-axis is plotted in a log10 scale. The false negative rate is calculated as the rate of reference or no calls for a variant genotype, in NA12892, who is a parent, and within positions presenting a true genotype configuration of 011. The false positive rate is calculated as the rate of variant calls for a reference genotype in NA12878, who is a child, and within positions presenting a true genotype configuration of 000. (*Figure reprint from Peng, G. et al PNAS, 2013*)

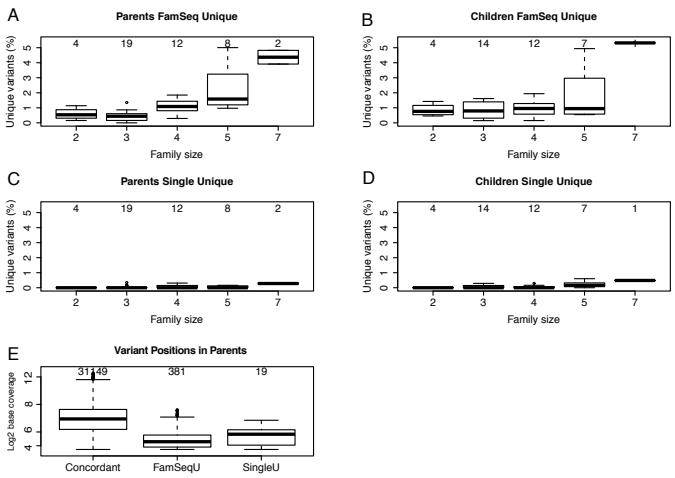


Figure 2.7 Summary of individual variant calls over family size in the target sequencing data. A-D) Distribution of unique variants added by either FamSeq or the Single method for each individual, 45 of whom are parents and 38 of whom are children. Shown is percentage of these unique calls in the concordant calls made by both methods. The total numbers of individuals at the corresponding family size are shown on top. E) Base coverage distribution for variant positions across the 45 individuals who are parents. The total numbers of variant positions for each category are shown on top. (*Figure reprint from Peng, G. et al PNAS, 2013*)

Table 2.4 Summary of individual variant calls from FamSeq and Single method across 26 families with mitochondrial disorders. The calls are made in 762KB of coding sequences. Here C denotes child, FM denotes father or mother, GP denotes grandparents. (*Table reprint from Peng, G. et al PNAS, 2013*)

FamilyID	MemberID	Relationship	Concordant	FamSeq-Unique	Single-Unique
Leigh	Leigh-Ca	C	689	7	2
Leigh	Leigh-F	FM	661	2	0
Leigh	Leigh-M	FM	638	2	0
Msy	Msy-Ca	C	640	4	0
Msy	Msy-M	FM	614	7	0
MTF01	MTF01-C1a	C	677	4	1
MTF01	MTF01-C2	C	634	6	1
MTF01	MTF01-F	FM	691	10	0
MTF01	MTF01-GM	GP	671	3	0
MTF01	MTF01-M	FM	624	7	1
MTF02	MTF02-C1a	C	668	2	0
MTF02	MTF02-C2a	C	634	5	0
MTF02	MTF02-F	FM	656	9	1
MTF02	MTF02-M	FM	595	11	0
MTF03	MTF03-Ca	C	700	2	1
MTF03	MTF03-F	FM	727	1	0
MTF03	MTF03-M	FM	700	6	0
MTF04-a	MTF04-Ca	C	671	5	1
MTF04-a	MTF04-F	FM	666	9	0
MTF04-a	MTF04-M	FM	628	3	0
MTF04-b	C	C	831	40	5
MTF04-b	F	FM	924	42	1
MTF04-b	FF	GP	929	12	1
MTF04-b	FM	GP	903	20	3
MTF04-b	M	FM	863	11	0
MTF04-c	C	C	830	41	3
MTF04-c	F	FM	927	9	0
MTF04-c	M	FM	859	43	0
MTF04-c	MF	GP	842	16	0
MTF04-c	MM	GP	851	20	1
MTF04-d	C	C	827	44	4
MTF04-d	F	FM	920	36	3
MTF04-d	FF	GP	924	15	3
MTF04-d	FM	GP	891	19	2
MTF04-d	M	FM	850	41	2
MTF04-d	MF	GP	838	16	0
MTF04-d	MM	GP	847	20	1
MTF05	MTF05-Ca	C	669	6	0
MTF05	MTF05-M	FM	644	1	0
MTF06	MTF06-C1	C	626	8	0
MTF06	MTF06-C2a	C	663	1	0
MTF06	MTF06-F	FM	699	10	0
MTF06	MTF06-M	FM	668	5	0
MTF07	MTF07-C1a	C	700	4	0
MTF07	MTF07-C2a	C	710	8	2

MTF07	MTF07-C3a	C	730	4	0
MTF07	MTF07-F	FM	697	12	1
MTF07	MTF07-M	FM	671	13	0
MTF08	MTF08-C1	C	696	9	2
MTF08	MTF08-C2a	C	705	5	0
MTF08	MTF08-F	FM	699	7	0
MTF08	MTF08-M	FM	692	8	0
MTF09	MTF09-C1a	C	685	5	1
MTF09	MTF09-C2	C	677	3	0
MTF09	MTF09-F	FM	694	6	0
MTF09	MTF09-M	FM	694	10	2
MTF10	MTF10-C1a	C	732	9	0
MTF10	MTF10-C2a	C	714	8	0
MTF10	MTF10-F	FM	685	10	0
MTF10	MTF10-M	FM	687	7	1
MTF11	MTF11-C1	C	615	11	1
MTF11	MTF11-C2a	C	567	11	0
MTF11	MTF11-F	FM	643	2	2
MTF11	MTF11-M	FM	686	2	0
MTF12	MTF12-Ca	C	689	2	0
MTF12	MTF12-F	FM	655	4	0
MTF12	MTF12-M	FM	697	1	0
MTF13	MTF13-C1a	C	681	11	0
MTF13	MTF13-C2a	C	640	7	0
MTF13	MTF13-M	FM	678	5	1
MTF14	MTF14-Ca	C	641	2	1
MTF14	MTF14-F	FM	663	0	0
MTF14	MTF14-M	FM	648	0	1
MTF15	MTF15-Ca	C	679	4	0
MTF15	MTF15-F	FM	680	3	0
MTF15	MTF15-M	FM	683	3	0
MTF16	MTF16-Ca	C	640	9	0
MTF16	MTF16-F	FM	660	4	0
MTF16	MTF16-M	FM	586	3	2
MTF17	MTF17-Ca	C	591	3	1
MTF17	MTF17-F	FM	579	1	1
MTF17	MTF17-M	FM	579	5	0
Myo	Myo-Ca	C	655	3	0
Myo	Myo-F	FM	658	4	0
OTC	OTC_m_g	FM	648	3	0
OTC	OTC_s_g	C	632	9	0
PDH	PDH-Ca	C	678	1	0
PDH	PDH-F	FM	646	2	0
PDH	PDH-M	FM	687	1	0
RCC4	RCC4-C1a	C	608	9	0
RCC4	RCC4-C2a	C	573	8	0
RCC4	RCC4-C3	C	596	5	0

Table 2.5 Summary of variant positions from FamSeq and Single method across 26 families with mitochondrial disorders. The calls are made in 762KB of coding sequences. The column of FamSeq- Unique gives the number of variant positions where more variants were called in family members by FamSeq. The column of Single-Unique gives the number of variant positions where more variants were called in family members by Single method. The column of Both-Unique gives the number of positions where more variants were called in different family members by FamSeq or Single method. (*Table reprint from Peng, G. et al PNAS, 2013*)

FamilyID	Size	Total Variant Positions	FamSeq-Unique	Single-Unique	Both-Unique	Concordant	New Positions	% FamSeq-Unique
Msy	2	766	10	0	0	755	1	0.014
OTC	2	767	12	0	0	755	0	0.016
Myo	2	809	7	0	0	802	0	0.009
MTF05	2	823	7	0	0	816	0	0.009
RCC4	3	786	20	0	0	766	0	0.025
MTF17	3	940	9	2	0	929	0	0.010
Leigh	3	958	10	1	1	946	0	0.010
MTF13	3	970	20	1	0	949	0	0.021
PDH	3	955	4	0	0	951	0	0.004
MTF04_a	3	975	17	1	0	957	0	0.017
MTF16	3	978	15	1	1	961	0	0.015
MTF14	3	987	2	2	0	983	0	0.002
MTF15	3	1012	9	0	0	1003	0	0.009
MTF12	3	1018	7	0	0	1011	0	0.007
MTF03	3	1035	8	0	1	1026	0	0.008
MTF02	4	970	23	1	0	946	0	0.024
MTF11	4	1030	25	3	0	1002	0	0.024
MTF09	4	1062	21	2	1	1038	0	0.020
MTF06	4	1085	24	0	0	1061	0	0.022
MTF10	4	1104	31	1	0	1071	1	0.029
MTF08	4	1118	24	0	2	1092	0	0.021
MTF07	5	1171	38	2	1	1130	0	0.032
MTF01	5	1179	27	3	0	1149	0	0.023
MTF04_c	5	1823	92	3	1	1721	6	0.054
MTF04_b	5	1966	87	8	2	1863	6	0.047
MTF04_d	7	2296	115	8	7	2162	4	0.052

2.3.10 Targeted Sequencing Data Analysis in Families with Mitochondrial Neurodevelopmental Disorders. These families vary in size from 2 to 7 and include single-parent, nuclear, as well as three-generation families (Table 2.1). In each individual, we sequenced 524 nuclear-encoded mitochondrial candidate genes[47,51] and focused our analysis on 962 Kb of coding regions in autosomes. We observed a significant increase in new variants called by FamSeq in the parents (Figure 2.5B and Table 2.4; FamSeq vs. Single method at size = 3: Kolmogorov-Smirnov test $P < 0.001$; FamSeq vs. Single method at size = 4: $P < 0.001$; FamSeq at size=3 vs.

size=4:P<0.001,FamSeq at size=4 vs. size > 4, P = 0.06). We measured the significantly increased number of variants as related to family size in a total of 45 individuals from 25 different families, thus accounting for biological and technological variations between different sequenced individuals. We are currently validating these positions using Sanger-based sequencing, which may facilitate finding the unknown gene defects in these families. We did not observe significant increases in variants in the children (Figure 2.3C and Figure 2.7). However, the approximate reduction in FNRs (estimated by % FamSeq-unique variants) in the three-generation pedigree was 1– 5%, which is substantially larger than the 0.1% observed at HapMap SNP positions (Table 2.5) indicating the power of FamSeq in detecting rare variants. In three of these families, we found 15 unique variant positions (Table 2.5) that are not reported in the Single Nucleotide Polymorphism Database (dbSNP) or the 1,000 Genomes Project, nine of which are non-synonymous. We also analyzed family MTF04 in three ways: trio, trio plus either pair of grandparents, and trio plus both pairs of grandparents. Interestingly, compared with the Single method for this family, only the extended pedigree (size = 5 or 7) analysis found new positions in the affected child. This illustrates the limitation of the Single method in detecting rare DNA variants and demonstrates the power of using multigeneration pedigrees to detect rare variants.

Table 2.6 Mean base coverage of all loci with HapMap heterozygous calls in FamSeq performance categories. (*Table reprint from Peng, G. et al PNAS, 2013*)

	FamSeq		
Single	Concordant	Discordant	*N*
Concordant	32 (sd = 10, *n* = 1.3M)	51 (sd = NA, *n* = 1)	16 (sd = 7, *n* = 126)
Discordant	**16 (sd = 7, *n* = 254)**	25 (sd = 11, *n* = 1784)	**14 (sd = 8, *n* = 74)**
N	**15 (sd = 7, *n* = 658)**	16 (sd = 8, *n* = 55)	14 (sd = 7, *n* = 758)

Cells in bold are where FamSeq improved on Single method (sd, standard deviation; *n*, the number of loci in each category).

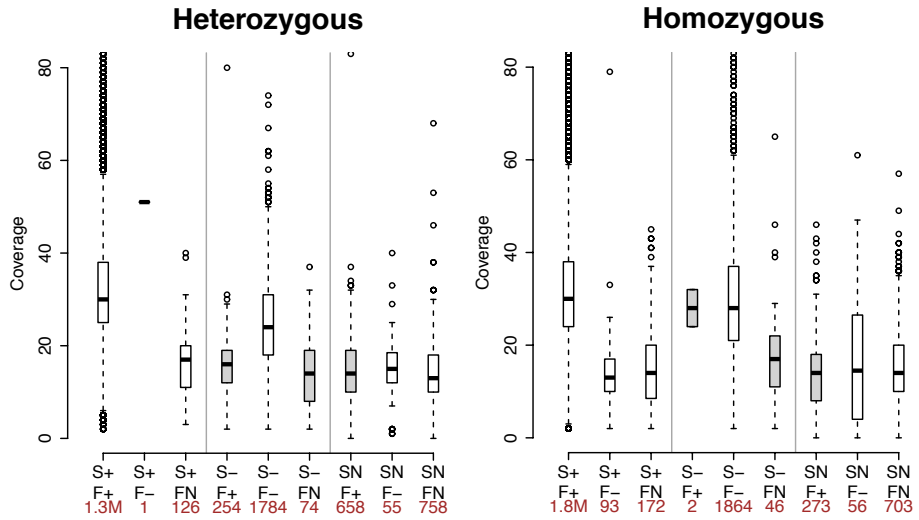


Figure 2.8 Distributions of coverage within positions, as categorized by their variant calling results from both the Single (individual-based) and FamSeq methods. Comparing HapMap SNP calls for "S", individual method, or for "F", FamSeq, we separated positions into nine categories using "+" suggesting concordance, "-" suggesting discordance and "N" suggesting no call. The total number of positions in each category is shown at the bottom of the boxplots. Improvements made by FamSeq correspond to categories "S-F+", "S-FN", and "SNF+" (grey boxes). In positions with heterozygous variants: for "S-F+", the median coverage is 16 and the range is 2 to 80; for "S-FN", the median coverage is 14, and the range is 2 to 37; for "SNF+", the median coverage is 14, and the range is 1 to 87. (*Figure reprint from Peng, G. et al PNAS, 2013*)

2.3.11 Coverage and Log Likelihood Ratios. FamSeq improved variant calling in both WGS and TS data at mean base coverages from 25× to 1,200×. In the HapMap WGS data (mean coverage 25–60×), FamSeq improved accuracy primarily at positions with low-to-moderate coverage (15–20×; Table 2.6 and Figure 2.8). NA12892 had the lowest mean coverage (25×) and presented the biggest reduction in error rates among the three samples (Figure 2.5A). Compared with the WGS data, the TS data have a wider range of mean coverage (200–1,200×). However, FamSeq still called 1.2% more variants overall, at coverage from 11 to 600× (median 24×; Figure 2.5C and Figure 2.7). To explore why, we correlated base coverage with log likelihood ratio (LLR) (input

for FamSeq) in all sequence data. We expected a genotype-specific linear relationship between LLR and coverage (Figure 2.9, r = 0.87 for heterozygotes, r = 0.80 for homozygous positions), which can be derived analytically from the underlying binomial distribution used by Samtools and GATK. FamSeq strengthens signals at positions with a low LLR (LLR < 10). Therefore, it can improve variant calling in sequencing data at positions with coverage 20× or lower. However, in TS data where most positions are at high coverage, FamSeq called more variants in 381 positions, 234 (61%) of which have high coverage (>20×) but still low LLR (<10), and thus show a relationship that varies from the expected linear relationship (Figure 2.5C and Figure 2.9).

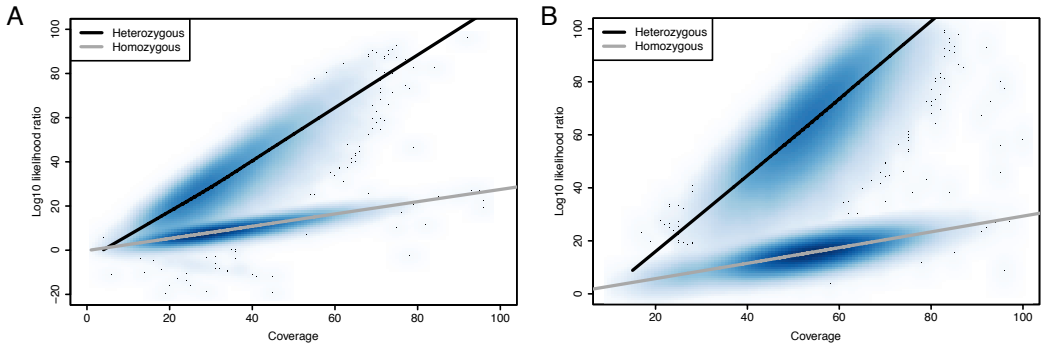


Figure 2.9 Smoothed scatterplot of log10 likelihood ratio over coverage for whole-genome sequencing data. A) shows data generated by Illumina GAII technology (NA12891, NA12892 and NA12878). B) shows data generated by Illumina HiSeq2000 technology (NA12882 and NA12877). For each panel, we randomly sampled values from 150,000 positions out of 3,000,000 positions for all samples combined. The two clusters observed are for positions with homozygous calls, and with heterozygous calls, separately. A loess curve for each cluster is shown. These lines suggest a linear relationship between log10 likelihood ratio and coverage, and present different slopes for heterozygous calls (slope=1.1 and 1.4, which is approximately -0.5log10(4e), with a corresponding error rate e of 0.002 and 0.0004 for the two technologies respectively) and homozygous calls (slope=0.27 and 0.29, which is approximately log10(2) for the two

technologies respectively). The mathematical derivation is based on a binominal distribution for variant calling, as described by Li et al. (2009). (*Figure reprint from Peng, G. et al PNAS, 2013*)

2.4 Discussion

We have developed a family-integrated method, FamSeq, which uses Mendelian transmission information to inform the calling of variants in raw sequence data. Such joint variant calling has been reported to improve variant detection using simulated data[21]. In simulations, we identified factors that may affect the level of improvements made by FamSeq, including family genotype con- figurations, the prior setting of MAF, as well as pedigree size and structure. Using actual sequence data from 28 families, we also evaluated the performance of FamSeq in practical settings, using WGS (WT family) or TS (families with mitochondrial disorders) to determine the effect of variables such as known SNPs versus unknown variants and moderate (20×) versus high sequencing coverage (>200×). By looking across 45 samples from 25 families, we accounted for biological and technical variations in real data and observed a statistically significant increase in variant detection with an increase in family size. From our comparative analysis of data with truth from two different studies using WGS, we found that FamSeq increased the sensitivity of variant calling, while still maintaining specificity. We project that the application of FamSeq to sequencing data for rare variant detection in families with heritable diseases will yield significant improvements at low, moderate, and high sequencing coverage.

