

**Statistical Methods and Analysis to
Identify Disease-Related Variants from
Genetics Studies**

by

Sai Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree
of Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2017

Doctoral Committee:

Professor Goncalo Abecasis, Chair
Assistant Professor Hyun Min Kang
Assistant Professor Jefferey M. Kidd
Associate Professor Jun Li
Professor Kerby Shedden

© Sai Chen 2017

saichen@umich.edu

ORCID: 0000-0003-3106-5643

To my love Bing

Acknowledgements

I would like to express my greatest gratitude to Dr. Goncalo Abecasis for his guidance in my research. When I meet difficulties in my research, his suggestions always opened a brand new view, and finally helped me solve the problem in a precise and simple way. He patiently gave me time and flexibility to try and to learn from such a broad range, and thanks to him, I finally found my true motivation and interests. I am so grateful to have such an understanding and brilliant mentor during my PhD. At the same time, I would like to thank Dr Hyun Min Kang. He guided me all the way through the GECCO project, and is always willing to teach me any new coding and analysis skills. I would like to thank my committee Dr Jun Li, Dr Jefferey Kidd and Dr Kerby Shedden. The rotations in their groups were my first step in Bioinformatics. I would like to thank Kerby Shedden for valuable suggestions on the statistical components in my thesis.

Apart from my committee members, I would like to thank all my collaborators. Without their contribution, I cannot complete the story. I would like to thank Mohammad Othman for helping me with experimental validation. I would like to thank Dr Ulrike Peters, who is the leader of GECCO, and Jeroen Huyghe, my collaborator and senior peer, and Yi Lin and Flora Qu, for collecting and preparing studies. It was always a great pleasure to meet them during the yearly

GECCO investigator meeting in Seattle. I would like to thank Jingjing Yang, Dajiang Liu, Xiaowei Zhan and Shuang Feng for their help and guidance in improving software Raremetal. I would like to thank people from Abecasis group, especially Alex Tsoi Lam, Scott Vrieze, Lars Frische, Goo Jun. I learnt so many from these passionate and smart scientists. I would like to express my appreciation for staffs in Center of Statistical Genetics (CSG), especially Laura Baker, Irene Felicetti, Sean Caron, Mary Kate Wing, Chris Schellner. Without their technical support, I cannot finish my work. I would like to thank Kirsten Harold, for proofreading my thesis all the way through.

I would like to thank all my friends at University of Michigan. I am very grateful to have Shiya Song, Xuefang Zhao, Shweta Ramdas as my peers and classmates. We helped each other through so many deadlines and homework. I would like to thank all my friends here at Ann Arbor. They make my five years here so happy and colorful.

I would like to thank my parents for raising me up. Although we had discrepancies in some aspects, I believe now they are very proud of me and happy to know I am having the life that I exactly want.

At last, I would like to thank my fiancé, and my soul mate, Bing Yang, for his love, support and inspirations. For the past seven years, everyday is so beautiful as the first day we met. I am looking forward to our new life at the West Coast.

TABLE OF CONTENTS

Dedication.....	ii
Acknowledgements	iii
List of Figures	viii
List of Tables.....	x
List of Abbreviations	xii
Abstract.....	xiii
CHAPTER	
I. Introduction	1
Advances and challenges in modern genetics studies.....	1
Genomic variant detection.....	2
Implications from sequencing studies	3
Associations and causality	4
Meta-analysis and existing problems	5
Outline of this thesis	6

II. LIME: a likelihood based Mobile Element Insertion detection for sequencing data.....	12
Introduction.....	12
Method	13
Results	18
Discussion	22
Figures	24
Tables.....	27
Supplementary Text	33
Supplementary Figures and Tables	37
III. Discovery of novel variants associated with colorectal cancer	44
Introduction.....	44
Materials and Methods	46
Results	50
Discussion	56
Figures and Tables.....	58
Supplementary Figures	64
Supplementary Tables	79
IV. Improvements to the meta-analysis software RAREMETAL.....	104
Introduction.....	104
Methods.....	107

Results	114
Discussion	116
Figures and Tables.....	119
V. Summary and Discussion.....	126
Summary	126
Future Directions	128
Conclusion.....	130

List of Figures

Figure 2.1 MEI detection power in simulated data across four methods (100 bp pair-ended reads)	24
Figure 2.2 Breakpoint detection accuracy of four methods in simulated data	25
Figure 2.3 Log-scale allele frequency spectrums of SNPs and MEIs in samples from Sardinia Whole-genome Sequencing Project (493 samples)	26
Supplementary Figure 2.1 MEI false discovery rate (FDR) in simulated data (100 bp pair-ended reads)	43
Supplementary Figure 2.2 MEI detection power in simulated dataset (75 bp pair-ended reads)	43
Supplementary Figure 2.3 MEI false discovery rate (FDR) in simulated dataset (75 bp pair-ended reads)	43
Figure 3.1 LocusZoom on the two regions with cluster of SNPs of $P < 5 \times 10^{-7}$	58
Figure 3.2 Sequencing characteristics of GECCO WGS samples	60
Supplementary Figure 3.1 Ancestry inference of the 26,903 individuals	64
Supplementary Figure 3.2 QQ plot of meta-analysis p values across all 22 million imputed variants	65

Supplementary Figure 3.3 QQ plot for single variant association analysis on each genotyping platform	66
Supplementary Figure 3.4 Manhattan plot of single-variant meta-analysis	67
Supplementary Figure 3.5 Manhattan plot of the gene-based test	68
Supplementary Figure 3.6 LocusZoom plots on the rest four regions with $P < 5 \times 10^{-7}$	69
Supplementary Figure 3.7 LocusZoom plot of HAO1	73
Supplementary Figure 3.8 Distribution of HCT-116 deltaSVM scores	74
Supplementary Figure 3.9 eQTL expression of the nearby genes of the 4 variants with $P < 0.05$ at Colon Sigmoid	75
Supplementary Figure 3.10 Singleton and rare variant counts per individual	77
Figure 4.1 P values from the standard and the optimized method for unbalanced studies in simulated dataset	119
Figure 4.2 Burden test p values from the standard and the optimized covariance matrix	120

List of Tables

Table 2.1 The 18 pre-defined read pair types in LIME	27
Table 2.2 Average proportions of read pair types in windows of homozygous MEIs in NA12878 of 1000 Genome Project Phase 3	29
Table 2.3 Number and percentage of covered PCR-validated MEIs in 1000 Genome Phase 3 deep sequenced trio samples (NA12878, NA12891, NA12892)	30
Table 2.4 Parent-child inconsistency of MEIs in 1000 Genome Phase 3 deep sequenced trio samples (NA12878, NA12891, NA12892)	31
Table 2.5 Summary of sequencing characteristics in Sardinia Whole Genome Sequencing samples (n = 493).....	32
Supplementary Table 2.1 Numbers of detected MEIs in 1000 Genome Phase 3 deep sequenced samples (ALUs, L1s and SVAs).....	40
Table 3.1 Newly identified SNPs associated with colorectal cancer with P values less than 5×10^{-7}	61
Table 3.2 Summary of Year 3 variant calls	63
Supplementary Table 3.1 Study information.....	79
Supplementary Table 3.2 Meta-analysis results on previously identified CRC risk loci	83

Supplementary Table 3.3 SNPs included in the Burden test of HAO1	88
Supplementary Table 3.4 Parameter estimates from fGWAS model	89
Supplementary Table 3.5 SNPs and genome blocks with causality predicted by fGWAS.....	91
Supplementary Table 3.6 eQTL test on predicted DNaseI sensitive sites.....	93
Table 4.1 Power of VT test using different methods to combine scores statistics generated by LMM and GLMM.....	121
Table 4.2 Power from the standard bi-allelic analysis and the new multi-allelic analysis method on multi-allelic sites	122

List of Abbreviations

GWAS: Genome-wide association studies

NGS: Next generation sequencing

LD: Linkage Disequilibrium

MAF: minor allele frequency

bp: base pair

SNP: Single nucleotide polymorphism

Indel: Short insertions and deletions (usually <50 bp)

MEI: Mobile Element Insertions

HRC: Haplotype Reference Consortium

GECCO: Genetics and Epidemiology of Colorectal Cancer Consortium

CRC: colorectal cancer

LMM: Linear mixed model

GLMM: Generalized linear mixed model

Abstract

Advances in genotyping and sequencing technologies have greatly revolutionized the analytic methods in genetics research. Due to the dramatically decreasing per-genotype cost, millions of variants have been detected and genotyped from population-scale data. The findings provide a new insight into the human genome, and are continuously shaping our understanding of the genetic basis for disease. In this dissertation, I focus on three topics related to discovering disease-related variants in genetics studies in the aspects of method development and dataset analysis.

In chapter 2, I develop a likelihood-based method, LIME, to detect and genotype mobile element insertions (MEIs), a specific type of large insertions, from sequencing data. The method generates genotype likelihoods for each MEI using simulation that mimics the distribution of reads in regions with and without MEIs. From both simulated and real sequence data, our method shows better sensitivity than existing methods, especially in low-coverage data.

In chapter 3, I present genome-wide association studies and a whole-genome sequencing effort of discovering potentially novel loci for colorectal cancer. Using an imputation-based meta-analysis strategy, I replicate many previous findings

and provide a list of novel variants and genes for colorectal cancer. In collaboration with Fred Hutch Cancer Research Center, we additionally sequenced ~3,000 individuals and generated a variant call set. By incorporating gene annotation, sequence function prediction and online gene expression database, I highlight potentially functional loci for colorectal cancer in the known region 12q12 and the novel region 6q21.31. Although it is difficult to obtain new significant variants in the absence of extremely large dataset, our analysis provides some practical examples to incorporate functional genomics data into association analysis and to prioritize potentially functional candidates under limited sample size. Additionally, from the variant calling of whole-genome sequencing samples, we identified over 50 million variants, half of them being novel to the dbSNP database.

In chapter 4, I describe a major update to the meta-analysis software RAREMETAL that brings in software engineering improvements and several useful new methods for rare variant analysis. The engineering improvements make RAREMETAL more computationally efficient. The new methods in addition preserve the ability to meta-analysis in unbalanced studies, multi-allelic sites and generalized linear mixed models.

CHAPTER I

Introduction

Advances and challenges in modern genetics studies

In the past decades, advances in genotyping and sequencing technologies have greatly revolutionized the methodology in genetics research. Due to the dramatically decreasing cost of genotyping, millions of variants have been detected and genotyped from population-scale data. Detection of common variants and SNPs has been successful during the past several years. The findings provide a new insight into the human genome, and are continuously shaping our understanding of the genetic basis for disease. However, detecting rare and complex variants, such as structural variants, is still difficult, and the detection accuracy is insufficient for applying in clinics¹.

Accurate detection of genomic variants is the first step in discovering the mechanism of diseases. As known from population genetics theory, most mutations and polymorphisms are functionally neutral, with little to no effect on phenotypes². With thousands and even millions of novel variants discovered in a genetics study, genome-wide association studies (GWAS) have rapidly become a standard method for detecting disease-related variants and genes³. During the

past decades, much work has been done to appropriately model the statistical association between variants and phenotypes, but for more complicated situations, current methods still need to be improved to avoid potential power loss.

Genomic variant detection

In recent years, the rapid development of High-throughput sequencing (HTS) provides a new way of detecting genomic variants besides the genotyping array. HTS approaches require chopping genomic DNA into shorter size fragments (sequencing reads) and sequential detection of the nucleotide composition of each fragment thorough sequencing machines. Due to the huge size of the human genome and relatively short length of sequencing reads, de novo assembly requires massive amounts of computational resources and is in fact not possible in practice¹. Therefore, to identify the genomic location of reads, a common approach is to map them back to reference genome sequences

By comparing sequences between reads and the reference genome, variants are discovered from sequencing data. Since sequencing-based studies do not require pre-requisite information, it greatly facilitates discovery of rare and novel variants. The 1000 Genome Project has discovered over 88 million variants, with over half of them novel, while over 70% of them are rare variants with allele frequency less than 0.5%¹¹. In this dataset, the heterozygous genotype accuracy

is estimated around 99.4% for Single Nucleotide Polymorphisms (SNPs) and 99.0% for small insertion and deletions (indels).

With an accurate read aligner and appropriate modeling, most SNPs and indels will be reliably discovered by sequence composition of the reads and the reference genome{GenomesProjectConsortium:2015gk}. However, Structural Variants (SVs), including large-size insertions and deletions, still have much lower detection accuracy in sequencing data than SNPs and indels^{12,13}. These SVs are mostly longer than the lengths of currently used sequencing reads, and as a result, the alternative alleles cannot be reliably reconstructed, which in turn leads to loss of detection accuracy as well. Among these SVs, Mobile Element Insertions (MEIs) is an important category, with a few reports about its association with specific diseases{WoodsSamuels:1989ug}{Stewart:2011bt}. Because of shared sequences each family MEIs are easier to detect than novel insertions. Therefore, one chapter of this dissertation will focus on improving detection accuracy of MEIs from sequencing data.

Imputation and the reference panel

Large-scale sequencing studies provide a valuable resource of variants, and enable researchers to generate reference panels from the variant database. In genetics, imputation refers to the statistical inference of unobserved genotypes, achieved by utilizing information from known variants and their haplotypes¹⁴.

With a well-established reference panel, genotypes of uncovered sites in microarray data may be accurately imputed^{15,16}.

Thus imputation provides another efficient and economical method to study rare and novel variants under a limited budget. A proportion of samples from a study cohort could be sequenced to build an enriched reference panel with rare mutations together with the currently existing reference panel. Then, a larger study cohort, usually microarray data, could be imputed to this combined reference panel. In this way, those variant sites that are not directly genotyped will be imputed. Application of this scheme in 1000 Genome studies, and later the Haplotype Reference Consortium (HRC) based studies have greatly enlarged the power of detecting novel and rare disease-related variants^{17,18}.

Association and causality

Today, with the greatly reduced cost of genotyping and development of large-scale variant detection methods, association studies, which compare the frequency of alleles in a particular variant between affected and unaffected individuals, have become more powerful than traditional linkage analysis¹⁹. With millions of variants available, a genome-wide association approach surveys most of the genome for causal variants, representing an unbiased yet fairly comprehensive option that can be attempted even in the absence of convincing evidence regarding the function or location of the causal genes²⁰.

However, SNPs identified by GWAS are expected to tag genomic regions containing correlated SNPs with the trait. It remains an open question of how to identify true causal SNPs from these tagged genomic. Fine-mapping efforts incorporating functional genomics data to narrow down the range of potentially causal SNPs will be cost-effective for laboratory evaluation, especially for LD extensive regions with multiple independent SNPs. Additionally, when sample size is not enough for rare SNPs GWAS signals to reach genome-wide significant threshold, functional genomics data will be helpful to identify true causal SNPs from false positives.

Meta-analysis and existing problems

Recently, meta-analysis, which naturally incorporates cross-study heterogeneity, has been successful in many large-size and collaborated GWAS studies²¹. Meta-analysis is performed by combining summary statistics across studies and thus avoids the data privacy issues in some situations on sharing raw genotypes.

Powerful meta-analysis methods have been proposed, and in ideal situation, they may obtain equal power as the joint analysis^{22,23}. However, the underlying assumptions of these methods are violated in more real-life situations. For example, the basic meta-analysis method suffers substantial power loss when the case and control ratio is unbalanced. Current bi-allelic models cannot appropriately deal with multi-allelic sites, and results in false-negative

associations in such variants. Thus, more work needs to be done to improve meta-analysis methods to accommodate different types of data.

Outline of this thesis

In this dissertation, I focus on three topics related to discovering disease-associated variants in genetics studies in the aspects of both method development and dataset analysis. First, I describe a novel method for detecting Mobile Element Insertions (MEIs), a special type of complex variants, from sequencing data. Second, I present a genome-wide association analysis of colorectal cancer using imputed and/or sequenced data. Third, I present method improvements on more complex situations, and covariance matrices storage optimizations to the meta-analysis software, RAREMETAL, making it more powerful for complicated situations.

In chapter 2, I describe a likelihood-based method, LIME, to detect MEIs from sequencing data. The method naturally accommodates cross-sample heterogeneity, and generates genotype likelihood to measure the probability of each MEI event. From evaluation on both simulated data and deeply-sequenced samples 1000 Genome Phase 3, our method shows better performance than existing methods, especially in low-coverage data. By applying LIME to 493 samples from the Sardinia Whole Genome Sequencing Project, I identified 6,537 MEIs, in which 20 were predicted of having high impact on nearby gene expression levels.

In chapter 3, I present a genome-wide association analysis of colorectal cancer in 26,903 samples imputed on HRC reference panel. Using a meta-analysis strategy, we replicated many previous findings and provide a list of novel associated variants and genes for colorectal cancer. By incorporating gene annotation, sequence function prediction and online gene expression database, we highlight potentially functional loci for colorectal cancer. Our results indicate that even with a well-established reference panel and superior imputation quality, a larger sample size is necessary to discover rare disease-related variants.

Additionally, in chapter 3, as part of the collaboration effort within the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), we conducted a variant calling on 3,061 whole-genome sequenced individuals, aiming to generate a consensus call set of variants that are potentially more correlated with colorectal cancer. We generated a dataset of 48.7 million variants with superior quality, and nearly half of them have never been reported before.

In chapter 4, I describe a major update to the meta-analysis software RAREMETAL that brings in software engineering improvements and several useful new methods for rare variant analysis. The engineering improvements make RAREMETAL more computationally efficient. The new methods, developed by Dr Dajiang Liu and Dr Jingjing Yang, in addition preserve the ability

to meta-analysis in unbalanced studies, multi-allelic sites and generalized linear mixed models. With these improvements, RAREMETAL becomes more powerful in analyzing real datasets.

In chapter 5, I summarize my work and propose potential future directions in genetics studies.

References

1. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
2. Nei, M., Suzuki, Y. & Nozawa, M. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* **11**, 265–289 (2010).
3. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
4. Altshuler, D. *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26**, 76–80 (2000).
5. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
6. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
7. Guan, W., Pluzhnikov, A., Cox, N. J., Boehnke, M. International Type 2 Diabetes Linkage Analysis Consortium. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum. Hered.* **66**, 35–49 (2008).
8. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human

- disease. *Science* **322**, 881–888 (2008).
9. Kretowski, A., Ruperez, F. J. & Ciborowski, M. Genomics and Metabolomics in Obesity and Type 2 Diabetes. *J Diabetes Res* **2016**, 9415645–2 (2016).
 10. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
 11. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 12. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
 13. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
 14. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
 15. Porcu, E., Sanna, S., Fuchsberger, C. & Fritsche, L. G. Genotype imputation in genome-wide association studies. *Curr Protoc Hum Genet* **Chapter 1**, Unit 1.25–1.25.14 (2013).
 16. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
 17. Hoffmann, T. J. *et al.* Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. *PLoS Genet.* **11**, e1004930

(2015).

18. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
20. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
21. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
22. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
23. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).

CHAPTER II

LIME: a likelihood-based Mobile Element Insertion detector for sequencing data

Introduction

Mobile elements are DNA sequences that can insert themselves along the genome¹. It is estimated over 40% of the human genome consists of mobile elements or mobile element derived sequences². Only a few of the mobile elements in the genome are still mobilized; the remaining elements are slowly “decaying” through mutation^{1,3,4}.

Based on their consensus sequences, mobile elements are classified into multiple families. Among them, ALU, LINE1 and SINE/VNTR/ALU (SVA) are the most actively transcribed ones in the human genome¹. A human genome is estimated to harbor ~1,500 non-reference mobile element insertions (MEIs), including ~1200 ALUs, ~200 LINEs, and ~100 SVAs⁵. These MEIs can disrupt

The work presented in Chapter 2 will be submitted as Chen S, Othman M, Abecasis G. “LIME: a likelihood-based Mobile Element detector from sequencing data”.

normal gene function or even generate new genes. Some have been shown to cause diseases such as hemophilia A⁶, Crohn's disease⁷, and cancer^{8,9}.

Next generation sequencing has greatly facilitated the study of MEIs at both individual and population levels. With appropriate bioinformatics approaches, MEIs could be detected out of the great amount of sequencing reads. The most common method for detecting MEIs is to identify reads that are partially aligned to known mobile element sequences, and then identify regions of the genome where clusters of these reads (or their mate pairs) align¹⁰. Unfortunately, this method suffers substantial power loss when samples are sequenced at low coverage and does not easily accommodate heterogeneity in sequence data between samples¹¹.

Here, we present our likelihood-based MEI detector (LIME). LIME compares the distribution of mapped reads between regions with and without an MEI, and generates genotype likelihoods that naturally accommodate heterogeneity between samples in coverage, read mapping rates, and insert sizes. LIME does not use fixed thresholds when examining reads supporting each event; thus detection power in low coverage samples is greatly improved. By applying LIME to a subset of Sardina Whole Genome Sequencing Project¹², we generated a call set of MEIs and made predictions for their potential functions.

Method

Basic scheme

Consider an experiment where short paired-end reads are sequenced and mapped to a reference genome. After mapping, most read pairs map close together with an expected orientation and distance. However, for regions with an MEI, read pair arrangements might deviate from this pattern. For example, some read pairs may have one end partially or entirely mapped to a mobile element sequence elsewhere in the genome, which may result in abnormal insert size or strand orientation between the two ends. To systematically summarize read pair distributions in each region, we classify mapped read pairs into 18 categories based on insert sizes, orientation and mapping to consensus mobile element sequences (Table 1.1). In an ideal situation, for the genomic regions with no MEI or other genomic variation, all reads will be properly mapped with a normal insert size or strand orientation. For regions with MEIs, there will be an excessive number of other types of reads (Table 1.2).

Construct known genotypes to model read type distribution

Our LIME method tries to fully interpret the distribution of the 18 read-types in each region. This distribution is specific to each sample because it depends on characteristics of library preparation such as read length and insert size. LIME uses simulation to construct a sample specific model for the distribution of read types in regions with and without MEIs. These distributions are then used to calculate the genotype likelihood for each potential MEI.

To model the distribution of read-types in regions with an MEI, we construct a modified reference genome sequence where several reference ALUs, L1s and SVAs are spliced out (sections below will discuss selection criteria). Relative to this modified reference sequences, samples are expected to be homozygous for these MEIs, one in each spliced region. For each sample, we remap reads to this modified reference and summarize the distribution of the 18 read types in each of the spliced regions.

Next, we bin the modified reference genome into 600-bp windows (typically, we recommend setting this window size to be at least double the average insert size). We label each window i in this modified genome that overlaps an artificially constructed MEI, as $G_i = 1/1$; we label remaining windows $G_i = 0/0$. Reads within each window are classified into the 18 read types. Then, for each constructed window i , we define the total number of reads as n_i and the number of reads belonging to type r as n_{ir} . Then, this window i specific frequency of read type r is

$$f_{ir} = \frac{n_{ir}}{n_i}.$$

Sliding window analysis in real sample

In original mapping results, we slide a fixed size window along the genome, with window size the same as described in sections before. In each window, reads are classified into the 18 read types, too. For each window j , we define the total number of reads as c_j and the number of reads of type r as c_{jr} . We calculate

genotype likelihoods for each window j based on total R reads within it -- first by comparing the read distribution of window j to artificially constructed windows with homozygous MEIs, then to windows with reference genotypes, and finally to a composite of the two. For each underlying genotype, we iterate through all N_i constructed windows i in the modified genome, and define:

$$L(C_j | G_j = 0/0) = \frac{1}{N_i} \sum_{G_i=0/0} \prod_R f_{ir}^{c_{jr}}$$

$$L(C_j | G_j = 1/1) = \frac{1}{N_i} \sum_{G_i=1/1} \prod_R f_{ir}^{c_{jr}}$$

Specifically, to construct windows that contain heterozygous MEIs (which we term as $G_i = 0/1$), we match constructed windows i_0 with $G_i = 0/0$ and windows i_1 with $G_i = 1/1$ based on sequence mappability and GC content. Read counts in the two windows are proportionally adjusted so that total read counts of the two windows are the same. Then we define window specific frequency of read type r in constructed heterozygotes window i as:

$$f_{ir} = 0.5f_{i_0r} + 0.5f_{i_1r}$$

Therefore, genotype likelihood for heterozygotes is defined as:

$$L(C_j | G_j = 0/1) = \frac{1}{N_i} \sum_{G_i=0/1} \prod_R f_{ir}^{c_{jr}}$$

For the specific window j in the originally mapped data, the prior probability of genotype G_j , $P(G_j)$, is derived using Expectation-Maximization algorithm with or without assuming Hardy-Weinberg Equilibrium¹³. Posterior probability of genotype G_j is calculated under a Bayesian framework¹⁴:

$$P(G_j | Read_{1,2,...,n}) = \frac{P(G_j)P(Read_{1,2,...,n} | G_j)}{\sum_{G_j} P(G_j)P(Read_{1,2,...,n} | G_j)}$$

The genotype of window j is finalized with the highest posterior probability.

Note that genotype likelihood of those constructed homozygous MEI windows in the modified reference genome can be calculated under the same scheme, which is used as an estimate of MEI detection power in each sample.

When remapping reads to the mobile element consensus sequence, LIME additionally generates site-level features of each MEI, such as length, depth and number of supporting reads. MEI length is estimated as the distance between 3' and 5' mapped positions in their corresponding mobile element consensus sequence. In our preliminary analysis, we found that non-MEI genomic variants might also be identified as non-reference genotypes, since their read type distribution is somewhat different from regions with reference genotypes. In order to filter out such false positives, we applied the following empirical rules that are adapted from other methods^{15,16}: 1. Estimated MEI length should be longer than 100 base pairs. 2. MEI supporting read counts on either side of the window should be no more than 4 times of the other side.

Generate simulated dataset to evaluate LIME performance

We generated a simulated sequence by randomly inserting 2,000 ALU and 2,000 L1 consensus sequences into GRCh37 chromosome 17. L1 sequences are randomly truncated from 3-prime ends. Pair-ended Illumina Hi-seq 2000 reads

were generated against this simulated sequence using illumine read simulator ART¹⁷. These simulated reads were mapped back to the original hs37d5 reference sequence using mapper BWA MEM¹⁸. BamUtil¹⁹ was applied to mapping results to mark PCR duplicates and recalibrate mapping quality. LIME and three published MEI detection methods, RetroSeq¹⁵, Tangram²⁰ and Mobster¹⁶, were applied to this dataset, with details in Supplementary Text.

With the detection results from the four methods, MEI detection power for each of them is measured as the number of detected MEIs that are within 600 base pair range of simulated MEIs, divided by the total number of simulated MEIs. The false discovery rate is measured as the number of detected MEIs that are not within this range of simulated MEIs, divided by the total number of detected MEIs by this method.

Results

Performance on simulated data

In the simulated dataset with read lengths of 100 base pairs, LIME, along with RetroSeq, showed better detection power over the other two methods (Figure 2.1). LIME and RetroSeq had similar performance in detecting ALUs, with 60% power in detecting homozygotes in 2x coverage. ALU detection powers of these two methods were saturated at the coverage of 5x in homozygotes, which is equivalent to 10x in heterozygotes. Tangram ALU detection power was saturated at a higher coverage of 15x. In low-coverage data, LIME also has improved

detection power for L1s. In 2x data, LIME had higher power of 30% in heterozygotes and 50% in homozygotes, compared to RetroSeq's 20% in heterozygotes and 41% in homozygotes. Results from 75 bp simulated data shows similar pattern for the four methods (Figure S2.2). On the other hand, LIME had less than 1% false discovery rate in all simulated data (Figure S2.1, S2.3).

Breakpoint detecting accuracy is an important measurement of variant calling method performance as well^{20,21}. In 20x-simulated data, we calculated the distance between estimated and true MEI breakpoints across all four methods (Figure 2.2). For LIME, Tangram and Mobster, most distances were smaller than 20 bp in simulated dataset, indicating their good performance in refining MEI breakpoints.