To be of practical use, variant calling algorithms that use family data should be computationally efficient and also account for marriage loops and/or consanguinity. FamSeq uses a Bayesian network to compute posterior probabilities, which results in fast computation (in minutes for analyzing WGS data) with a family size less than 7. The use of parallel probability calculations will extend the utility of the Bayesian network approach to larger families.

To allow for uncertainty in the estimates of LLRs, which will further improve variant calling accuracy, FamSeq includes an MCMC approach. The LLRs represent the signal-to-noise information from each family member. In 3,600 SNP positions where both the Single method and FamSeq made mistakes, the coverage as well as LLRs are higher than the average values, suggesting a possible bias in the LLR estimates (Figure 2.8). Thus, when variances on the LLR estimates are available, our MCMC approach may be useful to correct variant calling at more positions. Similarly, variances on the MAF estimates can be incorporated when available.

The overall improvement by FamSeq is measured on a continuous scale as increased confidence in the correct call for a variant or reference position. FamSeq gives a posterior probability as the confidence measure for variant calls. We compared this with two confidence measures derived from GATK. First, we examined the quality tranches that used HapMap SNP truths to define cutoffs; second, the individual-based posterior probability at positions that passed our hard-filtering criteria. We used the quality tranches in analyzing the WGS data for the WT family and used a cutoff at the last tranche level: 99.9–100 (less specific than other levels). We used posterior probabilities in the other analyses and a cutoff at 90%, because these analyses are either for comparison with HapMap SNPs or for TS data (where quality tranches cannot be reliably generated). We considered any call with a confidence measure at or below the cutoff as a "no call." In the HapMap data, FamSeq reduced the overall no call rate at the SNP positions by giving reference or variant calls at higher confidence (Table 2.3). Changing cutoffs can shift the balance between the FNR, FPR, and no call rate in the Single method and in the comparison with FamSeq. Regardless of the cutoffs, FamSeq provides a confidence measure that incorporates family information and, compared with the Single method, better describes the uncertainty of individual genotype calls, which improves the overall accuracy.

We identify two key questions for balancing cost with obtaining adequate data to identify the disease variant of interest. (i) Who should be selected from a large family for initial sequencing?

(ii) At what coverage depth should selected family members be sequenced? FN variant calls are of great concern in these types of gene identification studies. We found that adding both parents and then grandparents before adding more siblings was most effective. One explanation is when the LLRs for the parents are similar but only one parent has a heterozygous variant, adding data from one set of grandparents (the parents of one parent) can break the tie and help identify which parent carries the variant, whereas adding data from more children cannot. Additionally, we determined that WGS data generated at an average of 25–30× coverage per person will most benefit from FamSeq analysis. While overall coverage in the WT data were ~30×, about 5–20% of all base positions had a coverage of <20× nevertheless. FamSeq was highly beneficial in correcting calling errors made at these positions (Table 2.6). In sequencing data (especially TS data) generated at an overall high coverage (200– 1,200×), FamSeq is still valuable for variant calling as there will still be positions with low coverage and also positions with high coverage but small LLR (<10). These outlier positions are likely caused by sequence-specific technical errors, allelic imbalance, or other unobservable factors.

We identified factors that can facilitate the analysis and interpretation of family sequencing studies. Using simulations with FamSeq analysis, we showed the choice of MAF had little effect on the FNRs and FPRs in children, and the FNRs in parents, but can still affect the FPRs in parents (Figure 2.3C). This remaining effect can be alleviated in two ways: (i) setting an MAF of 0.001 or less to control for the FPR, while maintaining the power to detect true variants by using FamSeq, and (ii) prefiltering FP positions, which appear to be implemented in GATK for HapMap SNPs (Figure 2.6; little reduction in FPR by FamSeq was observed with the HapMap sample originally processed by GATK). For comparison, we observed FamSeq substantially reduced the FPR in HapMap SNPs on data generated by Samtools, which is less suited to removing FPs.

In both simulations and real data (Figure 2.3A and Figure 2.4 and 2.8 and Table 2.3), we showed that FamSeq can mistakenly change calls from the individual-based method, although this happens rarely compared with the corrections it makes (1–3% vs. 14–33% in HapMap SNPs, $P < 0.001$). Therefore, when comparing results from the Single method and FamSeq, we suggest giving high priority to positions at which FamSeq changed a de novo mutation to either a Mendelian mutation or to a reference position, or added variant calls in parents or removed them in children. This prioritization needs to be integrated into the generation of lists of validation variants. In general, family-based analysis improves both sensitivity and specificity of calling Mendelian mutations. However, in the case of de novo mutation calls, this decrease in FPs may increase FNs in some occasions.

We studied two diseases, one with a dominant trait and one with suspected recessive inheritance. For the family affected with WT (autosomal dominant), we took advantage of the large pedigree (Figure 2.2) and previous linkage mapping and used a 3+2 design: a family trio with affected parent and child and two affected distant relatives. Sequence variants identified in the affected mother and son and two other relatives but not in the unaffected parent are candidates for follow-up analysis. For further sequencing, we prioritized the grandparents of the trio to uncover additional variants. Linkage information was not available for the families with mitochondrial disease, which is a genetically and clinically heterogeneous group of disorders[51], making disease-related gene discovery very challenging. One approach relies on filtering against public SNP databases for genes with two rare functional variants (homozygous or compound heterozygous) present only in the affected individuals[2]. Notably, an analysis of our targeted sequence data of 524 genes identified relatively more recessive candidate genes in the larger families (e.g., MTF04) compared with smaller families. These positions are being validated.

Our method is implemented in a C++ based software called FamSeq, which is freely available. It can process variable pedi- gree structures and accommodate de novo mutations. It

contains three approaches: a Bayesian network, an MCMC algorithm, and the Elston-Stewart algorithm. For a variant call format (VCF) file containing 3.5 M variant positions for a pedigree of seven members without loops [on an Intel(R) Xeon(R) processor with a CPU at 2.93 GHz], the respective computing times are 550 s, 550 s, and 10,000 s (10,000 iterations) for the Bayesian network, Elston-Stewart, and MCMC, respectively. When a loop is added to this pedigree, we observe little change in computing times for the Bayesian network and MCMC methods, but can increase time of at least 20–50% for loop-cutting within the Elston-Stewart algorithm[52]. FamSeq is a stand-alone module that can be integrated with existing analysis pipelines of data generated from different high-throughput platforms, both sequencing-based and array-based data[13-15,50]. Our method can be extended to give joint posterior probabilities for calling short indels in sequenced families[15]. Thus, FamSeq provides a facile and flexible means of reducing FN sequence calls, and will greatly aid in identifying disease-causing variants in next-generation sequencing studies.

3. Implementation of Variant Calling for Family-Based Sequencing Data Using Graphics Processing Units

This chapter is based upon the journal paper: Peng G., Fan Y., Wang W. "FamSeq: A Variant Calling Program for Family-Based Sequencing Data Using Graphics Processing Units". PLoS Comput Biol 10(10): e1003880. doi:10.1371/journal.pcbi.1003880 (2014). According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.

3.1 Introduction

Next-generation sequencing technologies have been employed routinely in detecting DNA variants and unveiling the cause of genetic diseases[53]. The broad application of next-generation sequencing technologies has led to an accompanying rapid development in variant calling algorithms and related software[13-15,17]. However, the variant calling error rate remains relatively high for rare variants[54], even though many new methods have been employed to improve variant calling, such as calling multiple samples together and borrowing information from the dbSNP database[55].

Roach et al. suggested using pedigree information to reduce the false positive rate of variant calling by removing all variants that do not conform to Mendelian transmission[18]. However, this method cannot control the false negative rate and cannot find any de novo mutations. Pedigree information has also been used to improve the accuracy for haplotype phasing in small families[19,45]. Recent studies have shown that incorporating pedigree information into the variant calling reduces both false positive and false negative rates for family trios and extended families[16,21,56]. Peng et al. showed that in some HapMap families, incorporating pedigree information can reduce the false positive rates by 14–33% [56].

Several software packages have been implemented to incorpo- rate pedigree information for variant calling. SAMtools[16] and DeNovoGear[57] can process family trios together. The Elston-Stewart algorithm was used in PolyMutt[21] to incorporate extended families. However, the Elston-Stewart algorithm requires either loop-cutting techniques, which will substantially increase the computing time and give approximate answers that are not always close to the exact results[26], or the use of the method proposed by Cannings et al.[27] that is hard to implement and has large memory requirements. Peng et al. proposed additional computational solutions for implementing the Mendelian genetic model in sequence variant calling[56]. The Bayesian network algorithm, in particular, provides exact results for a family pedigree with inbreeding loops. In order to allow for uncertainty in the minor allele frequency estimation, we also implemented a Markov chain Monte Carlo algorithm[24] to perform the family-based variant calling. To incorporate pedigree information into variant calling, we provide a program, FamSeq, that allows users to choose among the four following approaches, the Elston-Stewart algorithm, the Bayesian network algorithm, the graphics processing unit (GPU) version of the Bayesian network algorithm and the Markov chain Monte Carlo algorithm. FamSeq further improves the computational efficiency by using the GPU.

In whole genome sequencing, there are billions of loci with millions of candidate variant positions, so computing time is always a problem. We therefore sought to parallelize the Bayesian network algorithm in order to make the computing time feasible for analyzing a large set of whole genome sequencing data. GPUs were originally designed to accelerate the processing of graphics. As GPUs have become more programmable and have performed powerfully in parallel computing, they have been widely used in general-purpose applications, including those used in bioinformatics[58-60]. The Bayesian network algorithm contains many homogeneous tasks that can be accomplished by GPU parallel computing. Therefore, we implemented the parallel computing

of the Bayesian network algorithm using the CUDA parallel computing platform on an NVIDIA GPU, which substantially increased the performance of that algorithm.

3.2 Design and Implementation

3.2.1 Design Overview. We developed a software package, FamSeq, which calls variants for family-based sequencing data. We used different methods to implement Mendelian transmission in FamSeq.

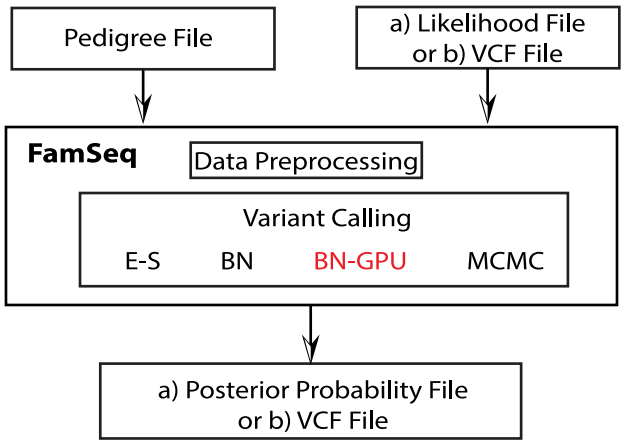
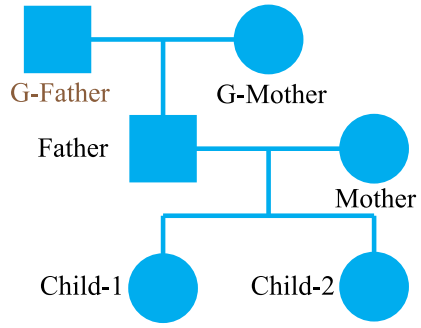


Figure 3.1 Workflow of FamSeq. We use a pedigree file and a file that includes the likelihood (Pr(D|G)) as the input to estimate the posterior probability (Pr(G|D)) for each variant genotype. (E-S: Elston-Stewart algorithm; BN: Bayesian network method; BN-GPU: The computer needs a GPU card installed to run the GPU version of the Bayesian network method; MCMC: Markov chain Monte Carlo method; VCF: variant call format.) (*Figure reprint from Peng, G. et al PLoS Comput Biol, 2014*)

As outlined in the workflow of FamSeq (Figure 3.1), two files are required as data input: a pedigree structure file and a file containing the genotype likelihood Pr(D|G), where D denotes the raw sequencing measurements, i.e., read counts, read quality and mapping quality, and G denotes the genotype of the individual. The pedigree file stores the individual identification (ID), parents'

IDs, and gender and sample name, as is used to denote samples in the likelihood data file (Figure 3.2). FamSeq accepts likelihood data files in two formats: a variant call format (VCF)[61] and a likelihood-only format (see description in our software manual). We introduced the likelihood-only format to allow for data generated from other sequencing platforms, with the requirement that the likelihood for each genotype is available.

A.



B.

ID	mID	fID	gender	IndividualName
1	0	0	2	G-Mother
2	0	0	1	NA
3	1	2	1	Father
4	0	0	2	Mother
5	4	3	2	Child-1
6	4	3	2	Child-2

C.

```
##fileformat=VCFv4.1
##FILTER=<ID=MQ0,Description="MQ0 > 40">
##FILTER=<ID=hard_to_validate,Description="MQ0 >= 4 && (DP> 0 && (MQ0 / (1.0 * DP)) > 0.1)">
##FILTER=<ID=qual,Description="QUAL < 10">
. . .
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Child-1 Child-2 Father G-Mother Mother
1 1337418 . T . 3924.52 PASS AC=0;AF=0.00;AN=10;DP=922;MQ=35.47;MQ0=0 GT:DP 0/0:196 0/0:185 0/0:107 0/0:253 0/0:181
1 1337419 . G . 3947.61 PASS AC=0;AF=0.00;AN=10;DP=937;MQ=35.48;MQ0=0 GT:DP 0/0:198 0/0:190 0/0:107 0/0:258 0/0:184
1 1337420 . C . 4081.22 PASS AC=0;AF=0.00;AN=10;DP=963;MQ=35.51;MQ0=0 GT:DP 0/0:203 0/0:195 0/0:109 0/0:268 0/0:188
1 1337421 . A . 4024.57 PASS AC=0;AF=0.00;AN=10;DP=976;MQ=35.53;MQ0=0 GT:DP 0/0:205 0/0:196 0/0:112 0/0:273 0/0:190
1 1337422 . A . 4291.47 PASS AC=0;AF=0.00;AN=10;DP=993;MQ=35.55;MQ0=0 GT:DP 0/0:206 0/0:196 0/0:116 0/0:281 0/0:194
1 1337423 . A . 4584.69 PASS AC=0;AF=0.00;AN=10;DP=1003;MQ=35.57;MQ0=0 GT:DP 0/0:206 0/0:197 0/0:121 0/0:284 0/0:195
1 1337424 . T . 4478.19 PASS AC=0;AF=0.00;AN=10;DP=1007;MQ=35.57;MQ0=0 GT:DP 0/0:208 0/0:197 0/0:122 0/0:285 0/0:195
1 1337425 . T . 4678.45 PASS AC=0;AF=0.00;AN=10;DP=1016;MQ=35.59;MQ0=0 GT:DP 0/0:208 0/0:199 0/0:123 0/0:289 0/0:197
```

Figure 3.2 Illustration of input files. A.) Pedigree structure. B.) Pedigree structure file storing the pedigree structure shown in Figure 3.2A. From the left-most column to the right-most column, the data are ID, mID (mother ID), fID (father ID), gender and sample name. C.) Part of VCF file. From the VCF file, we can find that the genome of the grandfather (G-Father) was not sequenced. We add his information to the pedigree structure file to avoid ambiguity. For example, if we include only one parent of two siblings in the pedigree structure file, it will be unclear whether they are full or half siblings. The sample name in the pedigree structure file should be the same as the sample name in the VCF file. When the actual genome was not sequenced, we set the corresponding sample name as NA in the pedigree structure file. (*Figure reprint from Peng, G. et al PLoS Comput Biol, 2014*)

FamSeq takes as input the two data files and settings of parameters (details on allele frequency and de novo mutation rate are shown hereafter). A data preprocessing feature of FamSeq will check whether there are any errors in the two input files. After that, FamSeq will implement the method the user chooses to call the variants.

FamSeq creates a new file as the output file, which follows the format of the input file but adds additional columns, with results on the posterior probability and the genotype, calculated using both the individual-based method and the family-based method.

3.2.2 Data Preprocessing. FamSeq first checks the pedigree file. FamSeq requires the input pedigree to be complete, which means that everyone listed in the pedigree should have both a father and a mother represented in the pedigree file, with the exception of the founders of the family (Figure 3.2). Otherwise, if two siblings have only one parent's information in the pedigree, FamSeq cannot determine whether they are full siblings or half siblings. FamSeq also checks for any inconsistency in the pedigree file, such as the father being erroneously listed as female. FamSeq extracts likelihood informa- tion from the Phred-scaled likelihood (PL) section in a VCF file or directly from a likelihood-only file.

3.2.3 Input of Allele and Genotype Frequency in the Population. We require Pr(G), which is the probability of the genotype in a population. In FamSeq, we consider a bi-allelic model with reference (R) and alternative (A) alleles. Consequently, there are three kinds of genotypes in a diploid genome: RR, RA and AA. Without compromising the detection of true variants, we set the default value of the frequency of three genotypes in the population at 0.9985, 0.001 and 0.0005 if the variant is not represented in dbSNP. The dbSNP information should be provided by the input VCF file. For dbSNPs, the default value is set at 0.45, 0.1 and 0.45. Users can choose to

set other values. When only the allele frequency is known, users can set genotype frequencies based on the Hardy-Weinberg equilibrium[62]. Based on findings from Peng et al., changes in the values of Pr(G) can affect the variant calling results of the founders, while its influence on offsprings in the family is small[56].

3.2.4 Rate of De Novo Mutation. In FamSeq, we require the input of the de novo mutation rate by assigning a probability of m for each parental allele to mutate into the other allele in the germline[56]. In other words, when the two parents have homozygous reference genotypes, there is still a probability that their child has a genotype with an alternative allele. We added the de novo mutation rate in the calculation of transmission probabilities (described under Model Implementation).