Analysis on 1000G deep-sequenced trio samples

The 1000G Phase 3 deep-sequenced European trio samples (NA12878, NA12891, NA12892) have been widely used to measure performance of genomic variants detecting methods²². We applied LIME and Mobster to these three samples using their default settings. RetroSeq and Tangram detecting results of this dataset were downloaded from the RetroSeq Wiki page¹⁵.

Overall, the four methods had similar numbers of detected MEIs in each individual, with RetroSeq had fewer number of detected MEIs than the other

three methods (Table S2.1). To estimate the percentage of covered PCR validated events as a rough estimate of detection power, we downloaded the list of these samples' PCR validated MEIs²² from 1000 Genome ftp and lifted over their chromosome positions accordingly. We show that LIME (99.1%) and Mobster (99.4%) have the highest percentage of covered validated MEIs (Table 2.3).

At the same time, we used the parent-child inconsistent rate of MEIs to roughly evaluate the false discovery rate of our method (Table 2.4). LIME has the lowest overall discordance rate (7.0%) across all four methods, compared to RetroSeq (7.3%), Mobster (9.6%) and Tangram (11.6%). At the same time, LIME's L1 discordance rate (9.5%) is greatly improved over RetroSeq (11.4%), Mobster (23.3%) and Tangram (11.4%).

Experimental validation

We applied LIME to 20 samples from the Age-related Macular Degeneration (AMD) Whole Genome Sequencing Project. All samples consist of 100 bp pair-ended reads, with sequencing coverage ranges from 2x to 7x. Basically, 3,634 ALUs, 417 L1s and 31 SVAs were identified from the dataset. 11 ALU sites were randomly picked and PCR-validated in all 20 samples. 220 PCR reactions were performed in total. The false positive rate, estimated as the number of false positive sites across all 20 samples divided by the total number of discovered

MEIs, is 5.4%. The detection power, measured as the number of discovered true positive sites divided by the total number of true positive sites, is 87.7%.

Additionally, for 8 homozygous MEIs, we compared their LIME estimated lengths to their Sanger-sequencing measured true lengths. In 7 out of 8 the MEIs, the difference between the true and the estimated MEI length was smaller than 20 base pairs. Only one MEI was estimated as 154 bp but was actually 303 bp in Sanger sequencing results.

Analysis on Sardinia dataset

We applied LIME to 493 low-coverage sequenced samples in Sardinia Whole Genome Sequencing Project¹². All samples were pair-ended sequenced with 100 bp read length on each end, with a median coverage of 4.6 (Table 2.5). In this dataset, LIME found 6,215 ALUs and 322 L1s. In average, we detected 757 ALUs and 69 L1s from each individual. Our MEI call set showed similar allele frequency spectrum as previously reported SNP call set¹² (Figure 2.3). Both call sets have an excessive number of low-frequency variants.

We annotated these MEIs and predicted their functions using SnpEff²³. Three MEIs were annotated as “high impact”, indicating their potentially disruptive impact on the coding genes. Seventeen MEIs were annotated as “moderate impact”, showing they’re probably non-disruptive variant that might change protein effectiveness. Around one thousand variants were annotated as “low

impact”, meaning most of them are likely harmless. Others are either intergenic or located in non-coding genes. This indicates potential functionality of some MEIs, which may be further revealed in additional analysis.

Discussion

We developed a likelihood-based method to detect MEIs from whole-genome sequencing data. Based on simulation using real reads from each sample, our likelihood incorporates sample-specific features and provides a quantitative way of measuring how likely an MEI event is true. On simulated datasets and 1000G deep-sequenced samples, LIME shows better performance to other methods, especially in low-coverage data.

Our modeling of MEI likelihood may be further implemented incorporating other features such as mapping scores. With the current classification methods, reads within same category contributes equally to an MEI’s likelihood, no matter how high their mapping score to the MEI consensus sequence are. A more detailed classification of reads incorporating mapping scores may be a future direction to improve LIME, and will be especially useful in situations where longer reads have various lengths of split-mapped ends.

Under current short-gun sequencing technology, MEI calling accuracy is still much lower than short variants. Factors such as data quality, batch effect and sequencing depth may greatly affect MEI detector performance. In population-

scale data, external haplotype information could be help to improve genotyping. Using high-quality SNPs and prior information of recombination rate, low-quality genotypes could be corrected by their nearby markers. With our LIME approach, MEIs could be combined with SNPs for further refinement, thus improved MEI detection quality in another aspect.

Figures

Figure 2.1 MEI detection power in simulated data across four methods (100 bp pair-ended reads)

The red line shows the power of LIME, while the other three colored lines show power of three most widely used methods. Power is calculated as the number of detected MEIs divided by the number of simulated MEIs in the modified reference genome sequences.

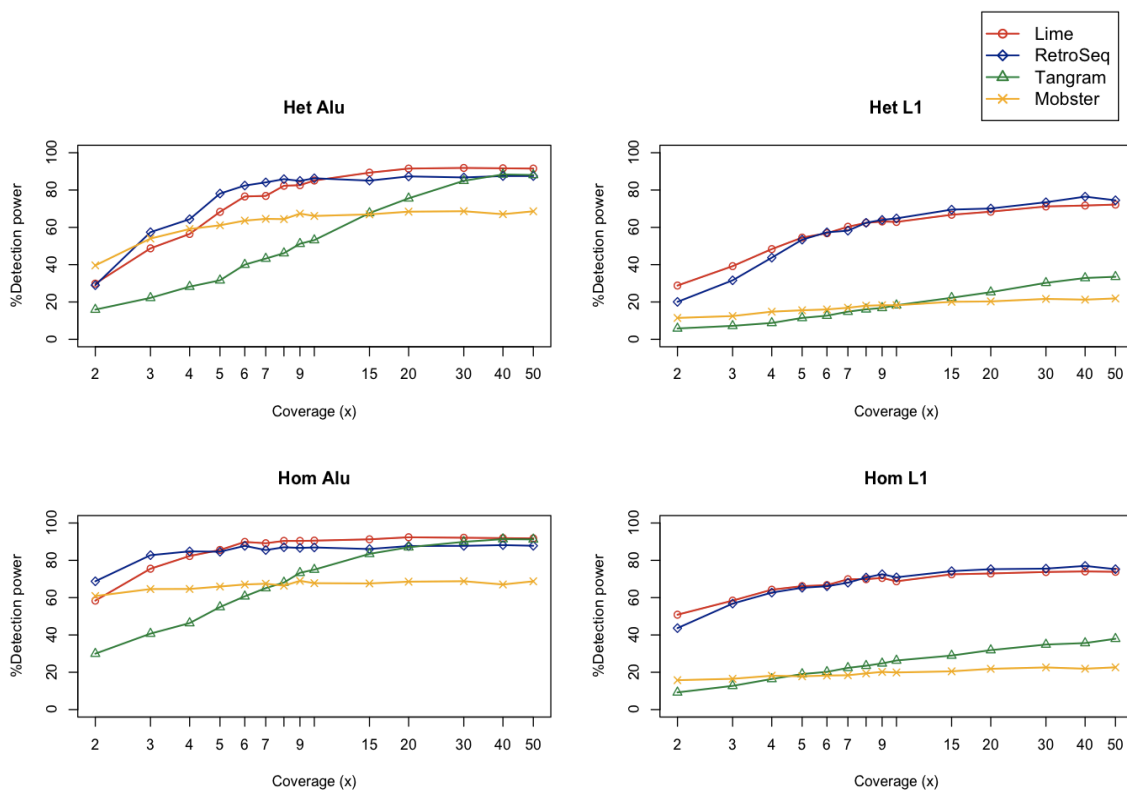


Figure 2.2 Breakpoint detection accuracy of four methods in simulated data

The breakpoint detection accuracy was estimated as the base pair distance between true breakpoint and the estimated breakpoint simulated data with homozygous ALUs and 100 bp pair-ended reads under 20x coverage. LIME shows good breakpoint detection accuracy as well as Tangram and Mobster, with most of distances between the true and the estimated breakpoints smaller than 20 base pair.

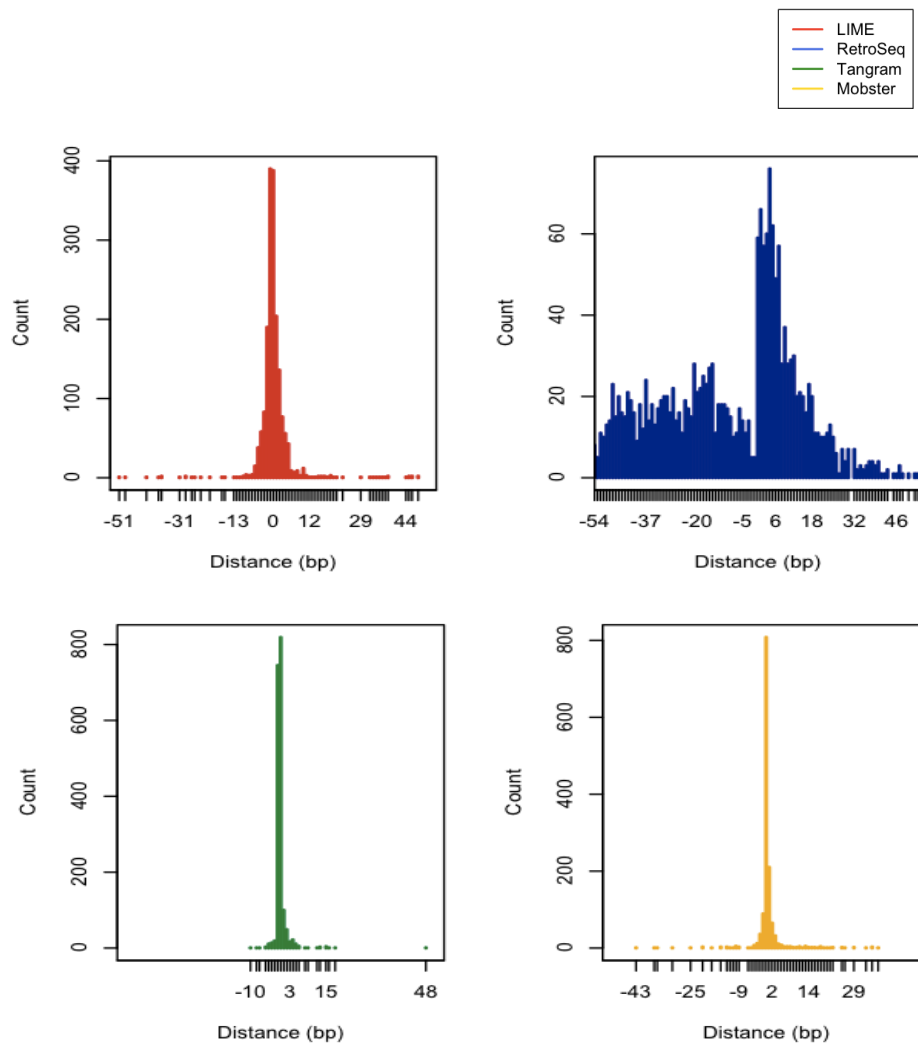
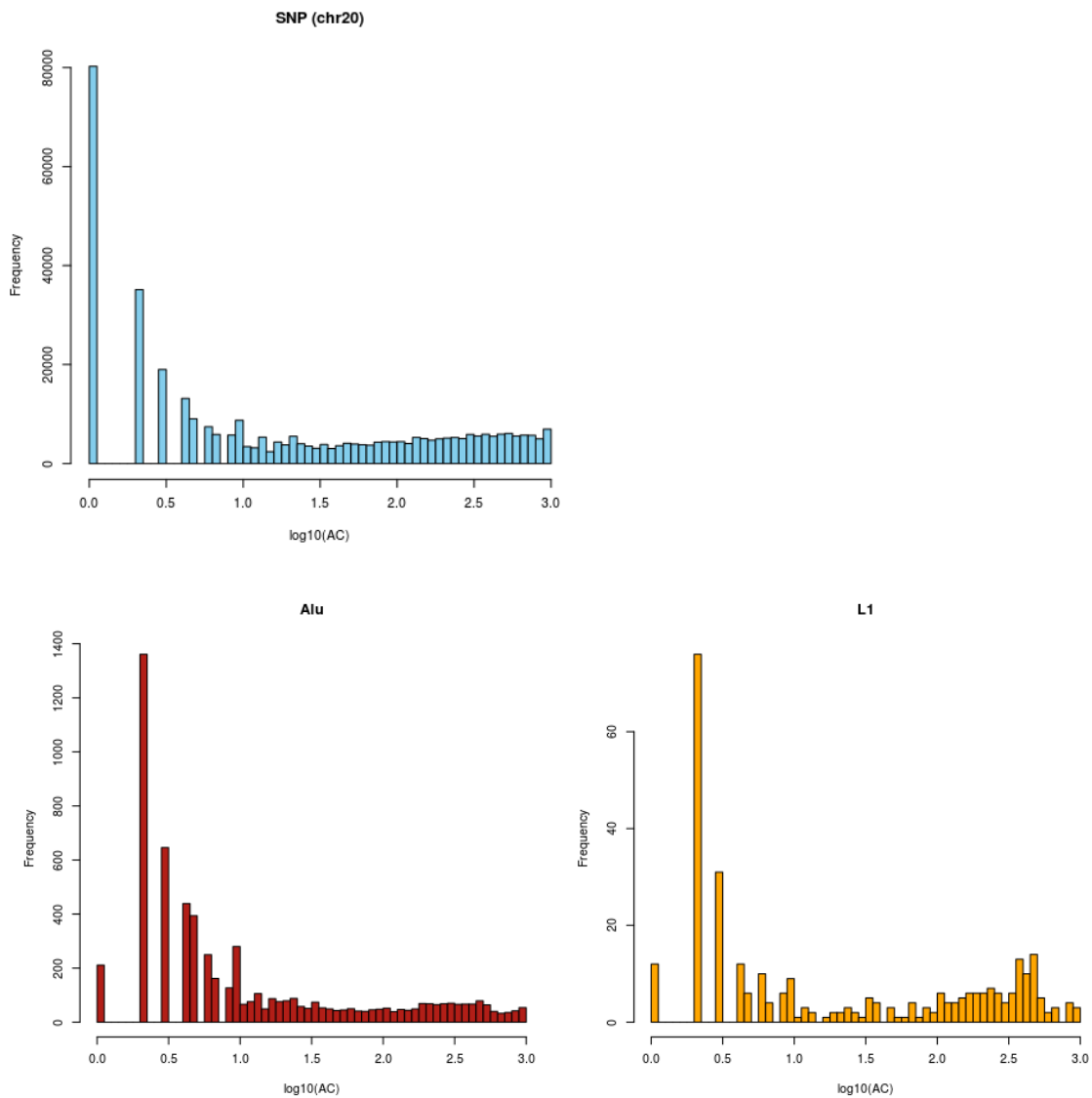


Figure 2.3 Log-scale allele frequency spectrums of SNPs and MEIs in samples from Sardinia Whole-genome Sequencing Project (493 samples)

Overall, frequency spectrums of MEIs and SNPs (provided by Sidore et al¹²) are very similar except for limited sensitivity of MEI singletons due to low coverage.