The de novo mutation rate has been estimated to be around 1e-8 per base per generation[20]. When we analyzed real data, we found that the rates of false positives and false negatives were better controlled when a de novo mutation rate was set at 1e-7. Thus, we set a de novo mutation rate of 1e-7 as the default in FamSeq. Users can set the de novo mutation rate according to their requirements. In general, when the de novo mutation rate is set to a large value, the influence of pedigree-to-variant calling is small and the identification of more de novo mutations is allowed during variant calling.

Even though we allow for de novo mutations in our model, we still may over-correct the variant calling at some loci by following Mendelian inheritance principles when there are true de novo mutations. Therefore, we provide the following option to alleviate the over-correction: when the likelihood ratio for all individuals in the pedigree is larger than a user-specified cutoff and the genotypes do not follow Mendel's law, FamSeq will call variants using the individual-based method instead of the family-based method.

3.2.5 Method Implementation

Markov chain Monte Carlo (MCMC) algorithm. We use the Gibbs sampler to derive the posterior probabilities for each genotype[24,25]. During Gibbs sampling, the genotype of each individual in the family is updated, one at a time, based on the condition of all other family members' genotypes, the family configuration and the raw sequencing measurements. According to Mendelian segregation principles, the genotype of the individual does not depend on those of all family members, but only on the individual's parents, spouse and children. We can write the full conditionals as follows:

$$\begin{aligned} f(G_i) &= \Pr(G_i|\mathbf{G}_{-i},\mathbf{D},\mathbf{P}) = \Pr\left(G_i|G_{f_i},G_{m_i},G_{c_i},G_{s_i},\mathbf{D}\right) \\ &\propto \Pr\left(G_i|G_{f_i},G_{m_i}\right)\prod_{j=1}^{J_i}\Pr\left(G_{c_{ij}}|G_{s_{ij}},G_i\right)\Pr(D_i|G_i) \end{aligned} \tag{3.1}$$

where G_i denotes the genotype for individual i, $\mathbf{G}_{-i}$ denotes the genotype for all family members, except individual i, **D** denotes the raw sequencing measurements, and **P** denotes the pedigree configuration. G_{fi} , G_{mi} , G_{ci} and G_{si} indicate the genotype of individual i's father, mother, child and spouse. $\Pr(G_i|G_{fi},G_{mi})$ is the transmission probability, which shows how the parents' genotypes influence the child's genotype.

To avoid a local maximization problem in the Gibbs sampler, we also implemented a heated-Metropolis algorithm in MCMC, as proposed by Lin et al.[25]. In the heated-Metropolis algorithm, G_i is sampled from a distribution of $f(G_i)^{1/T}$ instead of $f(G_i)$. The sampled $\dot{G}_i$ is accepted with the probability $\min\left\{\left[\frac{f(\dot{G}_i)}{f(G_i)}\right]^{1-1/T}, 1\right\}$.

The accuracy of the MCMC algorithm depends on the number of iterations. As is shown in Biswas et al.[24], the MCMC approach requires tens of thousands of iterations to converge for a large pedigree; therefore, the computing time will also increase. By default, we set the number of iterations at 20,000n, where n is the pedigree size. Users can specify the number of iterations according to their needs.

Elston-Stewart algorithm. This algorithm splits the whole pedigree into anterior and posterior parts according to the individual of interest[23]. The anterior part relates to the parents of the individual, and the posterior part relates to the child/children of the individual. The probability of the anterior and posterior parts can be estimated recursively, such that the posterior genotype probability is calculated according to the probability of the anterior part and the posterior part. The Elston-Stewart algorithm is especially complex when there are inbreeding loops in the pedigree because then the pedigree cannot be directly split into anterior and posterior parts. There are two methods to solve this problem. First, we can cut the loops according to complex criteria and obtain an approximate result[26,63]. Cannings et al. suggested using another method to obtain the analytical results[27]. However, their method has large memory requirements and is hard to implement.

Bayesian network algorithm. By treating the entire pedigree as a directed acyclic graph (a Bayesian network), the genotype of sample *i* depends on the genotypes of only his/her parents[22]. We can write the posterior probability as

$$\Pr(\mathbf{G}|\mathbf{P},\mathbf{D}) \propto \prod_{i=1}^{n} \Pr(D_i|G_i)\Pr\left(G_i|G_{f_i},G_{m_i}\right) \tag{3.2}$$

The Bayesian network approach directly calculates the joint probabilities for all the combinations of genotypes of the whole family, and allows for analytic calculations for pedigrees with inbreeding loops. The Bayesian network approach is straightforward and easy to implement; however, the computing time increases exponentially when the pedigree size increases, so a supplementary approach is needed for a larger pedigree.

Bayesian network parallelization. For variant calling using whole genome sequencing data, there are billions of loci. After filtering by FamSeq, there are still millions of candidate variant positions remaining; thus, we propose to parallelize the Bayesian network algorithm in order to reduce the computing time and make this approach feasible in the DNA sequencing data analysis.

In the Bayesian network method, we need to calculate the posterior probability for 3^n kinds of genotypes. This amounts to a large volume of homogeneous computing tasks that are suitable to parallel computing by GPUs.

Compared to central processing units (CPUs), GPUs have many advantages in parallel computing. A GPU usually has hundreds or thousands of core processors, while there are only several core processors for a CPU. Although the computing speed for each core processor of a GPU (about 1 GHz) is not as fast as that of a CPU (about 3 GHz), the total computing speed of a GPU is faster than that of a CPU. For a large amount of homogeneous computing tasks, we can assign one task to each GPU core to parallelize the computing.

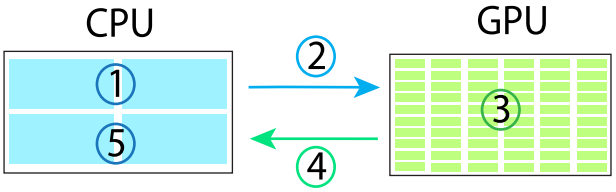


Figure 3.3 Illustration of GPU parallel computing in FamSeq. The program can be divided into two parts: a serial part and a parallel part. The serial part is processed in a CPU and the parallel part is processed in a GPU. The program: 1. Prepare the data for parallel computing in a CPU; 2. Copy the data from CPU memory to GPU memory; 3. Parallelize the 3n jobs computing in the GPU, where n is the pedigree size; 4. Copy the results from GPU memory to CPU memory; and 5. Summarize the results in the CPU.

In FamSeq, we use CUDA (version 5.0 or later) to parallelize the Bayesian network algorithm on a GPU. CUDA is a parallel computing platform and programming model developed by NVIDIA. It can be implemented on many CUDA-enabled GPUs (https://developer.nvidia.com/cuda-gpus). CUDA provides many application programming interfaces that can be easily incorporated into C++ language. A brief illustration of GPU programming in FamSeq is shown in Figure 3.3. For more details on GPU programming in

C/C++, see the NVIDIA CUDA C Programming Guide (http://docs.nvidia.com/cuda/cuda-c-programming-guide/index. html)

3.3 Results

We compared the computing time of the four different methods using real sequencing data with one million (1M) variants and a pedigree size that varied from 7 to 12. If there is no alternative allele at a position, this means that all individuals in the family have a homozygous reference genotype. If these positions are provided in the input VCF files, FamSeq will skip these positions and run joint calling on only the remaining potential variant positions. In order to estimate the actual computing time of FamSeq, we prepared a VCF file with 1M candidate variant positions as the input file. We tested the non-GPU version on a Linux server with Intel Xeon CPUs of 3.07 GHz. Only a single core of one CPU was used during testing. The GPU version was conducted on an NVIDIA Tesla M2090 with 512 cores of a 1.3 GHz GPU on a Linux server from Texas Advanced Computing Center (TACC). We used only one GPU during the comparison.

Table 3.1 Time needed for computation using FamSeq at one million positions.

Method	Loops	PU	Pedigree Size					
			7	8	9	10	11	12
E-S	N	CPU	13s	12s	15s	16s	22s	34s
MCMC[♯]	N	CPU	100,920s	129,030s	160,170s	177,740s	240,650s	296,600s
	Y	CPU	177,460s	233,490s	289,720s	362,630s	432,760s	496,750s
BN	N	CPU	242s	605s	2,003s	6,483s	23,404s	73,485s
	N	GPU*	2,472s(150s)	2907s(169s)	3,312s(239s)	3,856s(397s)	4,519s(946s)	6,452s(2,717s)
	Y	CPU	250s	902s	2,013s	6,731s	2,2078s	70,417s
	Y	GPU*	2,548s(150s)	2986s(170s)	3,123s(239s)	3,602s(399s)	4,396s(954s)	6,605s(2,726s)

PU: Processing Unit
E-S: Elston-Stewart algorithm
BN: Bayesian network algorithm
[♯]We only called 100,000 variants for MCMC. The time shown here is 10 times the time of calling 100,000 variants.
* The number in the parentheses is the GPU computing time.

Table 3.1 shows the computing time for FamSeq based on using the CPU versus the GPU. The Elston-Stewart algorithm was the fastest among the four methods we used, and was the best

choice when there were no inbreeding loops in the pedigree. The presence of inbreeding loops in the pedigree requires the use of loop cutting technology before calculating the probability of the anterior and posterior parts, which leads to algorithm complexity, increased computing time, and an approximation of the results. A big advantage of the Elston-Stewart algorithm is that the computing time increases almost linearly with increases in the pedigree size. When the pedigree size is large (greater than 12), the Elston-Stewart algorithm is almost the only computationally feasible method, especially for analyzing whole genome sequencing data.

Although the computing time for the Bayesian network algorithm increases exponentially when the family size increases, variant calling with this method can be completed in several hours for a pedigree with fewer than 10 individuals, based on whole genome sequencing data and assuming there are about 20 million candidate variant positions. When the pedigree size is small, the computing time difference between the Bayesian network algorithm and the Elston-Stewart algorithm is small. An advantage of the Bayesian network algorithm is that it can directly calculate posterior probabilities in pedigrees that have inbreeding loops. From Table 3.1, we show that the computing time for the Bayesian network algorithm is not affected by whether or not the pedigree has inbreeding loops.

We implemented the Bayesian network algorithm in both a CPU and GPU. Although we tried to increase the computing speed by parallelization at the GPU, the GPU version was slower than the CPU version when the pedigree size was less than 10. We found that transferring data between a CPU and GPU (steps 2 and 4 in Figure 3.3) requires a lot of time and becomes a bottleneck for speed improvement with GPU parallelization. The number in the parentheses in Table 3.1 is the actual GPU computing time, which is only about one tenth of the total computing time. Since the time required to copy the data increases linearly and the computing time increases exponentially, the advantage in speed improvement for GPU parallelization becomes evident when the pedigree size is larger than 10. When the pedigree size was 12, the GPU version became

10 times faster than the CPU version, which made it feasible to call variants for the whole genome sequencing data in ~36 hours as compared to more than 16 days for the CPU version. The actual improvement achieved from GPU computing will depend on its capacity, such as the total number of cores available in the GPU, which will vary from hundreds to thousands.

We ran FamSeq-GPU on a personal computer, a MacBook Pro with OS X 10.8.5, which has an NVIDIA GeForce GT 650M GPU containing 384 CUDA cores of up to 900 MHz. When the family size was 7, the corresponding GPU computing time was 360 s, which almost doubled the time needed by the TACC GPU server (Table 3.1), and the total computing time, including reading and writing between the CPU and GPU, was 3,060 s. We further observed that the GPU computing time increased to 1,970 s and the overall time increased to 7,300 s for a family size of 11. Our result shows if users do not have a professional computer server, they have an option of running FamSeq with parallel computing on a personal computer.

We also tested the computing time for our MCMC algorithm under the same settings (Table 3.1). Here, we set the total number of iterations at 20,000*n*, where n is the pedigree size. This option was the most time consuming and only provided approximate results. However, it can be used to analyze pedigrees with inbreeding loops and to incorporate uncertainty in the estimated alternative allele frequency, which is often not given as a set value, but as a value that follows a Beta distribution[57].

4. Germline Mutation Detection with Next Generation Sequencing Data from Multiple Platforms

4.1 Introduction

Next generation sequencing technology has been developed for 10 years since the first next generation sequencing platform Roche/454 occurred in 2005[6]. During the 10 years, many different kinds of next generation sequencing platforms have been developed, such as ABI/SOLiD[7], Illumina/Solexa[8,9], Life Technology Ion Torrent[10,11] and Pacific Biosciences PacBio[12]. For these platforms, they used different strategies to sequence the DNA. During template preparation, Roche/454 and Life Technology Ion Torrent platforms use emulsion PCR[64] to get enough short reads while the Illumina/Solexa platform uses solid-phase amplification[65]. After the template preparation, 100 – 200 million of clonally amplified templates are created. In each template, there are many same DNA molecules. In Pacific Biosciences PacBio platform, the templates are prepared without amplification. For each template, there is only one molecule[10,29]. To get the sequence information during DNA synthesis, most platforms use fluorescent dyelabeled nucleotides to generate light signals and use a camera to catch the signals. Life Technology Ion Torrent used a different method. It measures the voltage change in the well in which there is a bead containing the DNA template during DNA synthesis. The read length of these platforms is also different. The read length of ABI/SOLiD platform is only 35-75bp, while the read length of PacBio platform could be over 10Kb.

There are so many differences among these platforms. Their dominant error type is also different (Table 4.1). For each platform, it has its own strengths and weakness. For a single platform, it may always produce the same platform error during variant calling even we control the data quality and increase the read coverage because of the platform's special error type. If we combine the sequencing data from two or multiple platforms, we can correct some variant calling

errors by borrow the strength of advantage part of each other. Since the error type of each platform is different, we can assume the platform errors are independent for different platforms. When there is a variant calling error at a position in one platform, the other platform might give a high confident call to correct this error.

Table 4.1 Overview of major next generation sequencing platforms[29,31,66,67].

Platform	Amplification	Read Length	Sequence by Synthesis	Throughput per run	Run time	Dominant error type	Overall error rate
Roche/454	Emulsion PCR	400~1000 bp	Pyrosequencing	~700 Mb	23 h	Indel	0.50%
ABI/ SOLiD	Emulsion PCR	50~100bp	Sequencing-by-ligation	~100 Gb	10 d	Substitution	0.10%
Illumina/ Solexa	Bridge PCR	35~250bp	Reversible terminators	10~600Gb	3~12 d	Substitution	0.20%
Ion Torrent	Emulsion PCR	35~200bp	Ion semiconductor sequencing	~10 Gb	4 h	Indel	1%
PacBio	NONE	250bp~10 Kb	Single molecule sequencing	5 Gb	2 d	Indel	15%

In the past 10 years, some platforms have faded away, such as Roach/454 platform. A lot of data has been generated from these platforms. Some of these platforms are not widely used now. People usually focus on the latest technologies and platforms to sequence the individuals, even though some individuals have been sequence before by old platforms. For example, the individual NA12878 in 1000 genome project has been sequenced over 10 times with different platforms[32]. If we can combine the data from the old platforms to generate the result with comparable accuracy rate as the current platforms, we can save a lot of money and time.

We developed a method with Bayesian hierarchical model[68] to combine the next generation sequencing data from multiple platforms. In this method, we assumed that the platform error for each platform is independent. There are some features that decide the probability of platform error at different positions for each platform. The training data were used to estimate the parameters in the model. With the estimated parameters, we validated the method in testing data. In simulation, we investigated the factors that might influence the accuracy of the model,

including sample size, platform error rate, coverage depth, correlation between features and likelihood. The investigation showed the guidelines to multi-platform variant calling in real data. We also validated our method in real next generation sequencing data. We collected next generation sequencing data of NA12878 in 1000 genome project from two most popular platforms: Illumina Miseq platform and Life technology Ion Torrent platform. The sequence calls with high confidence for NA12878 provided by Zook et al.[32] were used as the true genotype in parameter estimation and model validation. In both simulation and real data, the method reduced the error rate of variant calling comparing to the single platform method when there are over 20 platform errors in both two platforms in training data set.

4.2 Method

4.2.1 Single Platform Variant Calling. For variant calling, let D_i denote the raw sequencing measurs at position i, such as read counts, read quality, mapping quality, etc. G_i denote the genotype at position i. The genotype at position i is called by estimating the posterior probability $\Pr(G_i|D_i)$. Following Bayes' rule, the posterior probability is calculated as $\Pr(G_i|D_i) \propto \Pr(D_i|G_i)\Pr(G_i)$. $\Pr(D_i|G_i)$ is the likelihood at position i. The prior $\Pr(G_i)$ is the genotype frequency in population.

4.2.2 Multi-platform Variant Calling. When there are sequencing data from 2 or more platforms, similar to the single platform variant calling, we can call the variant by estimating the posterior probability $\Pr(Gi|D_{i,1},D_{i,2},\cdots,D_{i,j},\cdots,D_{i,m})$, where $D_{i,j}$ is the raw sequencing measurements at position i from platform j, if there are total m platforms. Without loss of generality, we will use two-platform variant calling as an example. We write the model as following:

$$\Pr(G_i|D_{i,1},D_{i_2}) \propto \Pr(D_{i,1},D_{i,2}|G_i)\Pr(G_i) \tag{4.1}$$

Assuming the sequencing measurements of the two platforms are independent given the genotype, equation 4.1 can be rewritten as :

$$\begin{aligned}\Pr(G_i|D_{i,1}, D_{i_2}) &\propto \Pr(D_{i,1}, D_{i,2}|G_i)\Pr(G_i) \\ &= \Pr(D_{i,1}|G_i)\Pr(D_{i,2}|G_i)\Pr(G_i)\end{aligned} \tag{4.2}$$

In single platform variant calling, we usually assume the sequencing measurements reflect the true genotype without considering the platform error (Figure 4.1A). Here we introduce a latent variable Z to the model. If there is a platform error in the platform, Z = 1. Otherwise, Z = 0. Then for each platform we have

$$\Pr(D_{i,j}|G_i) = \Pr(D_{i,j}|G_i, Z_{i,j}=1)\Pr(Z_{i,j}=1) + \Pr(D_{i,j}|G_i, Z_{i,j}=0)\Pr(Z_{i,j}=0) \tag{4.3}$$

$Z_{i,j}$ is latent variable at position i for platform j, $j \in \{1, 2\}$. If Z = 0, there is no platform error, the sequencing measurements reflect the true genotype. Pr($D_{i,j}$ |G_i, $Z_{i,j}$ = 0) is the likelihood. When Z = 1, the platform fails. The sequencing measurements do not come from the true genotype G_T, but from another genotype $\tilde{G}$ (Figure 4.1B). We used two models to represent the platform error. To simplify the problem, we only considered two alleles in each position: reference allele R and alternative allele A. There are three kinds of genotypes: RR, RA and AA, coded as 0, 1 and 2.