Tables

Table 2.1 The 18 pre-defined read pair types in LIME

Read pair types were defined based on the following criteria:

1. We set the distance threshold as the sample's insert size plus 3 times of the standard deviation of the insert sizes of the sample. Of the mapped distance between two ends of the read pair is smaller than the distance threshold, and have the correct orientation, it is named "properly mapped"; otherwise if the distance is smaller than the distance threshold, it is named "discordantly mapped".
2. For properly mapped read pairs, if one end is split-mapped and the clip is longer than the defined minimum clip length: if the clip is at the 5' end of the read, it is named as "left clip", otherwise it is "right clip". If the clip is mapped to an MEI consensus sequence, it is marked as "Mapped to MEI", otherwise marked as "Not mapped to MEI".
3. Similarly, for read pairs with one end not mapped, the "Mapped to MEI" condition is decided by the mapping state to the MEI consensus sequence for the unmapped end.
4. For discordantly mapped read pairs, the end mapped to the examined genomic region is named "Anchor". Similarly, the "Mapped to MEI" condition is decided by the other end's mapping state to the MEI consensus sequence.

Properly mapped			
Properly mapped with clip length < minimum clip length			
Properly mapped with clip >= minimum clip length	Read mapped to plus strand	Left clip	Mapped to MEI
			Not mapped to MEI
		Right clip	Mapped to MEI
			Not mapped to MEI
	Read mapped to minus strand	Left clip	Mapped to MEI
			Not mapped to MEI
		Right clip	Mapped to MEI
			Not mapped to MEI
Discordantly mapped read pairs	Anchor mapped to plus strand		Mapped to MEI
			Not mapped to MEI
	Anchor mapped to minus strand		Mapped to MEI
			Not mapped to MEI
Read pairs with one end not mapped	Anchor mapped to plus strand		Mapped to MEI
			Not mapped to MEI
	Anchor mapped to minus strand		Mapped to MEI
			Not mapped to MEI

Table 2.2 Average proportions of read pair types in windows of homozygous MEIs in NA12878 of 1000 Genome Project Phase 3

For each identified MEI from deep-sequenced individual NA12878, we classified and counted read pair types within a 600 bp window (centered at each MEI breakpoint). An excess of clipped and discordantly mapped reads was observed in these windows. Note that “mapped to MEI” reads were majorly used to detect MEIs from most software.

Read pair type		Average proportions of pairs in window
Properly mapped		0.430
Properly mapped with clip length < minimum clip length		0.018
Properly mapped with clip \geq minimum clip length	Mapped to MEI	0.043
	Not mapped to MEI	0.000065
Discordantly mapped read pairs and Read pairs with one end not mapped	Mapped to MEI	0.336
	Not mapped to MEI	0.173

Table 2.3 Number and percentage of covered PCR-validated MEIs in 1000 Genome Phase 3 deep sequenced trio samples (NA12878, NA12891, NA12892)

LIME and the other three methods were applied to the three 80x samples. LIME and Mobster has better coverage of PCR-validated MEIs, which implies higher detection powers of these two methods.

	# Total Covered (%Covered)	# ALU Covered (%Covered)	# L1 Covered (%Covered)
Lime	460 (98.1%)	441 (99.1%)	20 (83.3%)
RetroSeq	451 (96.2%)	435 (97.8%)	18 (75.0%)
Tangram	455 (97.0%)	436 (98.0%)	20 (83.3%)
Mobster	464 (98.5%)	441 (99.1%)	21 (87.5%)

Table 2.4 Parent-child inconsistency of MEIs in 1000 Genome Phase 3 deep sequenced trio samples (NA12878, NA12891, NA12892)

The count of parent-child inconsistent MEIs is calculated as the number of MEIs in the offspring (NA12878) but not existing in any of the parents (NA12891 and NA12892). LIME has the lowest overall inconsistent rate, and the lowest inconsistent rate in L1.

Caller	# Inconsistent ALUs (%Inconsistency)	# Inconsistent L1s (%Inconsistency)	# Inconsistent MEIs Total (%Inconsistency)
Lime	95 (6.9%)	13 (9.5%)	107 (7.0%)
RetroSeq	51 (4.7%)	41 (23.5%)	92 (7.3%)
Tangram	155 (11.6%)	26 (11.4%)	181 (11.6%)
Mobster	94 (7.8%)	66 (23.3%)	151 (9.6%)

Table 2.5 Summary of sequencing characteristics in Sardinia Whole

Genome Sequencing samples (n = 493)

Overall the Sardinia samples were sequenced in low-coverage, with an average coverage of 4.6x. The samples have consistent insert size but vary in mapping read and the proportion of properly mapped reads, which implies batch effect and increased error rate in part of the samples.

	Mean	Median	Minimum	Maximum
Depth (x)	4.9	4.6	4.0	9.8
Mapping rate	97.4%	97.6%	94.3%	99.5%
Properly mapped read	94.3%	94.1%	86.6%	98.6%
Insert size (bp)	294	290	251	349

Supplementary Text

Optimizations on LIME internal simulation

When constructing known genotypes to mimic read type distribution, it is computationally expensive to remap all reads back to the modified reference genome. For efficiency, we only remap pair-end reads around each spliced region. This is substantially faster and results in similar accuracy as remapping all reads.

Secondly, considering there are less than 2,000 MEIs per individual genome, it is not cost-effective to analyze all windows across the genome. In our simulation results (discussed below), we found that, even when the coverage is as low as 2, all MEIs had at least one read pair with abnormal insert size and one end discordantly mapped to a mobile element sequence elsewhere in the genome. Therefore, instead of analyzing all windows along the genome, LIME only analyzes regions with at least one such read pair.

Thirdly, to measure impact of the number of constructed windows on genotyping accuracy, we used a simulation approach to generate 2000 ALUs on chromosome 17 under coverage of 2 and 5. (The simulation method will be discussed in section 3.1.) We tested LIME detection power with different number of constructed homozygous MEI windows and calculated the percentage of identified MEIs using 10,100,300 and 800 constructed homozygous MEI

windows. When using only 10 reference windows, 47% and 50% detection power was lost in 2x and 5x data, respectively. Using the same dataset, we then tested MEI detection power using 10,100,1000,5000, and all constructed reference genotype windows on chromosome 20. When using 10 reference windows, 37% and 3% detection power was lost in 2x and 5x data, respectively. These results indicate that an insufficient number of constructed windows will lower detection power. Therefore, to maintain detection power and minimize computational cost at the same time, we set 300 and 5000 as the default number of constructed windows with homozygous MEI and reference genotype in LIME, respectively.

Commands for evaluation of LIME, RetroSeq, Tangram and Mobster

Command for LIME:

```
LIME -SampleList <chr20-sample.list> -OutVcf <refstat> -SiteVcf <empty file> --  
statOnly
```

```
LIME -SampleList <sample.list> -OutVcf <output> -rSingle <refstat> -Chr 17
```

Command for RetroSeq:

```
retroseq.pl -discover -bam <bam> -output candidates.tab -q 10 -len 30 -refTEs  
ref_types.tab -eref probes.tab -align  
retroseq.pl -call -bam <bam> -input candidates.tab -ref hs37d5.fa -output  
<output> -reads 1 -q 10 -region 17 -hets
```

Tangram only works on MOSAIK mapped results, so *tangram_bam* needs to be used first to convert bwa mapped bams to MOSAIK mapped ones:

```
tangram_scan -in bam-list.txt -dir scan_out -mf 0
```

Tangram is applied using following command:

```
tangram_detect -lb scan_out/lib_table.dat -ht scan_out/hist.dat -in bam-list.txt -rg 17 -ref hs37d5.indexed.ref -out <out> -gt -bp -mcs 1 -mq 10 -smq 10 -rpf 1 -srf 1
```

Mobster throws an exception when processing reads that are marked as PCR duplicates or supplementary mapping. Therefore, we remove those reads from simulated data and apply Mobster using the following command:

```
java -jar Mobster.jar -properties <properties> -in <bam> -out <out> -sn <sample>
```

Here, in mobster property, we set minimum required supporting reads to one instead of default five to increase detection power in low-coverage data.

PCR protocols

DNA concentration was around 50 ng/ul. PCR primers were designed used NCBI Primer-Blast for the regions of interest spanning various chromosomal regions.

We used NEB 2x master mixes of OneTaq high fidelity hot start enzyme either with standard buffer or GC buffer based on the sequence of interest. Primers stocks were made to 100 uM and working dilutions were used at 10 uM. Strip tubes were centrifuges briefly and place in PCR machines PE 9700 with the following run program:

1 cycle at 94 °C, 2 minutes

10 cycles at: 94 °C, 20 seconds; 58 °C, 30 seconds; 68 °C, 45 seconds

Followed 25 cycles at: 94 °C, 20 seconds; 60 °C, 30 seconds; 68 °C, 45 seconds

1 cycle at 68 °C, 10 minutes

Hold at 4 °C

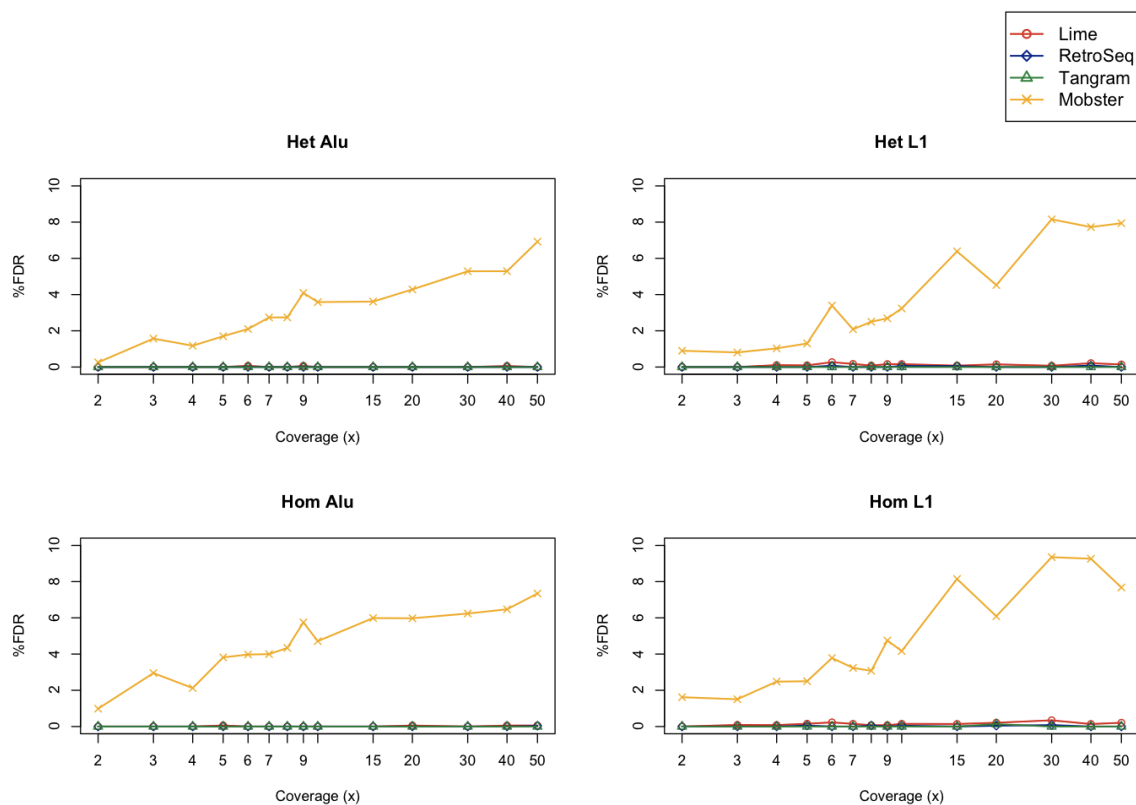
After PCR amplification, 1 ul of each was run on the Agilent Bioanalyzer 2200 TapeStation machine with the relevant ladders. Samples were submitted to the University of Michigan Sequencing Core for Sanger Sequencing using the forward and the reverse primers. Run files were analyzed using the GCC demo software Sequencer version 5.1.