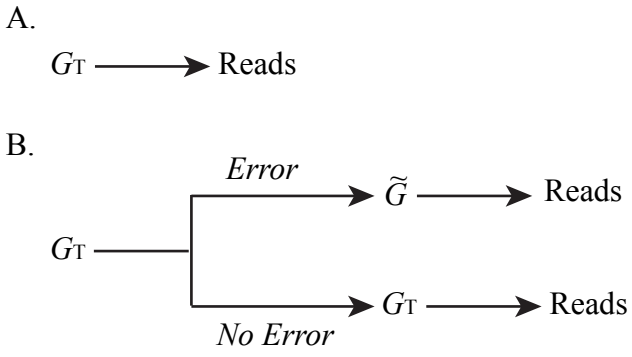


Figure 4.1 Error model illustration.

In the first model, the sequencing measurement is independent of the true genotype. The reads could come from any of the three genotypes with equal probability. In this case, we have

$$\begin{aligned}\Pr\left(D_{i,j}|G_i, Z_{i,j}=1\right) &= \sum_{k=0}^{2} \Pr\left(D_{i,j}|\tilde{G}_{i,j}=k\right) \Pr\left(\tilde{G}_{i,j}=k\right) \\ &= \frac{1}{3}\sum_{k=0}^{2} \Pr\left(D_{i,j}|\tilde{G}_{i,j}=k\right)\end{aligned} \tag{4.4}$$

$\tilde{G}_{i,j}$ is the genotype that the reads are generated from when there is a platform error at position *i* in platform *j*.

In the second model, we assume that the reads are generated from the other two kinds of genotype but the true genotype.

$$\begin{aligned}\Pr\left(D_{i,j}|G_i, Z_{i,j}=1\right) &= \sum_{\tilde{G}_{i,j}\neq G_i} \Pr\left(D_{i,j}|\tilde{G}_{i,j}\right) \Pr\left(\tilde{G}_{i,j}\right) \\ &= \frac{1}{2}\sum_{\tilde{G}_{i,j}\neq G_i} \Pr\left(D_{i,j}|\tilde{G}_{i,j}\right)\end{aligned} \tag{4.5}$$

$\Pr(Z_{i,j} = 0) + \Pr(Z_{i,j} = 1) = 1$. Here we assume the latent variable depends on some features of the platform, such as read depth, mapping quality and length of homopolymer. For example, the main error type in Ion Torrent platform is InDel[31], which is caused by homopolymer. The error rate increases according to the length of homopolymer[10]. Logistic model is used to imitate the relationship between the latent variable and features: $\Pr\left(Z_{i,j}=1\right) = \frac{\exp\left(\alpha_j + \boldsymbol{\beta_j}\, \mathbf{F}_{i,j}\right)}{1+\exp\left(\alpha_j + \boldsymbol{\beta_j}\, \mathbf{F}_{i,j}\right)}$. $F_{i,j}$ is a vector indicating the features related to the latent variable at position *i* for platform *j*. α and β are the coefficients.

4.2.3 Coefficients Estimation. If we know the latent variable Z at each position, we can use logistic regression to get the coefficients directly. However, we cannot get the latent variable directly. Instead, we use the training data with true genotype to estimate the coefficients through Markov chain Monte Carlo (MCMC) with Metropolis Hastings algorithm. Suppose the training

data has *n* positions. $G_{T_{i,j}}$ is the true genotype at position *i*. Let $\mathbf{G_T} = [G_{T_1}, G_{T_2}, \cdots, G_{T_n}]$, and $\mathbf{F}_j = [\mathbf{F}_{1,j}, \mathbf{F}_{2,j}, \cdots, \mathbf{F}_{n,j}]$. Assume the platform errors are independent at different positions, we have

$$\begin{aligned}
\Pr(\alpha_j, \boldsymbol{\beta}_j | \mathbf{G_T}, \mathbf{F}_j, \mathbf{D}_j) &\propto \Pr(\mathbf{D}_j | \alpha_j, \boldsymbol{\beta}_j, \mathbf{G_T}, \mathbf{F}_j) \Pr(\alpha_j, \boldsymbol{\beta}_j, \mathbf{G_T}, \mathbf{F}_j) \\
&\propto \Pr(\alpha_j, \boldsymbol{\beta}_j) \prod_{i=1}^{n} \Pr(D_{i,j} | \alpha_j, \boldsymbol{\beta}_j, G_{T_i}, F_{i,j}) \\
&\propto \Pr(\alpha_j, \boldsymbol{\beta}_j) \prod_{i=1}^{n} \left(\sum_{Z_{i,j}=0,1} \Pr(D_{i,j} | G_{T_i}, Z_{i,j}) \Pr(Z_{i,j} | \alpha_j, \boldsymbol{\beta}_j, F_{i,j}) \right)
\end{aligned} \tag{4.6}$$

To simplify the model, the prior distribution of the coefficients α_j and $\boldsymbol{\beta}_j$ are assumed to follow a normal distribution and are independent with each other:

$$\begin{aligned}
\Pr(\alpha_j, \boldsymbol{\beta}_j) &= \Pr(\alpha_j) \prod_{k=1}^{p_j} \Pr(\beta_{k,j}) \\
\alpha_j &\sim N(0, 100) \\
\beta_{k,j} &\sim N(0, 100)
\end{aligned} \tag{4.7}$$

Suppose there are p_j features related to the latent variable. According to the formula above, we adopted Metropolis-Hastings algorithm to estimate the parameters. During MCMC, we set burn-in as 10,000 and the iteration times as 100,000.

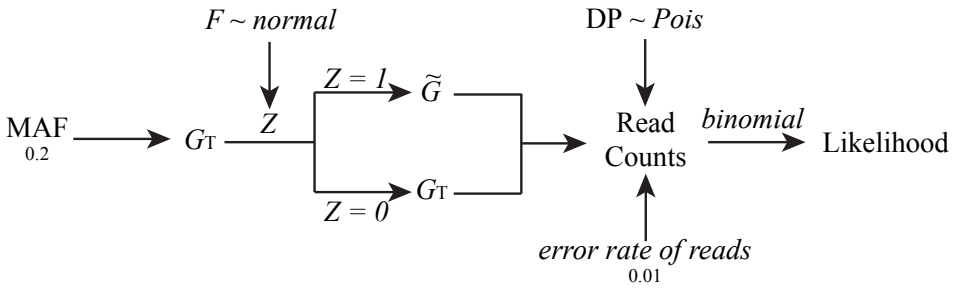


Figure 4.2 Illustration of data simulation.

4.2.4 Data Simulation. The data is simulated according to the process illustrated in Figure 4.2. The true genotype at position i in is simulated according to minor allele frequency (MAF). $\Pr(G_{T_i}=RR)=(1-MAF)^2$, $\Pr(G_{T_i}=RA)=2*MAF*(1-MAF)$, and $\Pr(G_{T_i}=RR)=MAF^2$. The reads for each position are generated according to the platform error at that position. The platform error depends on the features.

For platform j, $F_{i,j}$ with p_j features at position i is simulated according to a multi-variant normal distribution $N(0,\boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is the covariance matrix. The latent variable $Z_{i,j}$ is simulated from the distribution $\Pr(Z_{i,j}=1)=\frac{\exp(\alpha_j+\boldsymbol{\beta_j}\mathbf{F}_{i,j})}{1+\exp(\alpha_j+\boldsymbol{\beta_j}\mathbf{F}_{i,j})}$ with specified α_j and $\boldsymbol{\beta}_j$.

First we simulated the basic scenario. In the basic scenario, 2 features were related to the platform error in platform 1 and 3 features were related to the platform 2. The features at position i is simulated from multivariate normal distribution $N(0, \text{diag}(1, 1))$ in platform 1 and from $N(0, \text{diag}(1, 1, 1))$ in platform 2. We set $\alpha_1 = -8.89$, $\boldsymbol{\beta}_1 = [1.5, 3.0]^T$ for platform 1 and $\alpha_2 = -7.90$, $\boldsymbol{\beta}_2 = [1.06,1.59,2.12]^T$ for platform 2. In the basic scenario, the platform error rates for both two platforms are about 1%. The latent variable $Z_{i,j}$ is simulated according to $\Pr(Z_{i,j}=1)=\frac{\exp(\alpha_j+\boldsymbol{\beta_j}\mathbf{F}_{i,j})}{1+\exp(\alpha_j+\boldsymbol{\beta_j}\mathbf{F}_{i,j})}$. MAF is set to 0.2 to simulate the genotype and the average read depth for each platform is set to 40 to simulated the read counts for each platform. Then we changed average platform error rate, read depth and covariance between the features for each platform to check their influences to our method. In the previous simulation, the features are assumed to be independent of likelihood. However, in the next generation sequencing data, many features are related to the likelihood, which may influence the convergence and accuracy of MCMC. Instead of using features that are independent of likelihood to decide the latent variable Z, we added read depth with high correlation with likelihood as a feature influencing the latent variable Z. In

platform 1, the latent variable $Z_{i,1}$ is decided by $\Pr(Z_{i,1} = 1) = \frac{\exp(-27.8+1.5\mathrm{F}_{i_{1,1}} +0.47\mathrm{DP}_{i,1})}{1+\exp(-27.8+1.5\mathrm{F}_{i_{1,1}} +0.47\mathrm{DP}_{i,1})}$, while $Z_{i,2}$ is decided by $\Pr(Z_{i,2} = 1) = \frac{\exp(-21.6+1.06\mathrm{F}_{i_{1,2}} +1.59\mathrm{F}_{i_{2,2}}+0.34\mathrm{DP}_{i,2})}{1+\exp(-21.6+1.06\mathrm{F}_{i_{1,2}} +1.59\mathrm{F}_{i_{2,2}}+0.34\mathrm{DP}_{i,2})}$.

If $Z_{i,j} = 0$, there was no platform error. The reads were generated from the true genotype $G\mathrm{T}_i$. Otherwise, $\tilde{G}_{i,j}$ was randomly generated from three kinds of genotypes in model 1 and was randomly generated from the other two kinds of genotypes but the true genotype is model 2. After we got $\tilde{G}_{i,j}$, the reads were simulated from it. First, we simulated the read depth $d_{i,j}$ from a poisson distribution Pois($\lambda = \mathrm{DP}_j$). If $\tilde{G}_{i,j} = \mathrm{RR}$, the reference allele count $dr_{i,j}$ was simulated from a binomial distribution B($d_{i,j}$, $1 - e$). e is error rate of reads. The count of alternative allele is calculated by $da_{i,j} = d_{i,j} - dr_{i,j}$. If $\tilde{G}_{i,j} = \mathrm{AA}$, $dr_{i,j}$ was simulated from B($d_{i,j}$, e). When $\tilde{G}_{i,j} = \mathrm{RA}$, $dr_{i,j}$ was simulated from B($d_{i,j}$, 0.5).

Since the allele counts were simulated from binomial distribution, we could calculated the likelihood from the binomial model:

$$\Pr(D_{i,j}|G) = \begin{cases} \binom{d_{i,j}}{dr_{i,j}} (1-e)^{dr_{i,j}} e^{d_{i,j}-dr_{i,j}} & G = RR \\ \binom{d_{i,j}}{dr_{i,j}} 0.5^{d_{i,j}} & G = RA \\ \binom{d_{i,j}}{dr_{i,j}} e^{dr_{i,j}} (1-e)^{d_{i,j}-dr_{i,j}} & G = AA \end{cases} \quad (4.8)$$

4.2.5 Real Data Preparation. In order to validate our method, we also applied the method in real data. We downloaded fastq file of sequencing data for individual NA12878 from two platforms: IonProton Ion PI v2 and HiSeq2000. The IonProton data is whole exome sequencing data with average read depth of 80 and the HiSeq data is whole genome sequencing data with average read depth of 70. To combine the sequencing data from the two platforms, we only consider the whole exome region given in the bed file of IonProton data. The short reads from IonProton and HiSeq

were mapped to human genome reference of GRCH37 with TMAP (https://github.com/iontorrent/TS/tree/master/Analysis/TMAP) and bwa[69] respectively. After mapping, we processed the data following GTAK best practice[14]. Zook et al.[32] integrated 14 human genome sequencing data of NA12878 and provided a high confident variant data for NA12878. We downloaded their data and used it as the truth. In the intersection region of whole exome and high confident truth, there are 32,874,214 homo-reference positions and 21,663 variant positions after filtering out the positions with read depth less than 10 and 3 variant positions with more than 2 kinds of alleles. We compared the variant call from the two single platforms with the truth. At 32,871,024 positions (99.93% of total positions) both the two platform data were called as homo-reference. The real data is highly unbalanced. To balance the data, we randomly chose 20,163positions from the 32,871,024 positions and kept all the other positions to build the real data set. In the real data set, there are 21,663 variant positions and 23,353 homo-reference positions. Among these 45,016 positions, HiSeq2000 has 140 different variant calls with single platform method comparing to the truth, while there are 4988 different variant calls for IonProton Ion PI v2 data (Table 4.2).

Table 4.2 Summary of comparison of IonProton PI v2 and HiSeq2000 to high confidence sequencing call of NA12878.

		IonProton PI v2	
		Right	Wrong
HiSeq2000	Right	39923	4953
	Wrong	105	35

We randomly divided half of the data as training data and the other half as testing data. The training data was used to estimate coefficients of features for the two platforms. In the testing data, we used the estimated coefficients to call variants with two-platform data and compared the

result with variant calls from one platform. In testing data set, there are 59 single platform variant calls for HiSeq2000 and 2376 for IonProton Ion PI v2 that are different with the high confidence variant calls. Since there are much more different variant calls in IonProton data set than HiSeq2000 data set, we focused on whether we could make improvement for platform HiSeq2000 with our method. We also changed the sample size of training data to check whether it had similar result with the simulation data.

4.2.6 Features in Real Data. We collected 18 different features from the two VCF files for HiSeq2000 data and IonProton PI v2 generated by GATK (Table 4.3). Only some of these features are related to the platform error. Here we developed an ad hoc method to select features. All features were included in the model at first. Coefficients for each feature were estimated through MCMC. After 100,000 iterations, we check the trace plot of the coefficient for each feature. Some of them converged while the others did not. We only remained the features that converged and continue the previous steps until the coefficient for all the remaining features converged.

Table 4.3 Features collected from VCF files for the platforms IonProton PI v2 and HiSeq2000.

Feature	Data Type	Description
CHROM	Categorical	Chromosome, 1-22, X, Y
POS	Integer	Position of the locus
dbSNP	Boolean	In dbSNP or not
REF	Categorical	Reference allele (A, T, C, G)
ALT	Categorical	Alternative allale (A, T, C, G, .)
QUAL	Double	-10*LOG10(ALT is wrong)
DP	Integer	Read depth
RDF	Double	Reference allele frequency RD/ DP
ADF	Double	Alternative allele frequency. AD / DP
ODF	Double	Other allele frequency. (DP-RD-AD) / DP
GQ	Double	Genotype quality
Dels	Double	Fraction of Reads Containing Spanning deletions
FS	Double	-10*log10(p-value of strand bias)
MQ	Double	Mapping quality
MQRankSum	Double	Rank Sum Test for mapping qualities
QD	Double	Variant confidence normalized by unfiltered depth of variant samples
ReadPosRandSum Rank	Double	Sum Test for relative positioning of REF
numHomo	Integer	Length of homopolymer

4.3 Results

4.3.1 Sample Size And Platform Error Rate. In the simulation, we changed the sample size of training data to find out its influence to the result, while the sample size of testing data is always 20,000 to reduce the result variation. Table 4.4 shows the estimated coefficients and error rate of variant calling with 95% confidence interval for different training sample size. The last column of Table 4.4 is the error rate of variant calling with two platforms. If the called genotype is different from the truth, we tagged it as an error. The error rate is the percentage of errors in the testing data. If we only used the data from one platform, the error rate is 0.69% (0.57%, 0.82%) for platform 1 and 0.69% (0.58%, 0.81%) for platform 2 in model 1. In model 2, the error rate is 1.02% (0.88%, 1.16%) for platform 1 and 1.04% (0.88%, 1.18%) for platform 2. When the sample size is less than 500, the estimated coefficients by MCMC are far from the truth and the error rate of two platforms in larger or similar to the single platform. When the sample size is less than 500, there are about 5 or less platform errors occurring in the training data since we set the platform error as 1%. There is not enough information for MCMC to estimate the coefficients. When we checked the trace plots of MCMC, the Markov chain does not converge in some scenarios. When the training sample size is 1000, there about 10 platform errors in each platform. Even though the estimated coefficients are not very close to the truth, the error rate of variant calling for two platforms in smaller than the error rate of variant calling for single platform, especially for model 2. When the sample size is larger than 2000, the estimation is quite close to the truth and the error rate of variant calling of two platforms is only about 15% of the error rate of variant calling for single platform for both model 1 and model 2.