Supplementary Figures and Tables

Supplementary Figure 2.1 MEI false discovery rate (FDR) in simulated data (100 bp pair-ended reads)

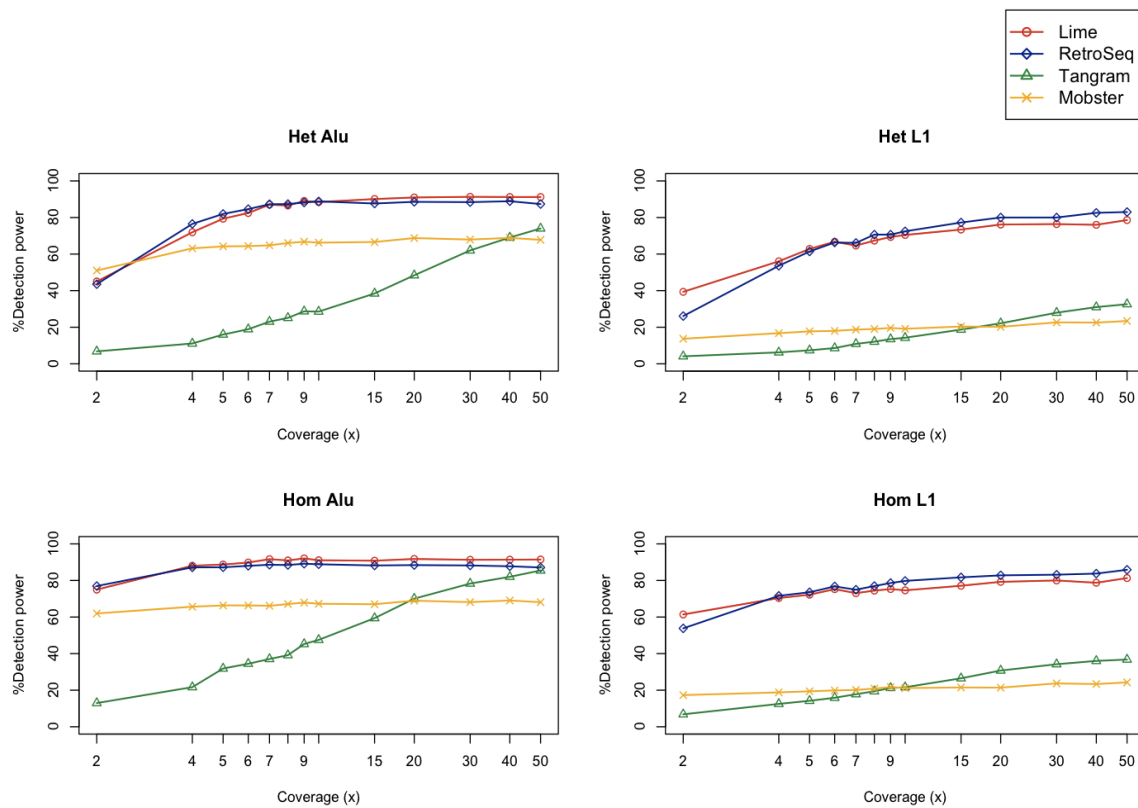
We estimated the false discovery rate as the number of detected MEIs not overlapped with simulated MEIs divided by the number of detected total MEIs.

LIME has very low FDR across all coverage.



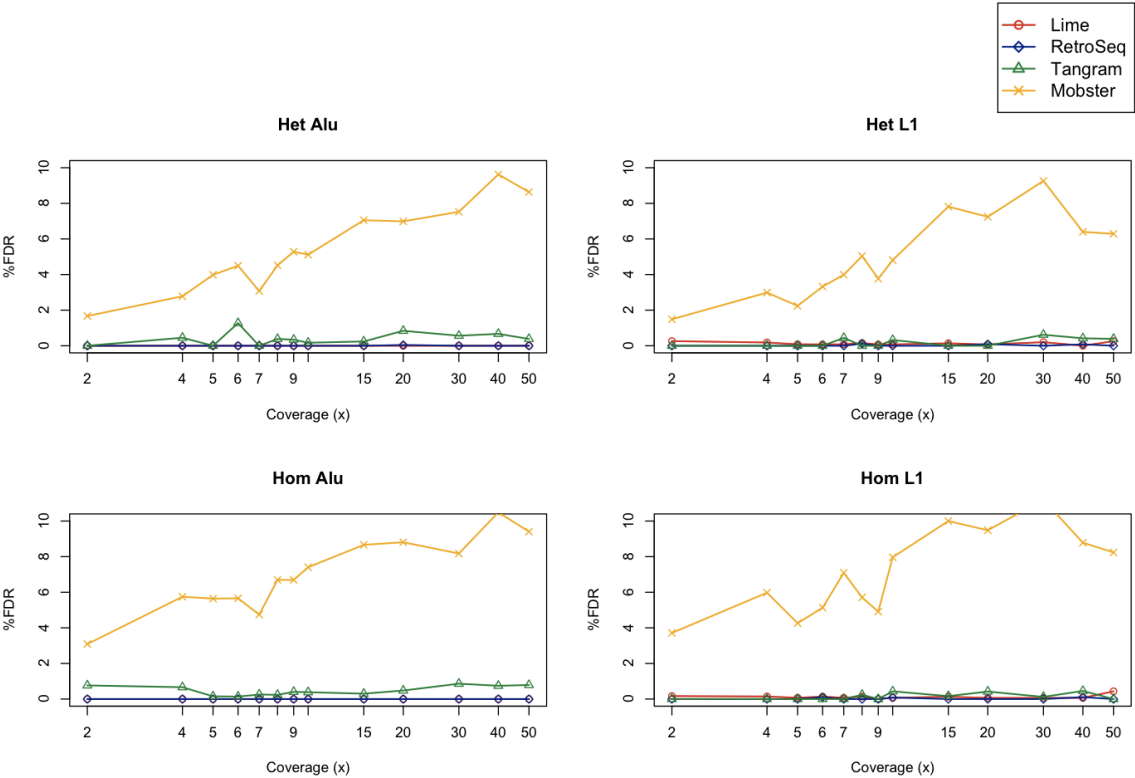
Supplementary Figure 2.2 MEI detection power in simulated dataset (75 bp pair-ended reads)

For all 4 methods, deeper coverage results in high MEI detection power. In lower coverage situations, LIME and RetroSeq have better detection power. Detection power of the 4 methods saturated at a certain coverage. These trends are similar to the 100 bp situations shown in Figure 2-1.



Supplementary Figure 2.3 MEI false discovery rate (FDR) in simulated dataset (75 bp pair-ended reads)

LIME has low FDR across all coverage, similar to Supplementary Figure 2-1.



Supplementary Table 2.1 Numbers of detected MEIs in 1000 Genome Phase

3 deep sequenced samples (ALUs, L1s and SVAs)

Overall the four methods have similar number of detected MEIs in each sample.

RetroSeq is most conservative while the other three methods have more similar number of detected MEIs.

Caller	NA12878	NA12891	NA12892
Lime	1549	1665	1588
RetroSeq	1222	1139	1153
Tangram	1511	1359	1369
Mobster	1568	1424	1392

References

1. Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
2. Solyom, S. & Kazazian, H. H. Mobile elements in the human genome: implications for disease. *Genome Med* **4**, 12 (2012).
3. Biéumont, C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **186**, 1085–1093 (2010).
4. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
5. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, e1002236 (2011).
6. Woods-Samuels, P. *et al.* Characterization of a nondeleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements. *Genomics* **4**, 290–296 (1989).
7. Green, E. P. *et al.* Sequence and characteristics of IS900, an insertion element identified in a human Crohn's disease isolate of *Mycobacterium paratuberculosis*. *Nucleic Acids Res.* **17**, 9063–9073 (1989).
8. Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T. & Nakamura, Y. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat. Genet.* **13**, 245–247 (1996).
9. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
10. Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).
11. Sims, D., Sudbery, I., Illott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).

12. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).
13. Kuk, A. Y. C., Zhang, H. & Yang, Y. Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium. *Bioinformatics* **25**, 379–386 (2009).
14. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
15. Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**, 389–390 (2013).
16. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* **15**, 488 (2014).
17. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
19. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. *An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data.* *Genome research* **25**, 918–925 (2015).
20. Wu, J. *et al.* Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* **15**, 795 (2014).
21. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
22. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).

23. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

CHAPTER III

Discovery of novel variants associated with colorectal cancer

Introduction

Colorectal cancer is the fourth most common type of cancer. With an estimate of 135,430 new cases in 2017, consisting of 8% all new cancer cases in the United States¹. In recent years, under the framework of genome-wide association studies (GWAS) researchers have discovered many common variants and genes statistically associated with colorectal cancer²⁻¹¹. For example, the mutation of rs16892766 at 8q23.3 leads to overexpression of cell growth factor EIF3H, which greatly increases the risk of colorectal cancer^{7,12,13}. The mutation of rs4939827 at gene SMAD7 has functional impact on the *Wnt* signaling pathway and is highly associated with colorectal cancer risk⁴. These findings provided new insights into

Part of this work in Chapter 3 will be submitted as:
Chen S, Huyghe J, Qu F, Lin Y, Peters U, Abecasis G. "Discovery of novel risk variants for colorectal cancer from an imputation-based study".
The WGS work in Chapter 3 has been included in the publication as:
McCarthy S, Das S, ..., Chen S, ..., Durbin R, Abecasis G. "A reference panel of 64,976 haplotypes for genotype imputation", *Nat Genet*, 2016
Rashkin S, Jun G, Chen S, Abecasis G, "Optimal sequencing strategies for identifying disease-associated sigletons", *Plos Genetics*, 2016
The HRC imputed dataset in Chapter 3 has been included and submitted as:
Bien S, Auer P, ..., Chen S, ..., Hsu L. "Enrichment of colorectal cancer associations in functional regions: insight for using epigenmocs data in the analysis of the whole genome sequence-imputed GWAS data". *Plos Genetics*, 2016

specific genes and regulatory regions, and revealed the importance of previously ignored pathways.

Even with these achievements, there still remain some unsolved questions. The heritability of colorectal cancer varies from 7.4% to 35%¹⁴⁻¹⁶, indicating the contribution of unknown variants. Recent progress in genotyping technology and statistical methods has made larger-scale imputation-based studies possible, but few studies have applied these kinds of methods to colorectal cancer.

Here, we conducted a genome-wide association analysis in 22 million Haplotype Consortium Reference Panel (HRC)¹⁷ imputed SNPs from 26,903 individuals across 14 studies. Compared to the 1000 Genome Phase 3 reference Panel, the newly released HRC panel has an unprecedented density and a higher rare variant imputation quality¹⁷. We identified six candidate regions and one gene potentially associated with colorectal cancer. By integrating functional annotation, sequence function prediction, tissue-specific expression quantitative trait loci (eQTL) data and meta-analysis, we highlighted several SNPs in known risk regions as potentially functional candidates.

Additionally, as part of the collaboration effort within the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), we conducted a variant calling on 3,061 whole-genome sequenced individuals, aiming to generate a consensus call set of variants that are potentially more correlated with

colorectal cancer. We generated a dataset of 48.7 million variants with superior quality, and half of them are novel to dbSNP¹⁸ build 145.

Materials and Methods

Data processing and imputation

Our dataset consists of 12,186 controls and 14,717 cases, coming from 6 genotyping platforms and 14 studies. Detailed information about the studies and genotyping methods can be found in Table S3.1. All individuals were identified as Europeans (Figure S3.1)^{19,20}. Samples from the same genotyping platform was merged and imputed onto Haplotype Reference Consortium (HRC) reference panel on Michigan Imputation Server²¹.

Statistical analysis

For phenotypes, individuals with or without colorectal cancer were coded as 0 or 1, respectively. Adjusted covariates included age, sex, study, batch information (if applicable) and first three genotype PCs estimated by PLINK²². Covariates were first regressed out from the phenotypes. We used the expected number of alternative alleles (dosage) for unbiased estimates²³. Single-variant analysis was performed under a linear mixed model framework of binary score test extensions²⁴ using EPACTS²⁵ (Figure S3.3). Meta-analysis was performed by summarizing p values and Z-scores in METAL²⁶ (Figure S3.2, S3.4).

For the gene-based test, we conducted a burden test with equal weight²⁷ using Raremetal^{28,29} (Figure S3.5). We used LocusZoom³⁰ to demonstrate potentially significant loci.

For single-variant meta-analysis, we used p value $< 5 \times 10^{-8}$ as the genome-wide significance threshold. As mentioned by previous studies, at p value $< 5 \times 10^{-7}$ there still exist possible associated variants in European populations³¹⁻³⁶.

Therefore, we also reported variants with p value smaller than 5×10^{-7} as potential candidates. For the gene-based test, as we tested 18,677 genes, 2.7×10^{-6} was used as the genome-wide significance threshold at the significance level of $P < 0.05$ after multiple corrections.

Functional annotation

Using EPACTS, we annotated all variants (both genotyped and imputed) into six categories: synonymous, nonsynonymous, exon, intron, stop gain and stop loss. To search for causal SNPs, we built and optimized an fGWAS³⁷ model incorporating genomic annotations and meta-analysis summary statistics. The final model, which has likelihood maximized, includes three annotations: exon, synonymous and stop gain. Previous fGWAS studies have suggested using posterior probability of association (PPA) greater than 0.9 as the threshold for causal SNPs. However, in other previous studies, $PPA > 0.2$ was also widely used^{38,39}. Thus, we highlighted SNPs and regions with $PPA > 0.2$ as potentially causal candidates.

Sequence functionality prediction

We used a Support Vector Machine (SVM) classifier, deltaSVM⁴⁰, to predict functions of each SNP based on its reference and alternative alleles, together with its 11 base-pair flanking sequences on both sides. We used ENCODE tissue-specific DNaseI peak information on the UCSC Genome Browser⁴¹⁻⁴³ to generate training sets for the classifier. Peaks were centered and normalized to a uniform length of 500 base pairs. Sorted by p values in ascending order, we selected the first 100,000 peaks as the positive training set. To construct negative training sets, we binned the rest of the genome into 500 base-pair windows, and randomly sampled sequences by their matching GC content and fraction of repeats against those in the positive training set.

Test eQTLs

First we selected variants with meta-analysis p value less than 10^{-5} and absolute value of deltaSVM score higher than 1. The deltaSVM score was predicted using the model from HCT-116 (a colorectal cancer cell line)⁴⁴. Second, for these variants, we obtained their rsID, the nearest gene Ensemble ID and histone modification information from the UCSC Genome Browser and HaploReg version 4.1⁴⁵. On GTEx eQTL website⁴⁶, we used their rsID, Ensemble ID to test eQTLs in tissue “Colon Sigmoid.” Box plots and p values were automatically generated by the webpage.

Whole genome sequencing and variant calling

Samples were collected through multiple studies from the Genetics and Epidemiology of Colorectal Cancer Consortium. 1187 samples were genotyped with different Illumina genotyping arrays as well. At the University of Washington Sequencing Center, sequencing reads were generated, were aligned to the human reference genome (GRCh37 assembly⁴⁷) using BWA⁴⁸ (v0.6.2 for Year 1 and 2, v0.7.10 for Year 3). After the alignment, samples were delivered to the University of Michigan for variant calling.