Table 4.4 Estimated coefficients and variant calling error rate with 95% confidence interval for different training sample size

	Sample Size	α_1	$\beta_{1,1}$	$\beta_{2,1}$	α_2	$\beta_{1,2}$	$\beta_{2,2}$	$\beta_{3,2}$	Error Rate(%)
Truth		-8.89	1.50	3.00	-7.90	1.06	1.59	2.12	
Model 1	100	-1389	128	276	-1611	124	158	259	1.38
		(-1830,-1021)	(-122,566)	(-117,848)	(-2076,-1177)	(-157,573)	(-167,626)	(-149,846)	(0.65,7.39)
	200	-1319	199	401	-1461	137	200	312	0.98
		(-1849,-15.5)	(-92.2,582)	(-48.3,798)	(-2212,-7.68)	(-242,573)	(-111,706)	(-99.9,793)	(0.43,2.47)
	500	-643	110	236	-851	120	172	265	0.51
		(-1670,-8.07)	(-0.28,391)	(2.12,647)	(-2014,-8.18)	(-54.6,489)	(0.11,551)	(1.14,746)	(0.095,1.05)
	1000	-173	22.2	67.4	-175	28.0	39.7	56.9	0.20
		(-1328,-7.59)	(0.59,237)	(2.33,506)	(-1401,-7.00)	(0.47,235)	(0.93,289)	(1.40,460)	(0.07,0.53)
	2000	-22.8	4.63	8.11	-9.42	1.25	1.88	2.61	0.12
		(-13.8,-7.97)	(1.04,2.69)	(2.38,5.12)	(-12.6,-7.20)	(0.52,2.42)	(1.17,3.13)	(1.52,3.68)	(0.067,0.20)
	5000	-9.30	1.59	3.16	-8.41	1.13	1.68	2.29	0.11
		(-11.1,-7.97)	(1.17,2.16)	(2.53,3.99)	(-9.96,-7.19)	(0.79,1.60)	(1.27,2.13)	(1.77,2.81)	(0.065,0.17)
	10000	-9.06	1.54	3.05	-8.12	1.08	1.63	2.20	0.11
		(-10.1,-8.19)	(1.23,1.92)	(2.66,3.59)	(-8.94,-7.36)	(0.82,1.37)	(1.33,1.94)	(1.90,2.50)	(0.07,0.17)
Model 2	100	-1156	99.9	213	-1303	87.6	161	194	1.60
		(-1913,-10.6)	(-158,502)	(-57.1,736)	(-2151,-9.92)	(-222,521)	(-124,687)	(-106,720)	(0.20,3.39)
	200	-617	71.5	144	-711	53.8	112	136	0.95
		(-1854,-8.11)	(-111,471)	(-69.9,592)	(-2087,-8.05)	(-163,402)	(-80.9,729)	(-33.7,665)	(0.18,1.98)
	500	-72.4	6.78	23.4	-93.0	14.7	9.87	20.9	0.30
		(-956,-7.83)	(0.38,85.5)	(2.18,225)	(-1504,-6.82)	(-0.11,217)	(0.77,152)	(1.23,380)	(0.11,1.14)
	1000	-11.0	1.83	3.80	-9.83	1.37	2.11	2.59	0.17
		(-18.6,-7.99)	(0.69,3.09)	(2.40,6.69)	(-15.9,-6.79)	(0.53,2.64)	(0.85,4.37)	(1.40,4.32)	(0.11,0.29)
	2000	-9.47	1.60	3.20	-8.58	1.19	1.79	2.25	0.16
		(-12.5,-7.70)	(0.93,2.48)	(2.33,4.41)	(-11.8,-6.91)	(0.65,1.88)	(1.06,2.75)	(1.65,3.15)	(0.10,0.24)
	5000	-9.11	1.53	3.08	-8.17	1.11	1.65	2.18	0.16
		(-10.5,-8.08)	(1.15,2.01)	(2.57,3.66)	(-9.53,-7.24)	(0.79,1.41)	(1.28,2.04)	(1.76,2.67)	(0.11,0.22)
	10000	-8.97	1.50	3.03	-8.02	1.08	1.63	2.15	0.16
		(-10.1,-8.02)	(1.24,1.78)	(2.60,3.52)	(-8.93,-7.41)	(0.84,1.31)	(1.37,1.93)	(1.88,2.41)	(0.11,0.21)

In real data, the overall platform error rate for different sequencing platform varies from 0.1% to 15%[31]. We changed the platform error rate from 10% to 0.1% to find out the sample size we need in training to estimate the coefficients accurately. Figure 4.3 shows the relationship between the error rate of two-platform variant calling and training sample size for different platform error rates. It seems that when there are about 10 errors in each platform in the training data, there is significant improvement for variant calling when we combine the two platforms together. When there are more than 20 errors in each platform, more errors help little to improve the variant calling accuracy. In the following simulation, we set the sample size of training data as 10,000 and the error rate to 1% to make sure that we estimated the coefficients accurately.

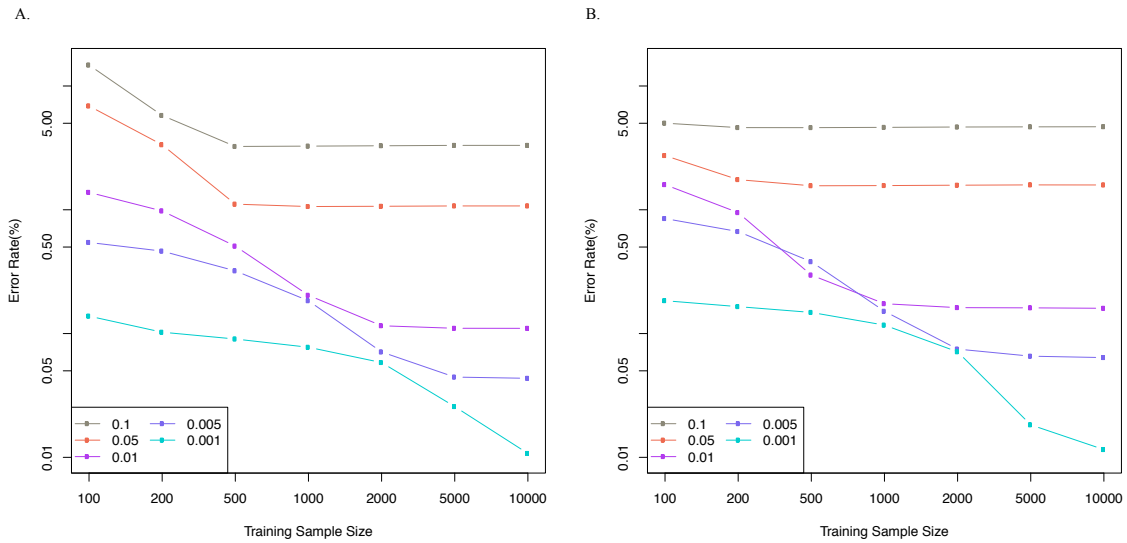


Figure 4.3 The two-platform variant calling error rate with different training sample size for different platform error rates. Each line indicates different platform error rate. A) Model 1. B) Model 2.

4.3.2 Read Depth. Read depth has a big influence to the variant calling of sequencing data. As the read depth increases, the error rate of variant calling usually decreases. When we combined the sequencing data from two platforms, it seemed that we doubled the read depth. To check whether the multi-platform variant calling method gets advantage from doubled read depth we changed read depth from 10 to 40 and compared multi-platform variant calling result with single platform result with read depth from 20 to 80. When the read depth is less than 20, the variant calling error rate decreases a lot when the read depth is doubled for single platform variant calling. The two-platform method has a little lower variant calling error rate than the single platform method with doubled read depth. When the read depth is more than 30, there is little difference of variant calling error rate between single platform method and single platform method with doubled read depth, while two-platform method still makes a big improvement to decreased the variant calling error rate (Figure 4.4). In the following simulations, we set the read depth to 40 for each platform

to make sure that multi-platform variant calling method didn't take advantage from doubled read depth.

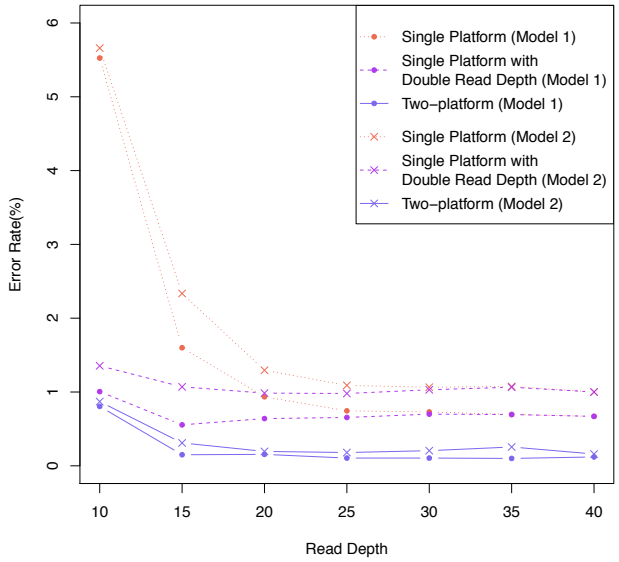


Figure 4.4 Variant calling error rate with different read depth.

4.3.3 Correlation Between Features. In real data, the features we selected to predict the platform error are usually correlated with each other, which may influence the estimation of coefficients. In order to find out the influence of the correlation between features and to check the robustness of our method, we changed the covariance between the features in the simulation. Two features in the first platform and the last two features in the second platform were set to be correlated. We changed the correlation coefficient of these two features in the two platforms from 0 to 0.95 and kept all other parameters the same. As it is shown in Figure 4.5, when correlation coefficient increases, the variant calling error rates for both single platform and two platforms also increases. However, the variation of correlation coefficient does not change the trend that the two-platform variant calling error rate decreases when the number of platform errors for each platform increases to about 20 and there is little changes after that.

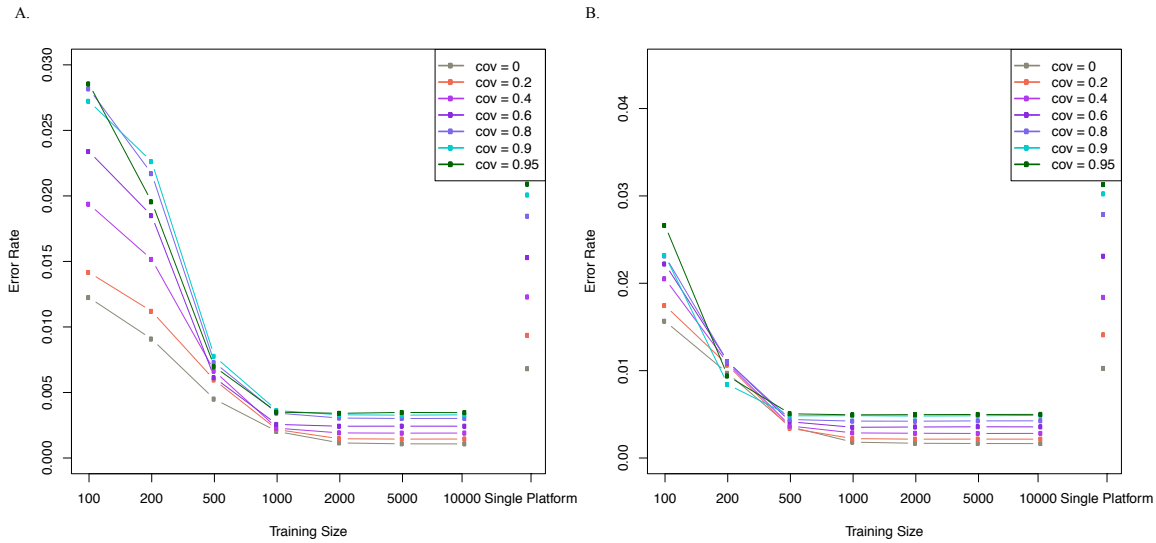


Figure 4.5 Variant calling error rate with different training sample size for different correlation coefficients between latent variable related features. Each line indicates the two-platform variant calling error rate with different correlation coefficients, while the dots on the right side of figure indicate the variant calling error rate for single platform. A) Model 1. B) Model 2.

4.3.4 Likelihood Related Features. When calculating the likelihood Pr(D|G), many different factors are considered, such as read depth and mapping quality, while these features may be selected to predict platform error at the same time. Read depth was selected as a feature related to both platform error and likelihood to check whether likelihood related features would influence the accuracy of the model. In the simulation, the average read depth was set to 40 and the training sample size was set to 10,000. The variant calling error rate for the two-platform method is estimated as 0.25% (0.18%, 0.36%) for model 1 and 0.39% (0.30%, 0.58%) for model 2, while the error rate for the single platform is 0.7% (0.59%, 0.84%) for model 1 and 1.10% (0.90%, 1.22%) for model 2.

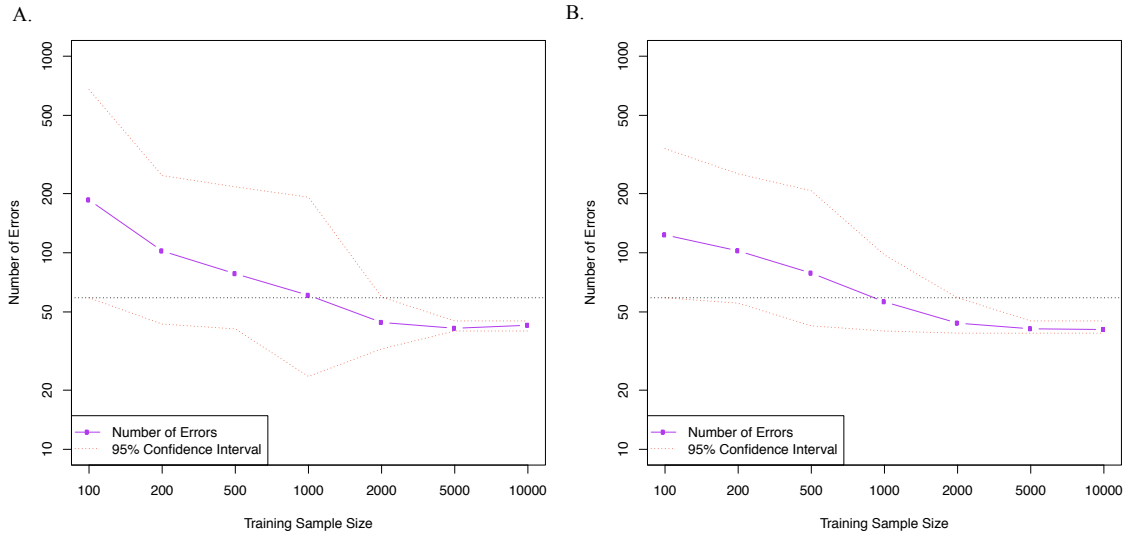


Figure 4.6 Number of two-platform variant calling errors in real data with training sample size. The horizontal line shows the number of variant call errors for single platform HiSeq2000 data set. A) Model 1. B) Model 2.

4.3.5 Real Data Validation. We include all the features in Table4.3 into the model and process the feature selection for each platform first. In HiSeq2000 platform, four features were selected: dbSNP, ADF, Dels and FS, while in IonProton PI v2 platform, there were also four kinds of features selected: dbSNP, ODF, Dels and numHomo. We include all the features in Table 4.3 into the model and process feature selection for each platform first. In HiSeq2000 platform, four features were selected: dbSNP, ADF, Dels and FS, while in IonProton PI v2 platform, the following features were selected: dbSNP, ODF, Dels and numHomo. In real data validation, we also changed training sample size from 100 to the number of all loci (22,508) in the training data set. The relationship between the variant calling error rate and the training sample size was shown in Figure 4.6. It seems that when there are less than 10 platform errors (training sample size less than 2000) in HiSeq2000 platform, we cannot estimate the parameters accurately. The variation of two-platform variant calling error rate is big. There is little improvement to combine the two

platforms. If there are more than 10 platform errors in platform HiSeq2000, there is improvement of combining the two platforms comparing to the single platform HiSeq2000. When there are over 20 platform errors in platform HiSeq2000, the estimated parameters are quite consisted with each other for different runs. The confidence interval is very narrow. We can get stable improvement when there are more than 20 platform errors for each platform. This result is similar to the result from the simulation. When we use all the training data to estimate the parameters, the number of variant calling error is 43, 27% of improvement comparing to the 59 variant calling error in HiSeq2000 data set.

4.4 Discussion

We have developed a variant calling method to integrate next generation sequencing data from different platforms with Bayesian hierarchical model. This method takes advantages of merit from one platform to overcome the weakness from another platform to reduce the error rate of variant calling. In both simulation and real data, our method performs better than the single platform method when we have enough training data. We identified the influence of training sample size, read depth, covariance between features and likelihood related features to the model with simulation. We also applied our method to the real sequencing data of NA12878 from the two most popular sequencing platforms: Illumina sequencing platform (HiSeq2000) and Ion Torrent platform (IonProton PI v2). The result shows that our method improves the accuracy for both platforms when we combine them together.

The result in both simulation and real data shows that our method makes improvements when there are more than 10 platform errors for each platform. When the number of platform error is more than 20 for each platform, the improvements are significant. More platform errors over 20 have very small improvements to reduce the variant calling error rate. That one platform has

many platform errors while the other have less than 10 platform errors does not help but might increase the variant calling error rate because the parameters estimated from the platforms with less errors might be far from the truth and causes more mistakes.

Usually, we can increase the read depth to improve the variant calling accuracy. High read depth reduces the variation during reads sampling to decrease the variant calling error rate. However, it will not help a lot when read depth reaches some point (Figure 4.4). Especially when the platform error is independent of the read depth. We cannot reduce the platform error rate by increasing the read depth. Our method partially solves this problem by combining the sequencing data from two different platforms. The platform errors for different platforms are usually independent from each other. When a platform error happens on a locus in one platform, it is unlikely that there is another platform error on the same locus in another platform. In this case, we can use the information from the platform without the platform error at this locus to correct the platform error in another platform. If the two sequencing data sets are from the same platform, the platform errors are not independent with each other. There might be platform errors on the same locus for the two data sets. Our model makes little improvement in this situation.

In our model, the read depth for the two platforms should be comparable or the read depth of the platform with lower read depth cannot be very small (less than 10). If the read depth for one platform is very small, it cannot provide the signal (likelihood ratio between the true genotype and other genotypes) strong enough to correct the platforms in the other platform with higher read depth. The result of two platform variant calling will mainly follow the platform with higher read depth.

The features in the next generation sequencing data are usually correlated. Some features might be calculated from one or a few other features. In simulation, it seems that the correlation does not affect the estimation of the coefficients but increase the variant calling error rate for both single platform method and two-platform method. It is more complex in real data. The result

might not as good as in simulation. However, in real data validation, dbSNP and ADF are high correlated in HiSeq2000 platform (correlation coefficient = 0.85, p-value < 2.2e-16). Our method still works well when the correlation is not very high.