On Year 3, to minimize batch effect coming from different versions of aligners, we performed a re-alignment for all samples. All BAM files were first converted back to FASTQ using BamToFastQ⁴⁹, then re-mapped to the human reference genome (GRCh37 assembly with decoy sequences as available from the 1000 Genomes Project) using BWA-MEM⁵⁰ v0.7.12. After this re-alignment, base qualities were recalibrated and duplicated reads were flagged with BamUtil⁵¹. We reviewed the summary metrics generated using QPLOT⁵² for each sample. We used LASER⁵³ for sample ancestry inference, and verifyBamID⁵⁴ for estimating contamination level. Population outliers and contaminated samples were identified but still included in variant calling.

Every year, variant call was performed together with all the samples that were delivered before. In total, three rounds of variant calls were performed during 2013 to 2016. On Year 1 and Year 2, we used GotCloud⁵⁵ pipeline to detect

SNPs and Indels. In brief, GotCloud automatically separate the analysis into many small jobs that are distributed on a high-performance computing (HPC) cluster. This pipeline includes annotating variants with various calling features, and both hard filters and a support-vector machine (SVM) classifier filter based on these features. Additionally, on Year 2, we used GenomSTRiP⁵⁶ to detect large deletions. We made a customized script to distribute GenomeSTRiP jobs on our HPC cluster. In Year 3, we used VT⁵⁷ to detect SNPs and Indels, and used an customized GotCloud-like pipeline for job distribution and variant filter. Finally, using the list of high-quality sites, we performed LD refinement for detected genotypes using the haplotype-aware calling algorithm in BEAGLE⁵⁸. The final list of variants was annotated with SNPeff⁵⁹.

Results

Genome-wide association analysis

In our large-scale meta-analysis, we successfully replicated many previous findings, especially in European populations (Table S3.2). After excluding regions with known risk SNPs, two signals remained at the level of $p < 5 \times 10^{-8}$, and four clusters of signals remained at the level of $p < 5 \times 10^{-7}$ (Table 3.1).

At the level of $p < 5 \times 10^{-8}$, we detected a novel signal at 7p12.1 (rs115561508, MAF=0.1, $P=3.6 \times 10^{-9}$) in an intergenic region, 120 Kb downstream of the stop codon of gene POM121L12. This gene has been reported as mutated in a rare cancer, Muscinous neoplasms of the appendix (MNA)⁶⁰. We also detected a

novel signal at 9q21.12 (rs138140376, MAF=0.003, $P=1.9 \times 10^{-9}$), which is a multi-allelic SNP; the other alternative allele was not significant in meta-analysis (MAF=0.0003, $P=0.55$). Rs138140376 is intronic to gene TRPM3, and 178 Kb downstream of gene KLF9. The relationship between these two genes and colorectal cancer has been reported in some experimental studies⁶¹⁻⁶⁴, but until now no GWAS studies have identified this relationship. Both rs115561508 and rs138140376 do not have nearby genome-wide significant SNPs, with the secondary lowest p value in their regions at the 10^{-6} level (Figure S3.6A,B).

At the level of $p < 5 \times 10^{-7}$, the two regions, 2p24.2 (rs78115417, MAF=0.007, $P=4.1 \times 10^{-7}$) and 5q35.1 (rs555044933, MAF=0.004, $P=1.7 \times 10^{-7}$) both had a rare significant SNP at the intergenic region. There is little LD information in these two regions (Figure S3.6C,D). However, in the other two regions, 1p21.3 and 6p21.31, we observed strong clusters of signals (Figure 3.2). In 1p21.3 (rs841684, MAF=0.47, $P=1.5 \times 10^{-7}$), there were 7 SNPs (rs841684, rs841689, rs1098724, rs1098725, rs2798940, rs866365, rs841708, rs1772895) from a strong LD block with p values smaller than 5×10^{-7} (Figure 3.1A). In 6q21.31 (rs12529688, MAF=0.085, $P=2.5 \times 10^{-7}$), the lead SNP rs12529688 was in complete linkage with other three significant SNPs (rs16877540, rs76489311, rs146718198) (Figure 3.1B). Rs12529688 is intronic to the promoter region of FKBP5, which has altered expression level in many different tumors⁶⁵. Additionally, one experimental study has shown that FKBP5 suppresses the proliferation of colorectal adenocarcinoma⁶⁶.

In the gene-based test, we tested 18,677 genes using coding variants only. With a Bonferroni corrected genome-wide significant threshold at $p = 2.7 \times 10^{-6}$, we identified one gene with p value above the threshold and two genes with p values near the threshold (Figure S3.5). Among them, POU5F1B ($P = 2.3 \times 10^{-12}$) and SH2B3 ($P = 4.5 \times 10^{-6}$) have been reported as associated with colorectal cancer in many previous studies^{2,3,5,7,9,13}. HAO1 ($P = 5.0 \times 10^{-6}$) contains 10 coding SNPs, and none of these SNPs reached a genome-wide significant threshold in single-variant meta-analysis (Table S3.3, Figure S3.7). A Korean GWAS study showed the correlation between HAO1 and childhood leukemia⁶⁷, but no previous studies have shown this gene's correlation with colorectal cancer. An East Asian GWAS study⁶⁸ identified a known risk SNP, rs2423729, which is 51 Kb away from the HAO1 transcription start site, as associated with colorectal cancer. However, in our meta-analysis, it was not significant ($P = 0.01$).

Functional annotation

We constructed an optimized fGWAS model that integrated three genomic annotations: exon, synonymous and stop gain (Table S3.4A). We estimated PPA for each genome block and each SNP to measure how likely a SNP is to be causal and how likely a genome block is to contain a causal SNP for colorectal cancer, respectively. Thirteen regions and 10 SNPs within them reached the threshold of causality of $PPA > 0.2$ (Table S3.5).

Regions at 1p22.1-1p21.3 (PPA=0.36), 6p21.31 (PPA=0.36), 7p12.1 (PPA=0.74) and 9q21.12(PPA=0.84) contained our newly identified significant signals. In our meta-analysis, the region at 4q22.2 (PPA=0.20) did not have genome-wide significant signals, with is consistent with the lack of reports about colorectal cancer-associated variants in this region. The other 8 regions all contained at least one previously identified risk SNP.

Among the 10 SNPs with PPA>0.2, rs115561508 (PPA=0.73) and rs138140376 (PPA=0.83) are our newly identified genome-wide significant signals. Others are previously identified SNPs or within the same LD block of the known risk SNPs. Additionally, rs841684 at region 1p22.1-1p21.3, with PPA of 0.16, is near the significant threshold of causal SNPs. It is worth notice that in region 15q13.3, we prioritized two SNPs in GREM1 gene body with significant PPA: rs1406389 (PPA=0.28, meta-analysis $p=2.0\times10^{-10}$) and rs2293581 (PPA=0.23, meta-analysis $p=2.5\times10^{-10}$). In previous studies, both SNPs were previously identified as causal variants in this region^{69,70}, and therefore a nice positive control for our approach.

Sequence function prediction and expression analysis

As we constructed fGWAS models with three tissues, GWAS signals were more enriched in DNaseI peaks from HCT-116 (a colorectal cancer cell line) from ENCODE database than from the control cell line GM12878 (Table S3.4B). We applied the deltaSVM model constructed from HCT-116 DNaseI peaks to our

imputed dataset, and generated DNaseI sensitivity prediction scores for each variant. Most scores were around zero, with a few variants having extreme scores, indicating their potential functionality (Figure S3.8).

To prioritize potential functional candidates, we combined the information from genome-wide meta-analysis and DNaseI sensitivity prediction. We extracted 22 SNPs with both low meta-analysis p values and predicted higher DNaseI sensitivity. In Colon Sigmoid tissue, half of the 22 SNPs have the allele-specific expression data of nearby genes in the GTEx database. After testing allelic-specific expression of these 11 SNPs, we obtained 4 signals (rs72894784, rs34645899, rs62072496, rs1741635) with expression p value smaller than 0.05 (Table S3.6, Figure S3.9 A-D).

Among the 4 signals, the first one, rs72894784, is in the same LD block as the genome-wide significant signal rs146718198 (meta-analysis $p=2.7 \times 10^{-7}$), and in this block we found 4 SNPs with p value at 10^{-7} level. The second signal, rs34645899, (meta-analysis $p=7.0 \times 10^{-5}$, eQTL $p=2.6 \times 10^{-6}$) was still significant even after multiple-testing correction. The SNP is intronic to ATF1, a gene reported as associated with carcinoma and melanoma⁹. A known risk SNP, rs11169552, is 46Kb upstream of rs34645899, but does not show strong correlation with ATF1 allelic-specific expression (eQTL $p=0.076$). This may be an interesting finding because the region has extensive LD and it is still unclear which gene is the effector gene. The third signal, rs62072496, is 4 Kb upstream

of LLGL1, a gene crucial in *Drosophila* neurogenesis⁷¹. There are no significant meta-analysis signals or known risk SNP in this region. The fourth signal, rs1741635, is 17KB downstream of known risk SNP rs2427038 and rs4925386. All these three SNPs are in the coding region of the colorectal cancer associated gene, LAMA5⁷². However, the two known SNP did not show a significant correlation with the LAMA5 expression level.

Variant calling from the whole-genome sequenced samples

We generated whole-genome shotgun sequencing data for 3,061 unrelated individuals. 10 of them are African Americans (all from Year 1) and the rest are Europeans. The average fold coverage of the genome is 6.6, 4.7, 35.2 from Year 1 to Year 3 (Figure 3.2A). Insert sizes are relatively consistent across samples (Figure 3.2B,C). We performed quality control, re-alignment, variant calling and LD refinement that efficiently handled this large dataset (see Methods for details).

In Year 3 call set, we identified 46.1 million SNPs and 2.6 million Indels from the 3,061 individuals (Table 3.2A). The average Transitions/Transversions (Ts/Tv) ratio of the detected SNPs is 2.37. In each individual, we identified an average of 3.4 million SNPs and 0.22 million Indels (Table 3.2B). Additionally, we observed an average of 10.3 thousand singletons and 47.7 thousand rare variants of frequency <0.5% in Year 3 deep samples, which is much higher than the numbers in Year 1 and Year 2 low-coverage samples, with an average of 4.6 thousand singletons and 39.3 thousand rare variants per sample (Figure S3.10).

For quality assessment, we compared genotypes from sequencing and microchip across 7,767 sites in 1,857 individuals. The average error rate is 0.30%.

In the 48.7 million variants discovered, 47.1% SNPs and 49.4% Indels are in protein coding regions. 0.25% Indels are predicted to cause coding frameshift. Compared to other sequencing studies, 58.2% of our identified SNPs exist in dbSNP build 145. At the same time, 46.4% of our identified SNPs and 40.6% Indels overlap with 1000 Genome Phase 3 findings⁷³.

Discussion

In this large-scale meta-analysis on HRC imputed data, we identified six regions and one gene that were significantly associated with colorectal cancer. In region 7p12.1 and 9q21.12, there were two SNPs with p values lower than the genome-wide significant threshold of 5×10^{-8} . In 2p24.2 and 5q35.1, there were two rare SNPS with p values lower than the 5×10^{-7} . In these four regions, there is little LD information and no other significant signals. In 1p21.3 and 6q21.31, there were two clusters of signals with p values lower than 5×10^{-7} , and these signals included directly genotyped SNPs. These findings make these two regions better candidates than the other four regions.

Without replication analysis, it is hard to tell if these signals are true. In addition, unless the sample size is extremely large, it is very difficult for rare variants to reach the genome-wide significant threshold. Thus, with the current available

data, we performed extra analysis to prioritize potentially functional variants, aiming to obtain more information from the current dataset. We observed substantial enrichment of significant p values in exon regions. We predicted several potentially causal SNPs in both novel regions and known regions. By combining meta-analysis, sequence function prediction and eQTL data, we highlighted four potentially functional variants as affecting nearby gene expression levels. It is worth notice that in known risk regions, some of these highlighted functional SNPs are not previously known risk SNPs, but are other SNPs that are in the same LD block.

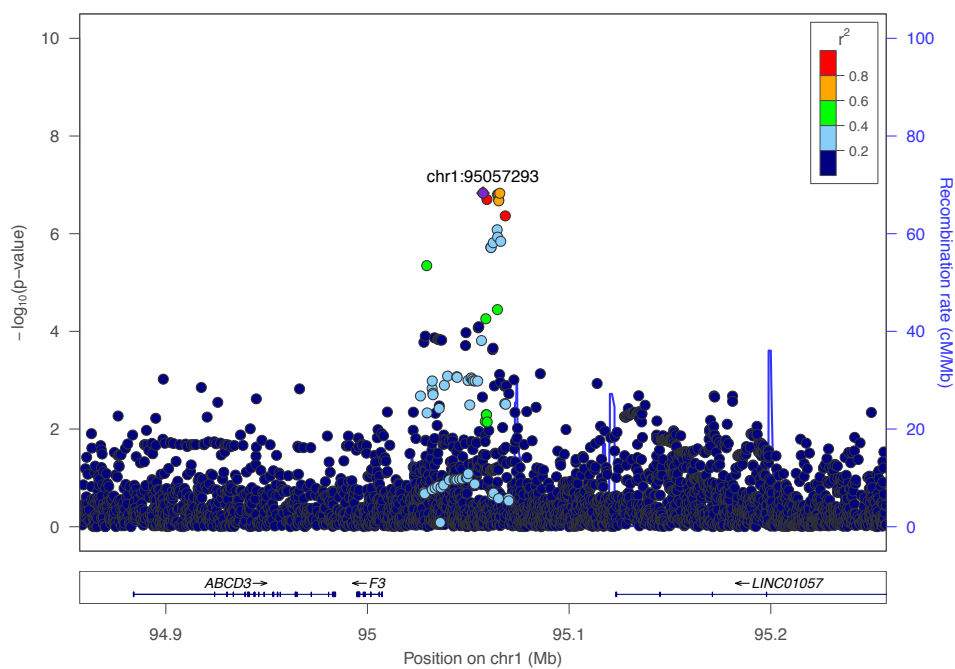
From 3,061 whole-genome sequenced individuals, we generated a dataset of 48.7 million variants. From this mixture of high and low-coverage data, we obtained an excessive number of rare variants from higher coverage samples. About half of the variants are novel to dbSNP, which means the dataset is very likely to contain undiscovered risk loci for colorectal cancer.

In conclusion, our association analysis results indicate that a larger variant set with higher marker density not only provides more potential for identifying novel variants, but also gives us a new understanding of variant functions in previously identified risk regions. To reliably discover novel risk variants, especially rare variants, an even larger sample size is necessary. Our whole-genome sequencing variant dataset has been contributed to the HRC project as well, and will be further used for more analysis for colorectal cancer.

Figures and Tables

Figure 3.1 LocusZoom on the two regions with cluster of SNPs of $P < 5 \times 10^{-7}$

(A) The region at 1p21.3. Lead SNP rs841684 with $P = 1.5 \times 10^{-7}$. In total, there are 8 SNPs with $P = 1.5 \times 10^{-7}$, all in the same LD block. Extensive LD is observed in this region. This cluster is downstream of F3 coding region.



(B) The region at 6p21.31. Lead SNP rs12529688 with $P = 2.5 \times 10^{-7}$. There are 4 SNPs with $P = 1.5 \times 10^{-7}$, all in the same LD block, with multiple genes clustered in this block. Three out of the four SNPs are upstream of FKBP5, and one is intronic to the promoter region of FKBP5.

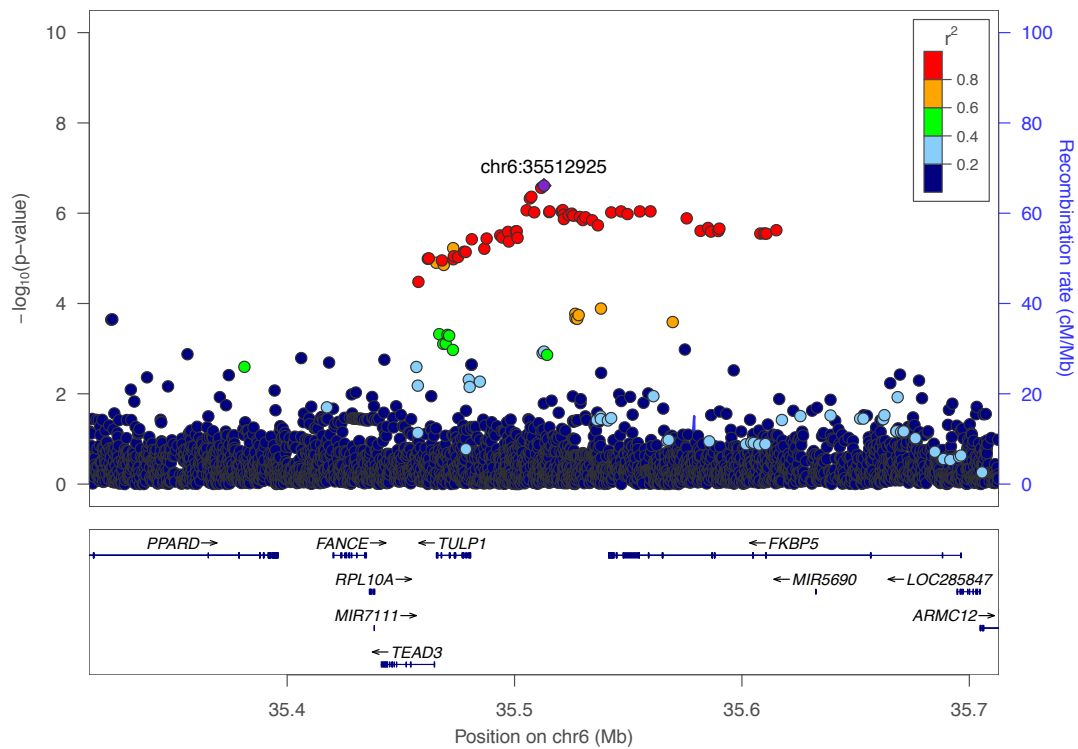


Figure 3.2 Sequencing characteristics of GECCO WGS samples

We show distributions of (A) mean depth (B) medium insert size (C) standard deviation of insert size. Overall the distributions of sequencing characteristics are consistent, with no obvious outliers. Year 3 samples are deeply sequenced, with more consistent insert size than the low-coverage sequenced samples from Year 1 and Year 2.

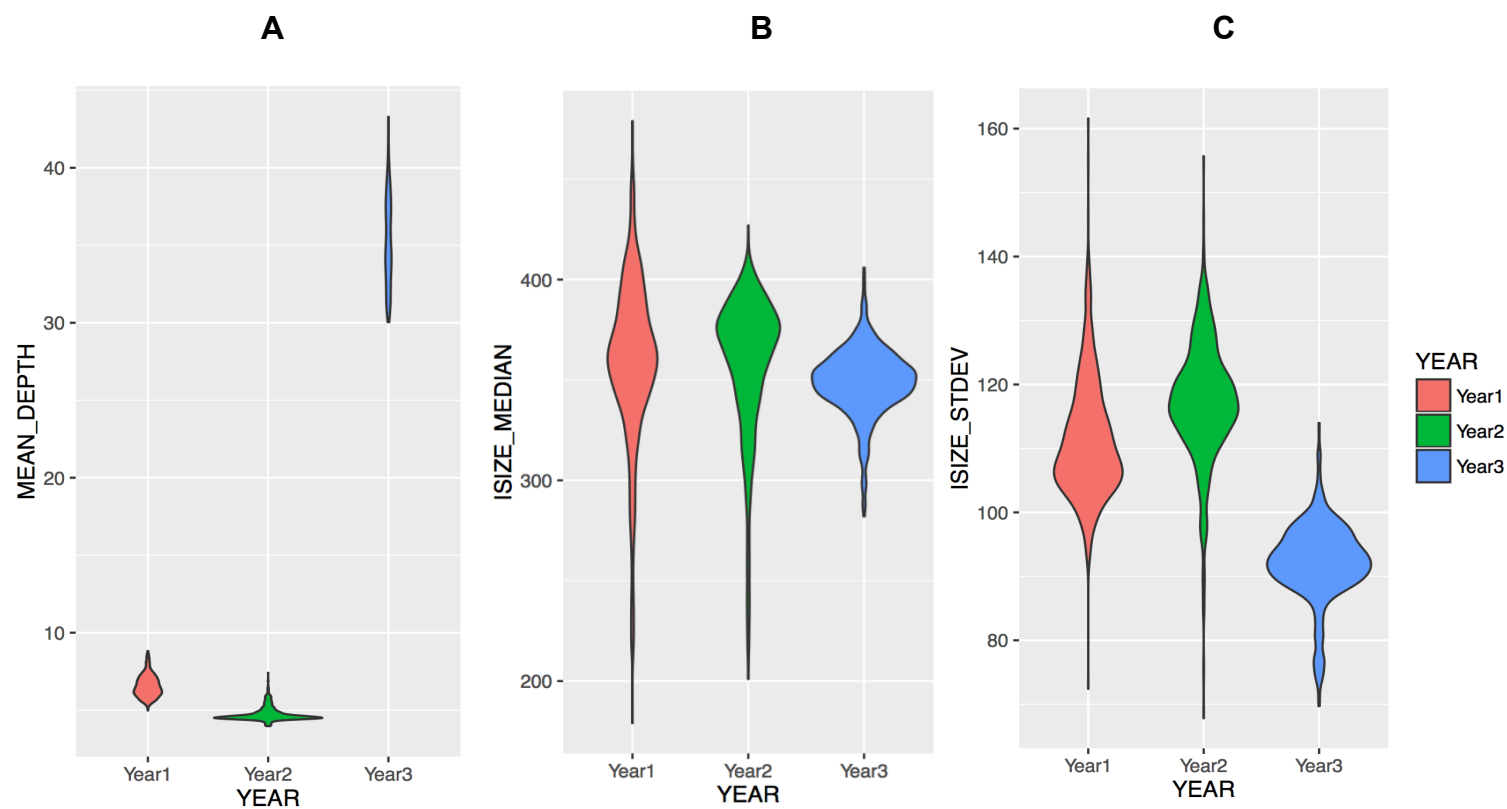


Table 3.1 Newly identified SNPs associated with colorectal cancer with P values less than 5×10^{-7}

Region	SNP	Position ^a	Allele ^b	MAF	P value	Imputation R2	Gene ^c	Annotation ^c
SNP with $P < 5 \times 10^{-8}$								
7p12.1	rs115561508	53226189	C/T	0.10	3.60E-09	0.43	<i>POM12L1</i> 2	Intergenic
9q21.12	rs138140376	73207739	C/T	3.00E-03	1.90E-09	0.61	<i>TRPM3</i>	Intron
SNP with $P < 5 \times 10^{-7}$								
1p21.3	rs841684	95057293	T/C	0.47	1.48E-07	0.95	<i>F3</i>	Intergenic
	rs841689	95059247	G/A		2.00E-07			
	rs1098724	95064512	T/C		1.58E-07			
	rs1098725	95064950	A/G		1.58E-07			
	rs2798940	95065092	C/T		2.04E-07			
	rs866365	95065110	G/C		2.12E-07			

	rs841708	95065633	T/G		1.49E-07			
	rs1772895	95068347	A/G		4.32E-07			
2p24.2	rs78115417	190304106	C/T	7.20E-03	4.10E-07	0.53	<i>WDR75</i>	Intergenic
5q35.1	rs555044933	170813604	A/C	3.60E-03	1.70E-07	0.47	<i>MIR3912,</i> <i>NPM1</i>	Intergenic
6p21.31	rs16877540	35506731	A/C	0.085	4.70E-07	0.99	<i>FKBP5</i>	Intergenic
	rs76489311	35507418	G/A		4.32E-07			Intergenic
	rs146718198	35511735	G/A		2.74E-07			
	rs12529688	35512925	C/T		2.50E-07			Intron

^aPosition is based on assembly GRCh37

^bAllele is annotated as reference allele / alternative allele

^cAnnotated with UCSC Genome Brows

Table 3.2 Summary of Year 3 variant calls

(A) Summary of variant sites detected from 3,061 individuals across all years. dbSNP Ts/Tv was calculated from the variants found in dbSNP, while the novel Ts/Tv was calculated from the variants not found in dbSNP.

#SNP	Overlap with dbSNP b145	dbSNP Ts/Tv	Novel Ts/Tv	Overlap with 1000G Phase 3
46.1 M	58.2%	2.37	1.72	46.4%
#Indel	Overlap with dbSNP b145	Insertion/ Deletion	Frameshift	Overlap with 1000G Phase 3
2.64 M	57.0%	0.37	0.25%	40.6%

(B) Average numbers of variants across all 3,061 individuals.

Type	#Variants ^a	#Singleton	#Doubleton	#HET ^b	#ALT ^c
SNP	3.43 M	4,708	3,130	2.11 M	1.33 M
Indel	216 K	198	107	139 K	77 K

^a Average number of variants in one individual

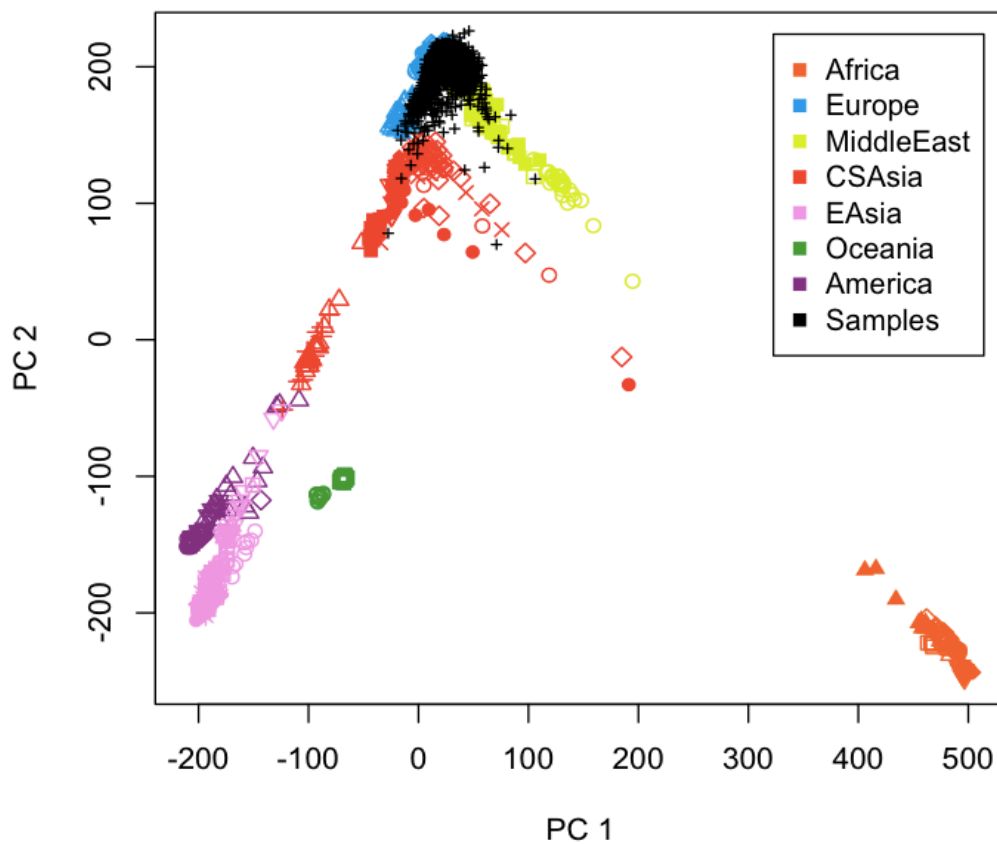
^b Average number of variants with heterozygous genotype in one individual

^c Average number of variants with homozygous alternative genotype in one individual