The correlation between the features and likelihood is another problem we need to consider. The likelihood is calculated from some of the features. The likelihood has high correlation with these features. These features influence the variant calling error rate through both the platform error and likelihood. In simulation, we set read depth positive correlated with platform error. The platform error increases when we have higher read depth. While on the other hand, high read depth increases the signal to noise ratio (likelihood ratio between the true genotype and other genotypes) to give more accurate result. These two effects counteract with each other. That is why the two-platform variant calling result in this scenario (0.25% in model 1 and 0.39% for model 2) is high than in the basic scenario (0.11% in model 1 and 0.16% for model 2).

Both the feature dbSNP and Dels are selected in HiSeq2000 and IonProton platform. As we know, variant calling is easier for common variants (in dbSNP) than rare variants. In our model, the negative coefficient indicates the probability of platform error decreases when the feature increases. That is why the coefficient estimated for dbSNP is negative for the two platforms. It is hard to call variants when there are many indels around, which explains the positive parameters of Dels in HiSeq2000. It is different in IonProton data set. In IonProton, the main cause of platform error is homopolymer. The longer the homopolymer is, the more likely there is a platform error. That is why numHomo is included in the IonProton data set and the parameter is position. But when indels are called, it will reduce the error of homopolyer. The parameter for Dels in IonProton is negative. There are usually two kinds of alleles at each locus. When there are more than two alleles, it has high possibility of a platform error like the feature of ODF in IonProton data set. The other two features selected in HiSeq2000 are ADF and FS. Their parameters are both positive. FS reflects the possibility of strand bias, which leads to platform

error. We are still investigating the reason why ADF in included in HiSeq2000 data set with positive parameters.

There are 9 out of 14 platforms that Zook[32] used to generate the high confidence variant calls coming from Illumina. According to Zook's criteria, the high confidence variant calls have preference on Illumina data to be right. It partially explained the high variant calling error rate in the IonProton data set. If it is the truth, our method will have more than 27% of improvement for HiSeq2000 data set.

5. Estimating *TP53* Mutation Carrier Probability in Families with Li-Fraumeni Syndrome Using LFSpro

5.1 Introduction

Li-Fraumeni syndrome (LFS) is an autosomal dominant inherited cancer predisposition syndrome. It was first described by Drs. Li and Fraumeni after they encountered unusual familial clustering of childhood soft tissue sarcomas with early onset breast and other cancers in relatives in four families[70],confirmed in a survey of over 600 childhood rhabdomyosarcoma patients and their families[71]. In 1988 Li et al. summarized the findings in 24 kindreds identified by a proband with a sarcoma before age 45 years, a first degree relative with cancer in that age interval, and another close (first or second degree) relative in the lineage with either cancer before age 45 or a sarcoma at any age. Using these criteria they identified the 6 most common tumors termed the component tumors including soft-tissue and bone sarcomas, breast cancer, brain tumors, leukemia and adrenal cortical carcinoma. These cancers occurring before age 45 accounted for almost 80% of all cancers in the 24 kindreds, and these criteria formed the basis for the classic LFS clinical criteria.

In 1990, germline mutations in the tumor suppressor gene *TP53* were identified in 5 out of 5 families studied with LFS[72]. However, since then it has been shown that only 60-80% of 'classic' LFS families have *TP53* germline mutations[36]. The lifetime cancer penetrance for carriers with TP53 mutation is about 90%[33,34]. For women, this penetrance increases to nearly 100% by age 70 years, mostly because of breast cancer[35]. A recent study demonstrated that with clinical surveillance, which included whole-body and brain magnetic resonance imaging, people with *TP53* mutations had significantly improved survival[73]. These findings indicate that early identification of *TP53* mutation carriers is beneficial as cancer screening can significantly improve survival.

To date, several sets of clinical criteria have been developed to identify individuals with *TP53* mutations. The first criteria proposed, termed classic LFS criteria, has high specificity, but low sensitivity in the identification of patients carrying *TP53* mutations[37]. Later, Birch[74] and Eeles[75] introduced Li-Fraumeni-like syndrome criteria, which are more sensitive and include some but not all components of the classic LFS criteria. However, both the classic LFS and the Li-Fraumeni-like syndrome criteria have bias in ascertainment and family selection[38]. To avoid this bias and allow for identification of *TP53* mutation carriers with a negative family history (e.g., cases in which the proband has an unaffected parent carrying a TP53 mutation or the proband has a de novo TP53 mutation), Chompret et al. proposed alternative criteria to identify TP53 mutation carriers[38,39]. The Chompret criteria are more specific with respect to cancer type and age of onset and focuses more on the cancer information of the individual. This can lead to a high false positive rate.

The classic LFS and Chompret criteria are all widely used in clinical practice as guidelines to identify patients with LFS. However, these criteria focus on affected family members without considering healthy individuals in the family and therefore much of the family information is discounted. In addition these criteria can only be applied to individuals and families with LFS spectrum cancers.

To address these limitations, we developed LFSpro, which can estimate an individual's *TP53* mutation probability on the basis of a detailed family history of cancer. LFSpro is built on a Mendelian risk prediction model[40], which has been successfully used in PancPRO[76] and MelaPRO[77]. To develop LFSpro, we introduced a de novo mutation rate into the original Mendelian risk prediction model to account for the relative frequent occurrence of de novo mutations in families with LFS[78]. We validated the performance of our model using 183 families prospectively collected at The University of Texas MD Anderson Cancer and 582 families from the International Sarcoma Kindred Study, Australia. We found that LFSpro had both higher

sensitivity and higher specificity in identification of *TP53* mutation carriers compared to the classic LFS and Chompret criteria.

5.2 Method

5.2.1 Model Development. LFSpro estimates the probability of any designated family member carrying a *TP53* mutation on the basis of the penetrance of *TP53* mutations and the detailed family history of cancer. The penetrance of *TP53* mutations is defined as the probability of developing any cancer at each age[35]. We built LFSpro based on Mendelian risk prediction models[40,76,77]. The probability of an individual carrying a *TP53* mutation is estimated via the following formula:

$$\Pr\left(G_i \mid \mathbf{D},\mathbf{P}\right) = \frac{\sum_{G_{-i}}\left[\Pr\left(D_i \mid G_i\right)\prod_{j\neq i}\Pr\left(D_j \mid G_j\right)\Pr\left(G_i \mid G_{fi},G_{mi}\right)\prod_{j\neq i}\Pr\left(G_j \mid G_{fj},G_{mj}\right)\right]}{\sum_{G_i}\left\{\sum_{G_{-i}}\left[\Pr\left(D_i \mid G_i\right)\prod_{j\neq i}\Pr\left(D_j \mid G_j\right)\Pr\left(G_i \mid G_{fi},G_{mi}\right)\prod_{j\neq i}\Pr\left(G_j \mid G_{fj},G_{mj}\right)\right]\right\}} \quad (5.1)$$

Here G_i denotes the genotype for individual i, G_{-i} denotes genotypes for all individuals in the pedigree except individual i, G_{fj} and G_{mj} are the genotypes of individual j's father and mother, D_i denotes the phenotype of individual i, and P denotes the pedigree structure. $\Pr(\mathrm{D}|\mathrm{G})$ is likelihood that can be estimated from the penetrance. $\Pr\left(G_{\mathrm{i}}|G_{fi},G_{mi}\right)$ denotes the transmission probability. To account for the substantial *de novo* mutation rate in LFS families, we incorporated a *de novo* mutation rate into the Mendelian risk prediction model. In other words, when both parents lack *TP53* mutations, there is a small probability (2 times *de novo* mutation rate) that their child is a *TP53* mutation carrier. During method validation, we used several different *de novo* mutation rates to check which rate was the best for our model.

It is computationally challenging to calculate the posterior probability from formula 1 directly. Instead, we used the Elston-Stewart algorithm to calculate the posterior probability for extended families[23].

5.2.2 Validation Study Population The study population used to validate our method consisted of two different groups of patients. The first group comprised of childhood soft-tissue sarcoma or osteosarcoma patients ('probands') treated at The University of Texas M.D. Anderson Cancer Center (Houston, TX) from 1944 to 1983 and their extended families members. The details on data collection, frequencies of site-specific cancers and germline testing of these patients have been described previously[33,34,79-81]. These extended kindreds were prospectively followed up for more than 30 years. Medical records and death certificates were used to confirm reported cancers, where possible. For the current analysis we defined case subjects as affected only if diagnosed with malignant tumors, excluding non-melanoma skin cancers, but including all adrenal cortical tumors, choroid plexus tumors, ovarian granulosa tumors and breast carcinoma *in situ*. Cancer diagnoses were further subdivided according to site: LFS spectrum (osteosarcomas, soft-tissue sarcomas, breast, brain, adrenal, lung and leukemia) and non-LFS spectrum cancers (prostate, colon, kidney, thyroid and others). Peripheral blood samples were collected from the probands and their relatives after gaining informed consent. Probands' *TP53* status was determined by PCR sequencing of exons 2-11. If *TP53* mutation was identified, then testing was extended to all first-degree relatives (both affected and unaffected with cancer) of the proband or any other family member found to have the familial mutation. This approach of extending the germline testing based on mutation status and not on phenotype of family members should not introduce an ascertainment bias during the analysis[33,35]. Individuals unavailable for testing (largely deceased) who were the links between confirmed mutation carriers were considered obligate mutation carriers. In kindreds in which the proband was not found to have a *TP53* mutation by sequencing,

no other family members were tested. This first group of patients (from here onwards referred to as 'pediatric sarcoma') consisted of 183 unrelated families with 2553 individual members (Table 5.1). 11 of the kindreds included at least one individual with a confirmed*TP53* mutation. Data from the other 172 *TP53* negative families was used to assess whether LFSpro can control for the false positive rate in our model.

Table 5.1 Overview of validation data sets by families.

	pediatric-Sarcoma		adult-Sarcoma	
	Family with TP53 Mutation Carrier		Family with TP53 Mutation Carrier	
	Yes	No	Yes	No
N_{Family}	11	172	19	563
$N_{individual}$	1256	1297	591	16386
$N_{individual}$ with $N_{primary\ cancer} = 1$	127(10.11%)	265(20.43%)	78(13.20%)	1788(10.91%)
$N_{individual}$ with $N_{primary\ cancer} = 2$	19(1.51%)	0	7(1.18%)	165(1.01%)
$N_{individual}$ with $N_{primary\ cancer} > 2$	7(0.56%)	0	8(1.35%)	48(0.29%)

The second group consisted of a prospective cohort of adult-onset sarcoma patients and their extended families recruited to the International Sarcoma Kindred Study from six major sarcoma treatment centers across Australia. All reported cancer diagnoses were verified by reference to medical records, Australian and New Zealand cancer registries or death certificates[82]. The proband's *TP53* status was determined by PCR sequencing, high resolution melt (HRM) analysis and multiplex ligation-dependent probe amplification (MLPA) analysis to detect large deletions or genomic rearrangements[82]. This second group of patients (referred to as 'adult sarcoma families') consisted of 582 separate families with 16,977 individuals (Table 5.1). 19 of these kindreds had at least one individual with a confirmed *TP53* mutation.

5.2.3 Validation Study Design. The *TP53* mutation carrier probability for the individuals with or without *TP53* mutations was estimated using LFSpro. The receiver operating characteristic

(ROC) curve was used to evaluate the model discrimination. Both classic LFS and Chompret criteria were used to estimate whether each individual had a *TP53* mutation. We then compared the results from LFSpro to the results from the classic LFS and Chompret criteria to evaluate our method.

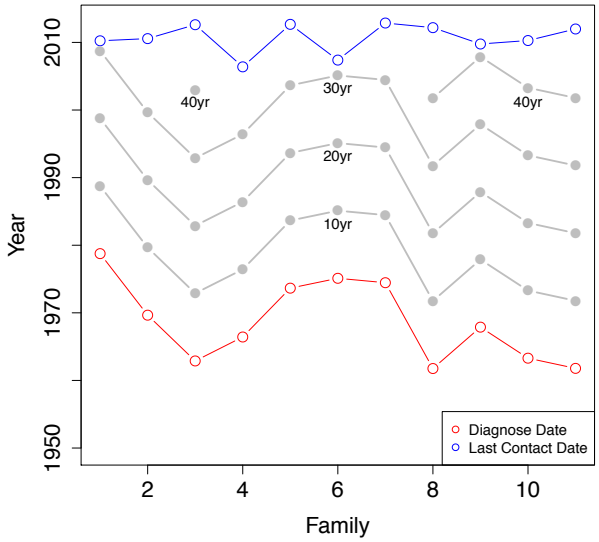


Figure 5.1 Illustration of time range of data collection for pediatric-sarcoma *TP53* positive families. Data was collected since the proband of the families was first diagnosed with LFS-spectrum cancer.

5.2.4 Roll Forward. Most of the *TP53*-positive pediatric-sarcoma data have been collected over the course of 40 years (Figure 5.1). To evaluate our model works with limited pedigree information and to evaluate whether our model could predict LFS-spectrum cancers, we re-arranged the validation population data by starting at the time when the proband was first diagnosed with an LFS-spectrum cancer and rolling forward every 10 years including only those individuals, their ages and their cancer affection status of that period. At each roll forward time point, classic LFS criteria, Chompret criteria and LFSpro were used to assess the influence of

time on the carrier estimates. Because we did not have a long observation time for *TP53*-negative families, we only rolled forward the data for *TP53*-positive families

5.3 Results

5.3.1 Clinical Illustration. Several scenarios of a hypothetical pedigree were used to illustrate how LFSpro provides information to support clinical decisions by estimating *TP53* mutation probability for counselees. In scenario 1 (Figure 5.2A), the *TP53* mutation probability for the counselee estimated by LFSpro is 39.47%, while this probability is about 0.01% in general population. This mutation carrier probability increases when it is more likely a Mendelian transmission occurs in the pedigree. For example, if we exchange the status between the counselee's grandmother and the grandmother's sister (the counselee's grandmother had breast cancer at age 18 while the grandmother's sister was healthy at age 42, scenario 2, Figure 5.2B), the *TP53* mutation probability for the counselee will increase to 77.78%. If more individuals in the family have cancer, the probability that the counselee is a *TP53* mutation carrier also increases. If both the counselee's grandmother and the grandmother's sister have cancer (scenario 3, Figure5.2C), the counselee's probability rises to 89.94%. In the first three scenarios, it is likely that the *TP53* mutation is transmitted from the counselee's great grandmother to the counselee. The probability of *de novo* mutation is small. If the pedigree is as shown but the counselee's grandmother is healthy at an advanced age and we exchange the status of the counselee's mother and her aunt (scenario 4, Figure 5.2D), the probability that the grandmother is a *TP53* mutation carrier is small. It is possible that there is a *de novo* mutation for the counselee's mother. If we consider the possibility of *de novo* mutation and set the *de novo* mutation rate to 0.0005, the carrier probability is 48.8% for the counselee, while if we ignore the *de novo* mutation possibility the carrier probability is 0.5% (Table 5.2). Consideration of the *de novo* mutation rate therefore

can have a significant impact on the carrier probability. The results of application of the classic LFS and Chompret criteria for the counselee are also shown in Table 5.2.

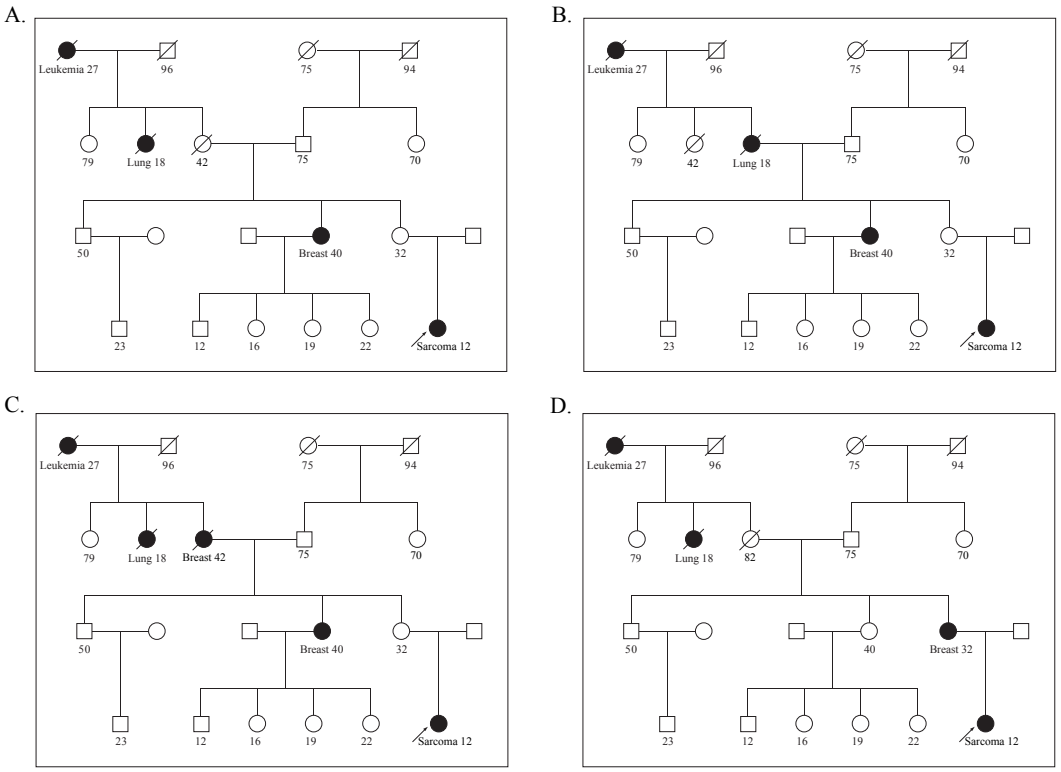


Figure 5.2 A hypothetical family pedigree to illustrate the clinical utility of LFSpro. An arrow points to the counselee, for whom the *TP53* mutation probability is calculated by LFSpro on the basis of her family history of cancer. The sites of cancer onset are provided below each of the affected family relatives, together with the age in years at which the cancer was diagnosed. A-D) Four variations in clinical scenarios.

Table 5.2 Clinical illustration of LFSpro.

	Carrier Probability[a]	Counselee's Absolute Risk of Developing Cancer		Classic Criteria	Chompret Criteria
		By Age 40 Years	By Age 70 Years		
General Population[b]	-	2.0%	21.9%	-	-
Shown in Fig 1A (Scenario 1)	39.5%	27.3%	47.1%	N	Y
Exchange the status of counselee's grandmother and the grandmother's sister (Scenario 2, Fig 1B)	77.8%	52.1%	80.5%	N	Y
Counselee's grandmother is also affected at age 42 years (Scenario 3, Fig 1C)	89.9%	59.9%	91.2%	N	Y
Counselee's grandmother is healthy at an andvanced age while her mother is affected (Scenario 4, Fig 1D)[c]	0.5% 48.8%	2.2% 33.3%	13.0% 55.2%	N	Y

[a] We set 0.0005 as the default for the de novo mutation rate.

[b] SEER 2009-2011 data

[c] In Scenario 4, we calculated the carrier probability at de novo mutation rate = 0, for comparison.

5.3.2 De Novo Mutation Rate. In LFSpro, we incorporated the de novo mutation rate into the Mendelian risk prediction model. Figures 5.3A, 5.4 and 5.4 show the influence of *de novo* mutation rate on the result. The area under the ROC curve (AUC) increases while the de novo mutation rate increases until the de novo mutation rate is about 0.0005 (Figure 5.3B), and then the AUC decreases. The observed expected ratio (OE ratio, the ratio between the number of observed *TP53* mutation carriers and the summation of probability of *TP53* mutations for all individuals estimated by LFSpro) is also close to 1 when the mutation rate is 0.0005. Therefore, we used 0.0005 as the default de novo mutation rate in our study.

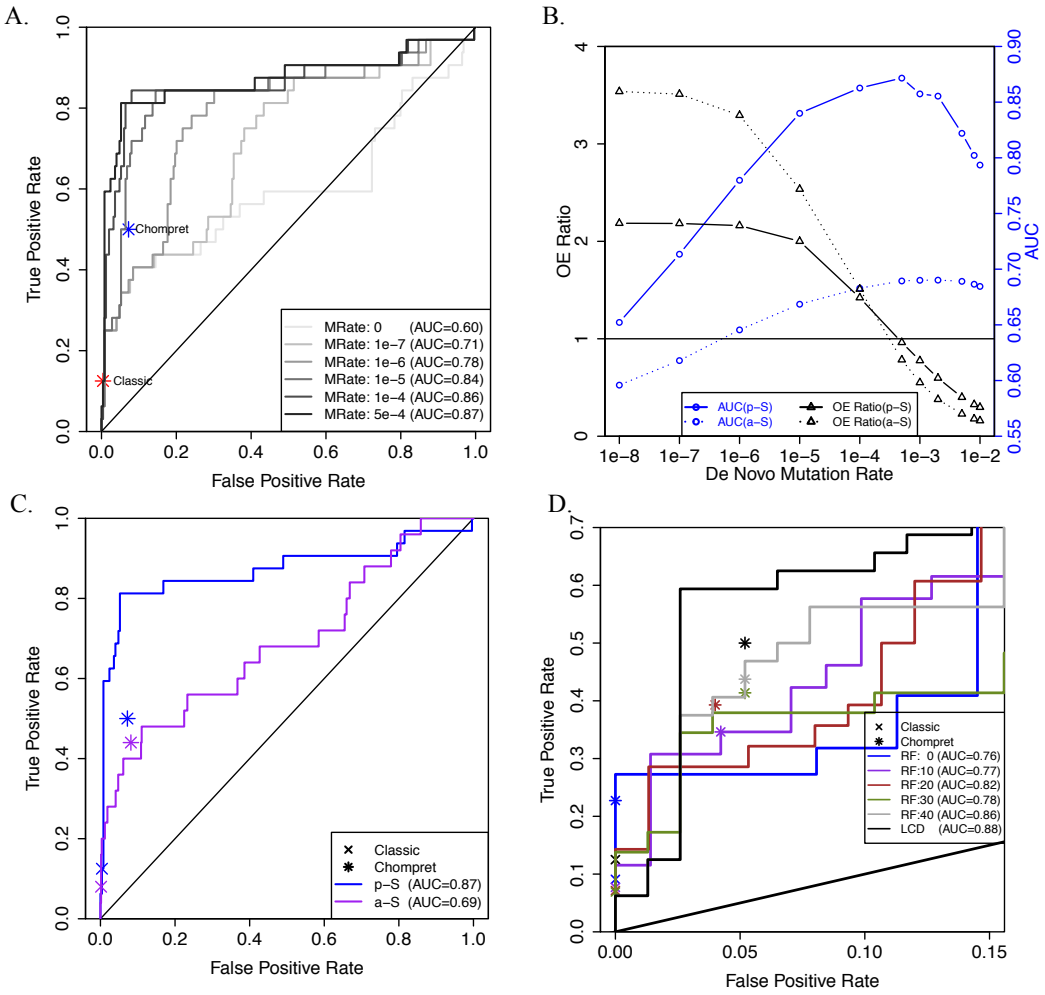


Figure 5.3. Validation results. A) Receiver operating characteristic (ROC) curves of LFSpro for data from 183 families collected through the pediatric sarcoma cohort, with de novo mutation rate changing from 0 to 0.0005. Also shown are results of applying the classic LFS and Chompret criteria to this data set. B) Influence of de novo mutation rate on observed/expected (O/E) ratio and area under the receiver operating characteristic curve (AUC). p-S: pediatric-sarcoma families. a-S: adult-sarcoma families. C) ROC curves of LFSpro for three data sets when we set the de novo mutation rate as 0.0005. Also shown are true positive and false positive rate of classic LFS and Chompret criteria for the three data sets. D) Part of the ROC curves of different roll forward time. RF: roll forward. LCD: last contact date.

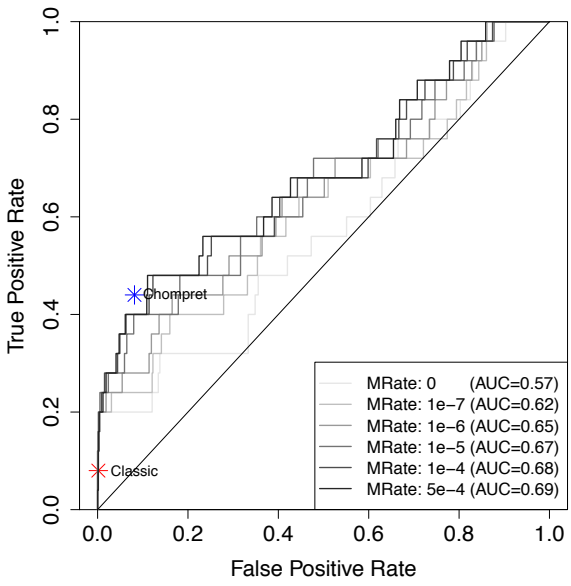


Figure 5.4 ROC curves of LFSpro for adult-sarcoma families with de novo mutation rate changing from 0 to 0.0005.

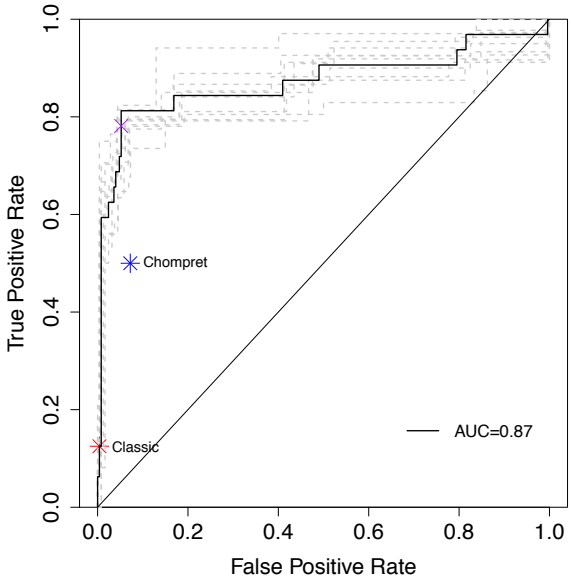


Figure 5.5 ROC curve of LFSpro with de novo mutation rate of 0.0005 for pediatric-sarcoma families. Variability is shown by 10 bootstraps. Also shown are true positive and false negative rate of classic LFS and Chompret criteria. The purple x indicates the cut-off we used to distinguish *TP53*-positive from *TP53*-negative individuals in LFSpro (posterior probability of TP53 mutations: 0.2).

5.3.3 Validation Result. Figure 5.3C presents the results of LFSpro with the results of both classic LFS and Chompret criteria for two data sets. The classic LFS criteria are very conservative. The false positive rate is almost 0, but its sensitivity is not very high. The sensitivity is only around 10% for the two data sets. The Chompret criteria have much higher sensitivity but also a relatively higher false positive rate. For pediatric-sarcoma families, both the points of the classic LFS and Chompret criteria are under or on the ROC curve, indicating that LFSpro performs better than the classic LFS and Chompret criteria. For the adult-sarcoma families, the point of the Chompret criteria is a little higher than the ROC curve of LFSpro. Table 5.3 summarizes AUC and OE estimates by LFSpro and the 95% confidence intervals.

Table 5.3 Summary of validation results

	AUC	OE
p-Sarcoma	0.87(0.77,0.95)	0.96(0.74,1.20)
a-Sarcoma	0.69(0.57,0.81)	0.78(0.51,1.12)

Table 5.4 Reclassification of *TP53* mutation carriers using LFSpro.

	LFSpro	Carriers		Noncarriers		
		Pr≥0.2	Pr≥0.2	Pr≥0.2	Pr≥0.2	Reclassification rate[c]
p-Sarcoma	Classic					
	+	4	0[b]	0	1[a]	
	-	21[a]	7	13[b]	235	3.20%
	Chompret					
	+	15	1[b]	3	15[a]	
	-	10[a]	6	10[b]	221	4.98%
a-Sarcoma	Classic					
	+	2	0[b]	1	0[a]	
	-	8[a]	15	39[b]	548	-5.06%
	Chompret					
	+	7	4[b]	11	37[a]	
	-	3[a]	11	29[b]	511	1.14%

[a] LFSpro corrected classifications of carrier or noncarrier, as compared to clinical criteria.

[b] The clinical criteria corrected classifications of carrier or noncarrier, as compared to LFSpro.

[c] Reclassification rate = (total # of correcte reclassification - total # of incorrecte reclassification) / total # events

We also compared 3 different methods in the reclassification table (Table 5.4). We used 0.2, which is the turning point in the ROC curve, as the cutoff to classify *TP53*-positive and *TP53*-negative individuals in LFSpro (Figure 5.5). The reclassification rate was defined as the percentage of individuals whose classification by LFSpro differed from their classification by classic LFS or Chompret criteria. A positive value means that LFSpro makes a net improvement. For pediatric-sarcoma families, LFSpro performs better than both classic LFS and Chompret criteria. For adult-sarcoma families, LFSpro is better than the Chompret criteria, while the reclassification rate is negative when comparing to the classic LFS criteria. However, the sensitivity for the classic LFS criteria is too low (about 10%) which makes it difficult to detect *TP53* mutations carriers in the population using the classic LFS criteria only.

5.3.4 Roll Forward. Figure 5.3D shows how the results of LFSpro and the other two clinical criteria changed when we rolled forward the data for *TP53*-postive pediatric-sarcoma families. The false positive rate for the classic LFS criteria is always 0. The sensitivity of classic LFS criteria decreased initially and reached the lowest point when we rolled forward 30 years and then improved when we rolled forward 40 years (The true positive rate and false positive rate are the same for rolling forward 40 years and the last contact date). The sensitivity of the Chompret criteria increased smoothly as we had more family information. The AUC for LFSpro increased at first and then dropped when rolling forward 30 years and then increased again until the last contact date. At the 20 year and 30 year time point, Chompret criteria performed better than LFSpro. And at all other rolling forward time points, LFSpro performs better or as well as the Chompret criteria.

5.4 Discussion

Our validation showed that in most cases LFSpro had both higher sensitivity and higher specificity than both the classic LFS and Chompret criteria, indicating that LFSpro provides a better method to identify the *TP53* mutation carriers.

Unlike the classic LFS and Chompret criteria, which only take into account the information about the counselee's first and second degree relatives with cancer, LFSpro uses information from all family members. The information from the family members with cancers helped LFSpro to predict *TP53* mutation carriers with high sensitivity, while the information about healthy family members prevented LFSpro from being associated with a high false positive rate. This was evident when we added *TP53* negative families into the validation population. In many *TP53* negative families, the proband and close relatives had early onset of LFS-spectrum cancers, a significant predictor of *TP53* mutations. This explains why the Chompret criteria classified the probands in such families as *TP53* mutation carriers in error. In contrast, the inclusion of healthy family members (especially the parents) helped LFSpro to avoid these false positives.

Whereas the classic LFS and Chompret criteria provide a binary result—i.e., to test or not to test for germline *TP53* mutations—LFSpro provides a probability that a certain individual carries a germline *TP53* mutation. We can change the cut-off for the probability to balance sensitivity and specificity. Another limitation of the classic LFS and Chompret criteria is prediction of *TP53* mutation carrier status only when the individual has had a LFS spectrum cancer. If the individual is healthy, the prediction of the classic LFS and Chompret criteria will always be non-carrier. However, LFSpro can give an estimation of the carrier probability even if the individual is healthy.

The prevalence of *de novo TP53* germline mutations have been reported to be at least 7% to as high as 20%[41], which is more substantial than previously appreciated. We found that when we

incorporated the de novo mutation rate into our model, the predictive power of the model improved substantially. AUC increased from 0.60 to 0.71 for pediatric-sarcoma families (Figure 5.3A) and from 0.57 to 0.62 for all adult-sarcoma families (Figure 5.4) even though the mutation rate only increased from 0 to 10^{-7}. The hypothetical illustration shows that there is not significant difference to incorporate *de novo* mutation when it is a Mendelian transmission. However, when there is negative family history or a true *de novo* mutation occurs, a big increase in carrier probability occurs, which underscores the importance of *de novo* during *LFS analysis.* According to the validation population of our study, 0.0005 seems to be a good estimation of the *de novo* mutation rate in LFS families.

In the adult-sarcoma data set, the Chompret criteria is a little better than LFSpro (Figure 5.3C). This may be due to the fact that Chompret criteria can identify a *TP53* mutation carrier using only the information from that individual, rather than depending on strong family cancer history. However, this also allows for the Chompret criteria to result in many false-positives. In the *TP53* positive adult-sarcoma families, there are 7 families with only 1 or 2 individuals with cancers. When we removed all these individuals and only considered the adult-sarcoma families with more than 2 individuals with cancers, LFSpro performed better than the Chompret criteria. Therefore, when family information is limited (< 3 individuals with cancer in the family), we can use the Chompret criteria to ensure that we do not miss any *TP53* mutation carriers. In contrast, when family information is sufficient, it is better to use LFSpro to identify *TP53* mutation carriers since LFSpro has higher sensitivity and specificity than the Chompret criteria.

Another reason LFSpro does not perform as well as Chompret criteria is because in LFSpro, if a patient has more than 1 cancer, we consider only the first cancer in the calculation. However, in the Chompret criteria, one patient with multiple primary cancers is a very important criterion for identifying *TP53* mutations. As previously reported, early onset of multiple LFS-spectrum

cancers is one of the characteristics of LFS[83] Therefore, we plan to incorporate information about multiple primary cancers in a single individual into our model.

In addition, in the roll-forward study of the *TP53*-positive pediatric-sarcoma families, LFSpro did not perform as well as the Chompret criteria when there was not enough family information. The AUC of LFSpro increased first and then there was a drop at 30 years and then increased again (Figure 5.3D). This can be explained by the observation that when we rolled forward the years, we had more information about the existing family members and therefore we could make the estimation more accurately. After about 20 to 30 years a new generation was born, with most of them being healthy at first, which reduced the probability that they are *TP53* mutation carriers and this in turn reduced the probability that their first degree relatives are *TP53* mutation carriers. And during this period, some *TP53* mutation carriers developed multiple cancers. The Chompret criteria used the information to improve the prediction.

In summary, with the advent of population and clinic-based ascertainment of mutations in *TP53* in the context of next generation screening strategies, new methods that incorporate *de novo* mutations will become increasingly relevant clinically. To address this need, we developed LFSpro and validated it with both *TP53*-positive and *TP53*-negative families. LFSpro provides an accurate estimate of *TP53* mutation carriers on the basis of family history and *de novo* mutation rates. With more accurate identification of *TP53* mutation carriers, we can initiate appropriate screening and health management thereby reducing disease burden and cancer mortality.

6. Conclusions and Future Research

6.1 Conclusions

In this dissertation, we focused on developing methods to improve the accuracy of germline mutation detection in next generation sequencing data and estimating the *TP53* mutation carrier probability in families with Li-Fraumeni syndrome. We investigated the factors that influence the performance of our methods and validate the methods in both simulation and real data. The results show that the new methods improve the accuracy comparing to the existing methods.

We first introduced the method, FamSeq, using pedigree information to improve the variant calling accuracy in Chapter 2. In simulation and real data, FamSeq reduces both false positive and false negative rate. The improvement level depends on the read depth, pedigree structure, pedigree size and the position the individual in the pedigree. Sometimes, sequencing the individuals in a family with moderate read depth gives more and better results than sequencing one individual with high read depth. It is better if we can sequence large families. When large families are not available, the simulation and real data show that sequencing a family with 7 individuals of three generations has big improvement in variant calling accuracy.

In family-base sequencing analysis, pedigree information data is usually very complex. One method cannot meet the requirements all the time. We implement four different methods in FamSeq to avoid these problems. We compared the difference among these four methods in Chapter 3. We focused on the application of GPU to improve the computing speed. When family size is relatively large, GPU has 10-fold improvement comparing to CPU. In next generation sequencing analysis, the computing tasks are usually homogeneous. It is very suitable for GPU since it has much more computing cores than CPU.

In Chapter 4, we introduced another method to improve the germline mutation detection in next generation sequencing data. The new method used the Bayesian hierarchical model to

combine the data from multiple platforms to reduce variant calling error by using the information from other platforms when one platform failed. In the application of the methods in the two most popular sequencing platforms, variant calling error rate decreased about 25% percent when we have enough training data. This method can also be applied to the sequencing data generated from old platforms. We can use the old sequencing data with the new data to improve the variant calling accuracy. We can also combine the old sequencing data together to get the result has similar accuracy rate as the new data, when some individuals are sequenced multiple times by old platforms, such as NA12878. It will save a lot time and money without sequencing these individuals again.

We also applied the pedigree information to estimate the *TP53* mutation carrier probability. Comparing to the existing clinical criteria, our method, LFSpro, gives an estimation of probability instead of a binary result. With the probability, we can change the cut-off to balance the sensitivity and specificity. The de novo mutation plays an important role in Li-Fraumeny syndrome. We incorporated de novo mutation rate into the Mendelian risk prediction model and estimate the de novo mutation rate in two different data sets.

6.2 Future Research

A key character in Li-Fraumeni syndrome is multiple primary cancers. In LFSpro, we only consider the earliest primary cancer, which will lose a lot useful information. It partially explained the reason why LFSpro does not perform better than the Chompret criteria in adult sarcoma data set. Another reason is the penetrance used in LFSpro is estimated from the population in the United States, while the adult sarcoma data is collected in Australia. The penetrance might be very different between the two countries. To improve the performance of LFSpro, we should build a new model to consider multiple primary cancers. On the other hand,

we can collect more Li-Fraumeni syndrome family data from Australia and use these data to estimate the penetrance specific for Australia.

We need training data with true variants to estimate the parameters to build the model combining the data sets form multiple platforms to call the variant with next generation sequencing data. However, we cannot get training data for most data sets. It cost a lot of money and time to generate true variants in next generations sequencing data. In future, I will develop a method that does not need the training data for better application on the basis of current model.

How to integrating different biology data together is a difficult and popular problem in biology research recently. The model we build to combine sequencing data from multiple platforms can be adapted to integrate different biology data. For example, we can predict cancer stage from gene expression and MRI image by calculating the posterior probability of Pr(Cancer Stage | Gene Expression, MRI image).

Bibliography

1 Network, T. C. G. A. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690, doi:10.1016/j.cell.2014.09.050 (2014).

2 Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J. & Bamshad, M. J. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-35, doi:10.1038/ng.499 (2010).

3 Victoria, J. G., Wang, C., Jones, M. S., Jaing, C., McLoughlin, K., Gardner, S. & Delwart, E. L. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *Journal of virology* **84**, 6033-6040, doi:10.1128/jvi.02690-09 (2010).

4 Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K. A., Hosoda, F., Nguyen, H. H., Aoki, M., Hosono, N., Kubo, M., Miya, F., Arai, Y., Takahashi, H., Shirakihara, T., Nagasaki, M., Shibuya, T., Nakano, K., Watanabe-Makino, K., Tanaka, H., Nakamura, H., Kusuda, J., Ojima, H., Shimada, K., Okusaka, T., Ueno, M., Shigekawa, Y., Kawakami, Y., Arihiro, K., Ohdan, H., Gotoh, K., Ishikawa, O., Ariizumi, S., Yamamoto, M., Yamada, T., Chayama, K., Kosuge, T., Yamaue, H., Kamatani, N., Miyano, S., Nakagama, H., Nakamura, Y., Tsunoda, T., Shibata, T. & Nakagawa, H. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nature genetics* **44**, 760-764, doi:10.1038/ng.2291 (2012).

5 Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

6 Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:10.1038/nature03959 (2005).

7 Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

8 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu,

A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. & Smith, A. J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

9 Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A. P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research* **36**, e25, doi:10.1093/nar/gkn021 (2008).

10 Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, doi:10.1038/nature10242 (2011).

11 Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology* **9**, e1003031, doi:10.1371/journal.pcbi.1003031 (2013).

12 Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).

13 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

14 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

15 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

16 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

17 Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. & Wang, J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967, doi:10.1093/bioinformatics/btp336 (2009).

18 Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L. & Galas, D. J. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639, doi:10.1126/science.1186802 (2010).

19 Zhou, B. & Whittemore, A. S. Improving sequence-based genotype calls with linkage disequilibrium and pedigree information. *The Annals of Applied Statistics* **6**, 457-475 (2012).

20 Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E. & Awadalla, P. Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-714, doi:10.1038/ng.862 (2011).

21 Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H. M. & Abecasis, G. R. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS genetics* **8**, e1002944, doi:10.1371/journal.pgen.1002944 (2012).

22 Fishelson, M. & Geiger, D. Exact genetic linkage computations for general pedigrees. *Bioinformatics* **18 Suppl 1**, S189-198 (2002).

23 Elston, R. C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Human heredity* **21**, 523-542 (1971).

24 Biswas, S. & Berry, D. A. Determining joint carrier probabilities of cancer-causing genes using Markov chain Monte Carlo methods. *Genetic epidemiology* **29**, 141-154, doi:10.1002/gepi.20082 (2005).

25 Lin, S., Thompson, E. & Wijsman, E. An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees. *Annals of human genetics* **58**, 343-357 (1994).

26 Stricker, C., Fernando, R. & Elston, R. An algorithm to approximate the likelihood for pedigree data with loops by cutting. *Theoretical and Applied Genetics* **91**, 1054-1063 (1995).

27 Cannings, C., Thompson, E. & Skolnick, M. Probability functions on complex pedigrees [domesticated mammals, laboratory animals]. *Advances in Applied Probability* (1978).

28 van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* **30**, 418-426, doi:10.1016/j.tig.2014.07.001 (2014).

29 Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46, doi:10.1038/nrg2626 (2010).

30 Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology* **2012**, 251364, doi:10.1155/2012/251364 (2012).

31 Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: promises and challenges. *Cancer letters* **340**, 284-295, doi:10.1016/j.canlet.2012.11.025 (2013).

32 Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. & Salit, M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32**, 246-251, doi:10.1038/nbt.2835 (2014).

33 Hwang, S. J., Lozano, G., Amos, C. I. & Strong, L. C. Germline p53 mutations in a cohort with childhood sarcoma: sex differences in cancer risk. *American journal of human genetics* **72**, 975-983, doi:10.1086/374567 (2003).

34 Wu, C. C., Shete, S., Amos, C. I. & Strong, L. C. Joint effects of germ-line p53 mutation and sex on cancer risk in Li-Fraumeni syndrome. *Cancer research* **66**, 8287-8292, doi:10.1158/0008-5472.can-05-4247 (2006).

35 Wu, C. C., Strong, L. C. & Shete, S. Effects of measured susceptibility genes on cancer risk in family studies. *Human genetics* **127**, 349-357, doi:10.1007/s00439-009-0774-y (2010).

36 Olivier, M., Goldgar, D. E., Sodha, N., Ohgaki, H., Kleihues, P., Hainaut, P. & Eeles, R. A. Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. *Cancer research* **63**, 6643-6650 (2003).

37 Li, F. P., Fraumeni, J. F., Jr., Mulvihill, J. J., Blattner, W. A., Dreyfus, M. G., Tucker, M. A. & Miller, R. W. A cancer family syndrome in twenty-four kindreds. *Cancer research* **48**, 5358-5362 (1988).

38 Chompret, A., Abel, A., Stoppa-Lyonnet, D., Brugieres, L., Pages, S., Feunteun, J. & Bonaiti-Pellie, C. Sensitivity and predictive value of criteria for p53 germline mutation screening. *Journal of medical genetics* **38**, 43-47 (2001).

39 Tinat, J., Bougeard, G., Baert-Desurmont, S., Vasseur, S., Martin, C., Bouvignies, E., Caron, O., Bressac-de Paillerets, B., Berthet, P., Dugast, C., Bonaiti-Pellie, C., Stoppa-Lyonnet, D. & Frebourg, T. 2009 version of the Chompret criteria for Li Fraumeni syndrome. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, e108-109; author reply e110, doi:10.1200/jco.2009.22.7967 (2009).

40 Chen, S., Wang, W., Broman, K. W., Katki, H. A. & Parmigiani, G. BayesMendel: an R environment for Mendelian risk prediction. *Statistical applications in genetics and molecular biology* **3**, Article21, doi:10.2202/1544-6115.1063 (2004).

41 Gonzalez, K. D., Buzin, C. H., Noltner, K. A., Gu, D., Li, W., Malkin, D. & Sommer, S. S. High frequency of de novo mutations in Li-Fraumeni syndrome. *Journal of medical genetics* **46**, 689-693, doi:10.1136/jmg.2008.058958 (2009).

42 Siva, N. 1000 Genomes project. *Nature biotechnology* **26**, 256, doi:10.1038/nbt0308-256b (2008).

43 Glazov, E. A., Zankl, A., Donskoi, M., Kenna, T. J., Thomas, G. P., Clark, G. R., Duncan, E. L. & Brown, M. A. Whole-exome re-sequencing in a family quartet identifies POP1 mutations as the cause of a novel skeletal dysplasia. *PLoS genetics* **7**, e1002027, doi:10.1371/journal.pgen.1002027 (2011).

44 Koenekoop, R. K., Wang, H., Majewski, J., Wang, X., Lopez, I., Ren, H., Chen, Y., Li, Y., Fishman, G. A., Genead, M., Schwartzentruber, J., Solanki, N., Traboulsi, E. I.,

Cheng, J., Logan, C. V., McKibbin, M., Hayward, B. E., Parry, D. A., Johnson, C. A., Nageeb, M., Poulter, J. A., Mohamed, M. D., Jafri, H., Rashid, Y., Taylor, G. R., Keser, V., Mardon, G., Xu, H., Inglehearn, C. F., Fu, Q., Toomes, C. & Chen, R. Mutations in NMNAT1 cause Leber congenital amaurosis and identify a new disease pathway for retinal degeneration. *Nature genetics* **44**, 1035-1039, doi:10.1038/ng.2356 (2012).

45 Roach, J. C., Glusman, G., Hubley, R., Montsaroff, S. Z., Holloway, A. K., Mauldin, D. E., Srivastava, D., Garg, V., Pollard, K. S., Galas, D. J., Hood, L. & Smit, A. F. Chromosomal haplotypes by genetic phasing of human families. *American journal of human genetics* **89**, 382-397, doi:10.1016/j.ajhg.2011.07.023 (2011).

46 Pearl, J. *Causality*. (Cambridge university press, 2009).

47 Shen, P., Wang, W., Krishnakumar, S., Palm, C., Chi, A. K., Enns, G. M., Davis, R. W., Speed, T. P., Mindrinos, M. N. & Scharfe, C. High-quality DNA sequence capture of 524 disease candidate genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6549-6554, doi:10.1073/pnas.1018981108 (2011).

48 McDonald, J. M., Douglass, E. C., Fisher, R., Geiser, C. F., Krill, C. E., Strong, L. C., Virshup, D. & Huff, V. Linkage of familial Wilms' tumor predisposition to chromosome 19 and a two-locus model for the etiology of familial tumors. *Cancer research* **58**, 1387-1390 (1998).

49 Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L.,

Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D. & McEwen, J. E. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

50 Wang, W., Shen, P., Thiyagarajan, S., Lin, S., Palm, C., Horvath, R., Klopstock, T., Cutler, D., Pique, L., Schrijver, I., Davis, R. W., Mindrinos, M., Speed, T. P. & Scharfe, C. Identification of rare DNA variants in mitochondrial disorders with improved array-based sequencing. *Nucleic acids research* **39**, 44-58, doi:10.1093/nar/gkq750 (2011).

51 Scharfe, C., Lu, H. H., Neuenburg, J. K., Allen, E. A., Li, G. C., Klopstock, T., Cowan, T. M., Enns, G. M. & Davis, R. W. Mapping gene associations in human mitochondria using clinical disease phenotypes. *PLoS computational biology* **5**, e1000374, doi:10.1371/journal.pcbi.1000374 (2009).

52 Stricker, C., Fernando, R. L. & Elston, R. C. Linkage analysis with an alternative formulation for the mixed model of inheritance: the finite polygenic mixed model. *Genetics* **141**, 1651-1656 (1995).

53 Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. & Sonstegard, T. S. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods* **5**, 247-252, doi:10.1038/nmeth.1185 (2008).

54 Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome research* **21**, 940-951, doi:10.1101/gr.117259.110 (2011).

55 Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics* **12**, 443-451, doi:10.1038/nrg2986 (2011).

56 Peng, G., Fan, Y., Palculict, T. B., Shen, P., Ruteshouser, E. C., Chi, A. K., Davis, R. W., Huff, V., Scharfe, C. & Wang, W. Rare variant detection using family-based sequencing analysis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 3985-3990, doi:10.1073/pnas.1222158110 (2013).

57 Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A. & Conrad, D. F. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature methods* **10**, 985-987, doi:10.1038/nmeth.2611 (2013).

58 Buckner, J., Wilson, J., Seligman, M., Athey, B., Watson, S. & Meng, F. The gputools package enables GPU computing in R. *Bioinformatics* **26**, 134-135, doi:10.1093/bioinformatics/btp608 (2010).

59 Schatz, M. C., Trapnell, C., Delcher, A. L. & Varshney, A. High-throughput sequence alignment using Graphics Processing Units. *BMC bioinformatics* **8**, 474, doi:10.1186/1471-2105-8-474 (2007).

60 Zandevakili, P., Hu, M. & Qin, Z. GPUmotif: an ultra-fast and energy-efficient motif analysis program using graphics processing units. *PloS one* **7**, e36865, doi:10.1371/journal.pone.0036865 (2012).

61 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G. & Durbin, R. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

62 Hardy, G. H. Mendelian proportions in a mixed population. *Science* **28**, 49-50 (1908).

63 Totir, L. R., Fernando, R. L. & Abraham, J. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics, selection, evolution : GSE* **41**, 52, doi:10.1186/1297-9686-41-52 (2009).

64 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8817-8822, doi:10.1073/pnas.1133470100 (2003).

65 Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34**, e22, doi:10.1093/nar/gnj023 (2006).

66 Buermans, H. P. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* **1842**, 1932-1941, doi:10.1016/j.bbadis.2014.06.015 (2014).

67 Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J Genet Genomics* **38**, 95-109, doi:10.1016/j.jgg.2011.02.003 (2011).

68 Allenby, G. M., Rossi, P. E. & McCulloch, R. E. Hierarchical Bayes models: a practitioners guide. *Available at SSRN 655541* (2005).

69 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

70 Li, F. P. & Fraumeni, J. F., Jr. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Annals of internal medicine* **71**, 747-752 (1969).

71 Li, F. P. & Fraumeni, J. F., Jr. Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome. *Journal of the National Cancer Institute* **43**, 1365-1373 (1969).

72 Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Jr., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A. & et al. Germ line p53

mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233-1238 (1990).

73 Villani, A., Tabori, U., Schiffman, J., Shlien, A., Beyene, J., Druker, H., Novokmet, A., Finlay, J. & Malkin, D. Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: a prospective observational study. *The lancet oncology* **12**, 559-567, doi:10.1016/s1470-2045(11)70119-x (2011).

74 Birch, J. M., Hartley, A. L., Tricker, K. J., Prosser, J., Condie, A., Kelsey, A. M., Harris, M., Jones, P. H., Binchy, A., Crowther, D. & et al. Prevalence and diversity of constitutional mutations in the p53 gene among 21 Li-Fraumeni families. *Cancer research* **54**, 1298-1304 (1994).

75 Eeles, R. A. Germline mutations in the TP53 gene. *Cancer surveys* **25**, 101-124 (1995).

76 Wang, W., Chen, S., Brune, K. A., Hruban, R. H., Parmigiani, G. & Klein, A. P. PancPRO: risk assessment for individuals with a family history of pancreatic cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **25**, 1417-1422, doi:10.1200/jco.2006.09.2452 (2007).

77 Wang, W., Niendorf, K. B., Patel, D., Blackford, A., Marroni, F., Sober, A. J., Parmigiani, G. & Tsao, H. Estimating CDKN2A carrier probability and personalizing cancer risk assessments in hereditary melanoma using MelaPRO. *Cancer research* **70**, 552-559, doi:10.1158/0008-5472.can-09-2653 (2010).

78 Gonzalez, K. D., Noltner, K. A., Buzin, C. H., Gu, D., Wen-Fong, C. Y., Nguyen, V. Q., Han, J. H., Lowstuter, K., Longmate, J., Sommer, S. S. & Weitzel, J. N. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1250-1256, doi:10.1200/jco.2008.16.6959 (2009).

79 Strong, L. C. & Williams, W. R. The genetic implications of long-term survival of childhood cancer. A conceptual framework. *The American journal of pediatric hematology/oncology* **9**, 99-103 (1987).

80 Bondy, M. L., Lustbader, E. D., Strom, S. S. & Strong, L. C. Segregation analysis of 159 soft tissue sarcoma kindreds: comparison of fixed and sequential sampling schemes. *Genetic epidemiology* **9**, 291-304, doi:10.1002/gepi.1370090502 (1992).

81 Lustbader, E. D., Williams, W. R., Bondy, M. L., Strom, S. & Strong, L. C. Segregation analysis of cancer in families of childhood soft-tissue-sarcoma patients. *American journal of human genetics* **51**, 344-356 (1992).

82 Mitchell, G., Ballinger, M. L., Wong, S., Hewitt, C., James, P., Young, M. A., Cipponi, A., Pang, T., Goode, D. L., Dobrovic, A. & Thomas, D. M. High frequency of germline TP53 mutations in a prospective adult-onset sarcoma cohort. *PloS one* **8**, e69026, doi:10.1371/journal.pone.0069026 (2013).

83 Hisada, M., Garber, J. E., Fung, C. Y., Fraumeni, J. F., Jr. & Li, F. P. Multiple primary cancers in families with Li-Fraumeni syndrome. *Journal of the National Cancer Institute* **90**, 606-611 (1998).

VITA

Gang Peng is the son of Shufang Yan and Peiqing Peng. Gang was born in Nantong, Jiangsu, China where he graduated from Tongzhou High School in 2001. He entered Fudan University in Shanghai, China and obtained a Bachelor of Science degree in Physics in 2005. After graduation, he worked as a statistical analyst at School of Life Science in Fudan University for four years. In 2009 he left Shanghai for Houston to pursue a Ph.D. degree in Biostatistics and Bioinformatics in the University of Texas Graduate School of Biomedical Science at Houston. At the beginning of 2010, he joined Dr. Wenyi Wang's laboratory to prepare his Ph.D. dissertation in the Department of Bioinformatics and Computational Biology at the University of Texas MD. Anderson Cancer Center. He expects to get his Doctor of Philosophy Degree in Biostatistics and Bioinformatics in August 2015.