Supplementary Figures

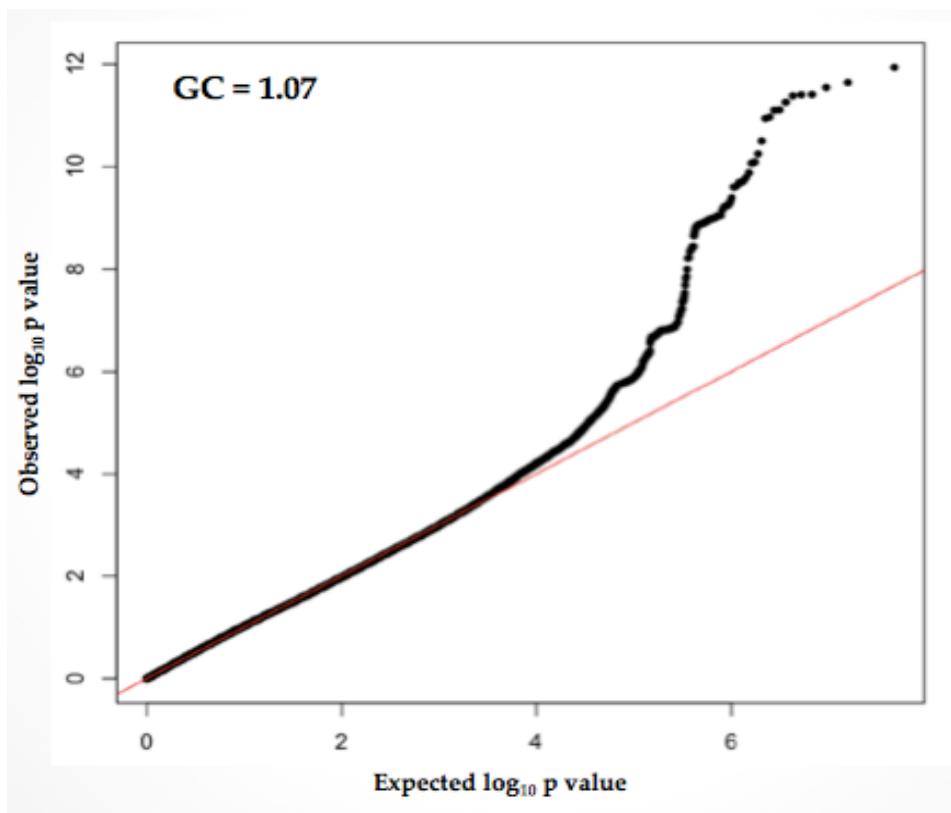
Supplementary Figure 3.1 Ancestry inference of the 26,903 individuals

Dark spots represent the tested individuals while the other colors represent the reference population from Human Genome Diversity Project (HGDP) reference panel. Most individuals are estimated as Europeans, with only a few outliers at the cluster of Middle East and Central Asia, but not far from the European clusters.



Supplementary Figure 3.2 QQ plot of meta-analysis p values across all 22 million imputed variants

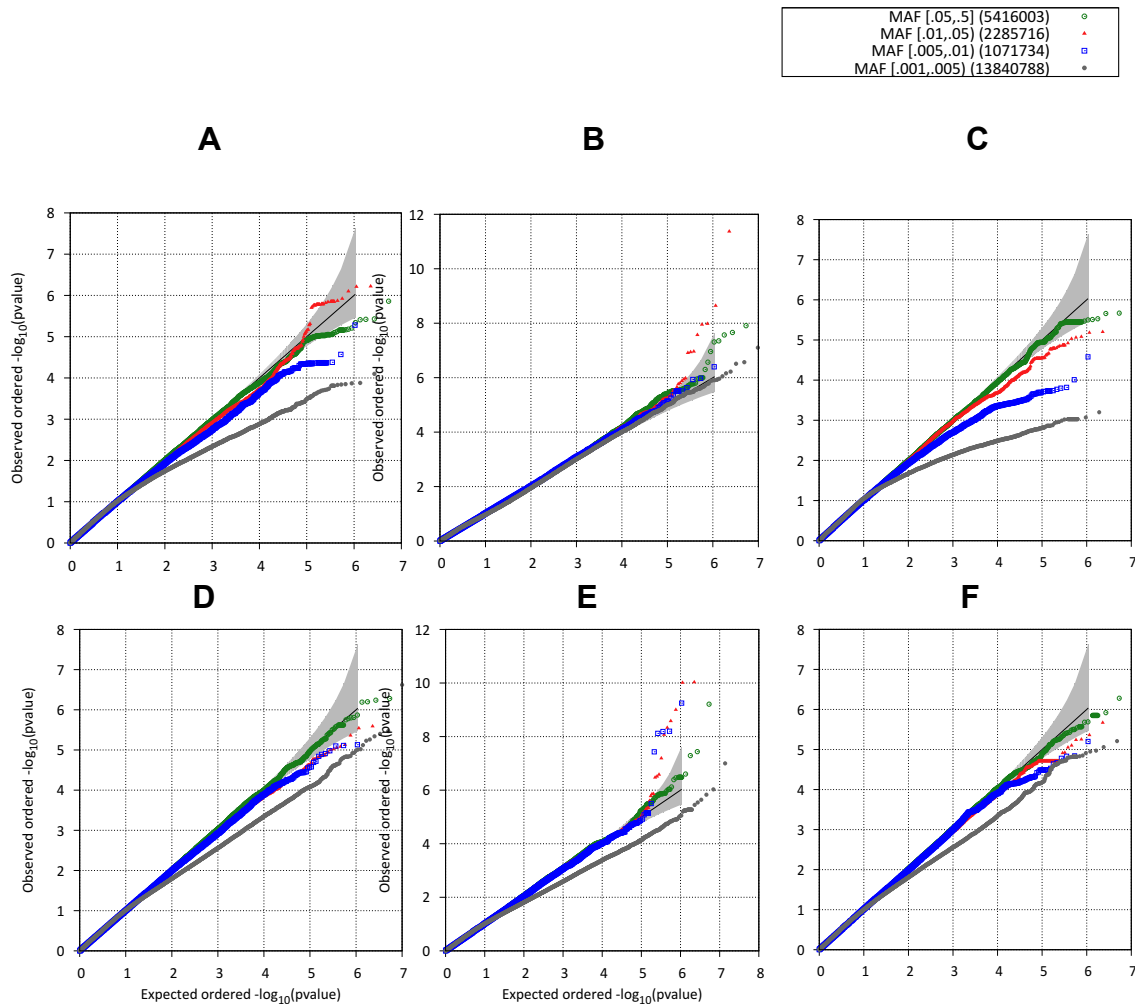
The meta-analysis has a well-controlled inflation rate. The significant signals mostly come from the known risk loci associated with colorectal cancer (shown in Supplementary Table 3-1)



Supplementary Figure 3.3 QQ plot for single variant association analysis on each genotyping platform

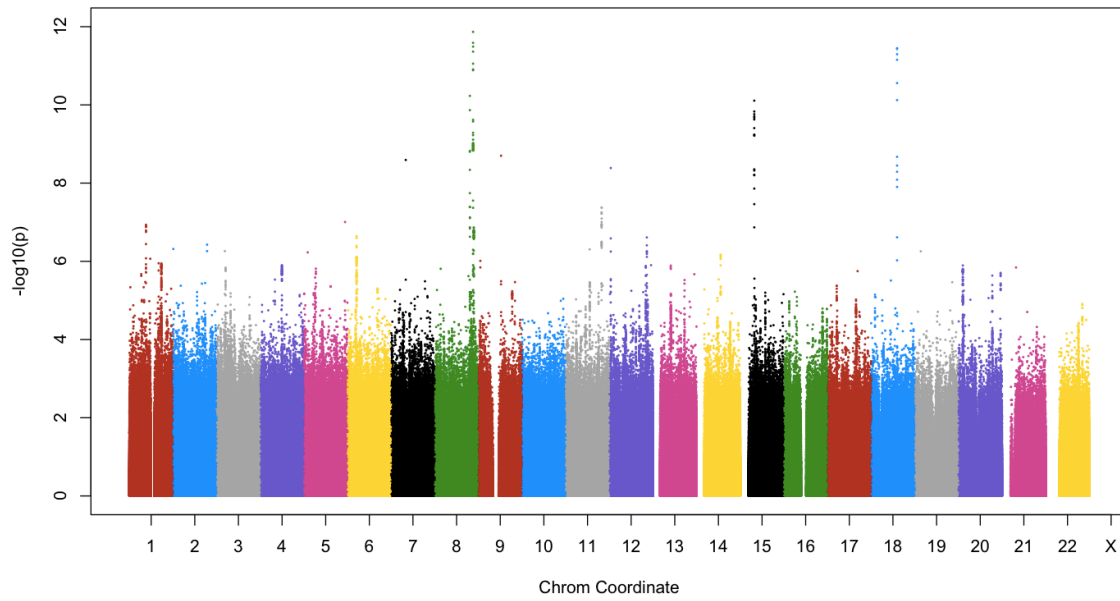
- (A) AffyMetrix (n=1,124) (B) Initial (n=5,924)
 (C) CCFR USC (n=2,151) (D) OmniExpress (n=5,986)
 (E) CytoSNP (n=10,908) (F) CCFR Subset3 (n=811)

We observed strong association signals in larger dataset (B) (D) and (E), especially in common variants. Due to the limitation of sample size, p values of rare variants are deflated, especially in smaller studies (A) (C) and (E)



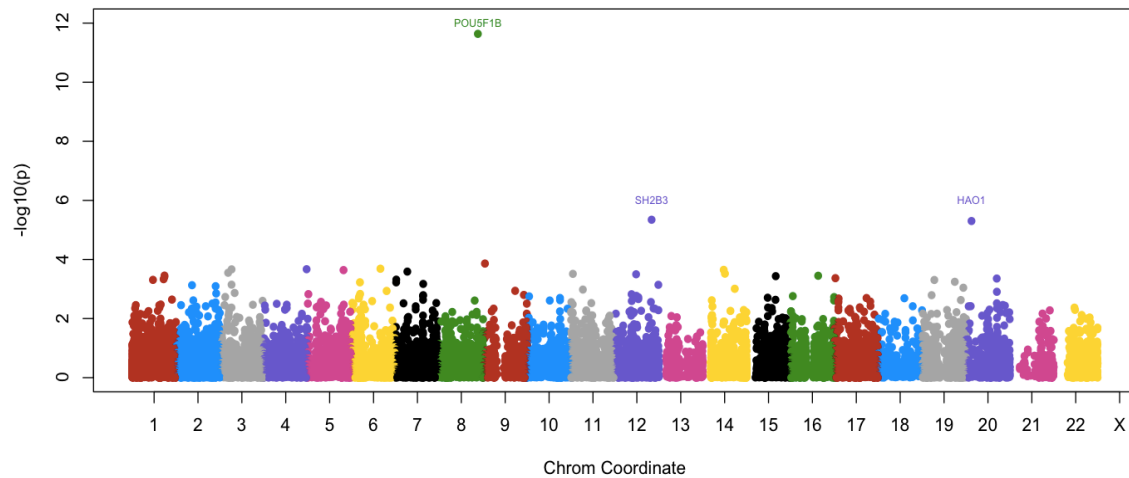
Supplementary Figure 3.4 Manhattan plot of single-variant meta-analysis

22 million variants from were tested. We observed strong cluster of signals at known risk regions, such as the 8q24, 15q13 and 18q21.



Supplementary Figure 3.5 Manhattan plot of the gene-based test

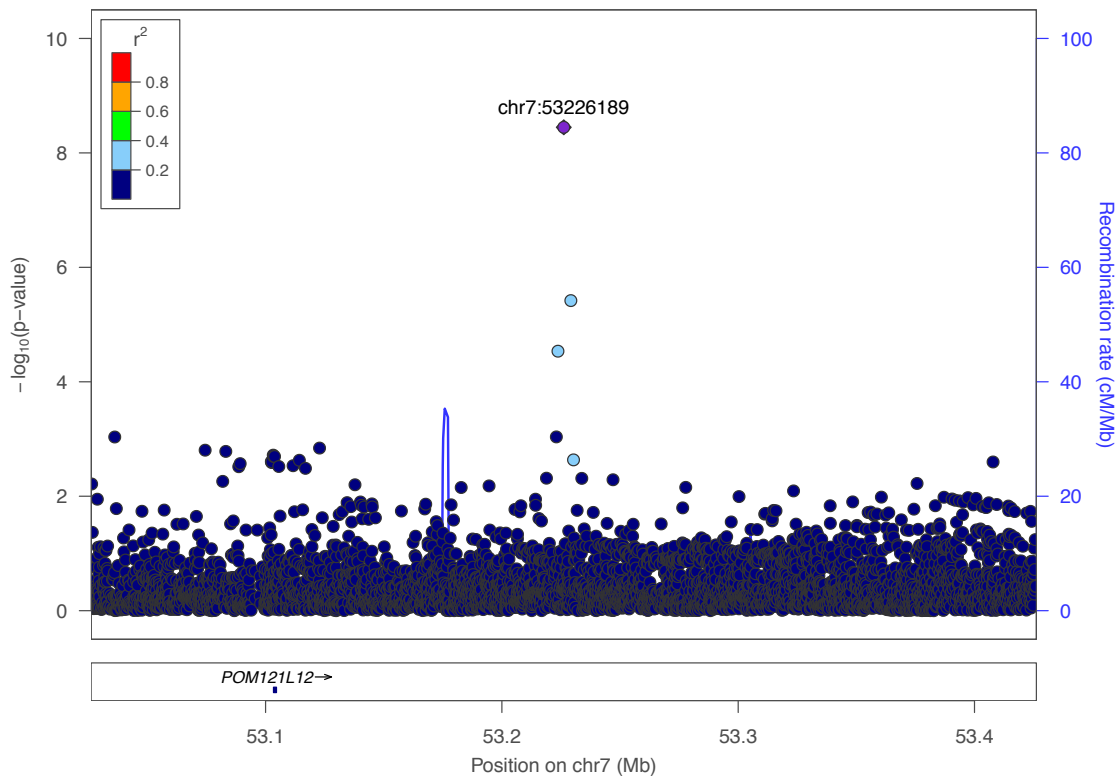
18,677 genes were tested with Burden test assuming equal weights. We observed strong signals as known gene POU5F18 and SH2B3, and a novel gene HAO1



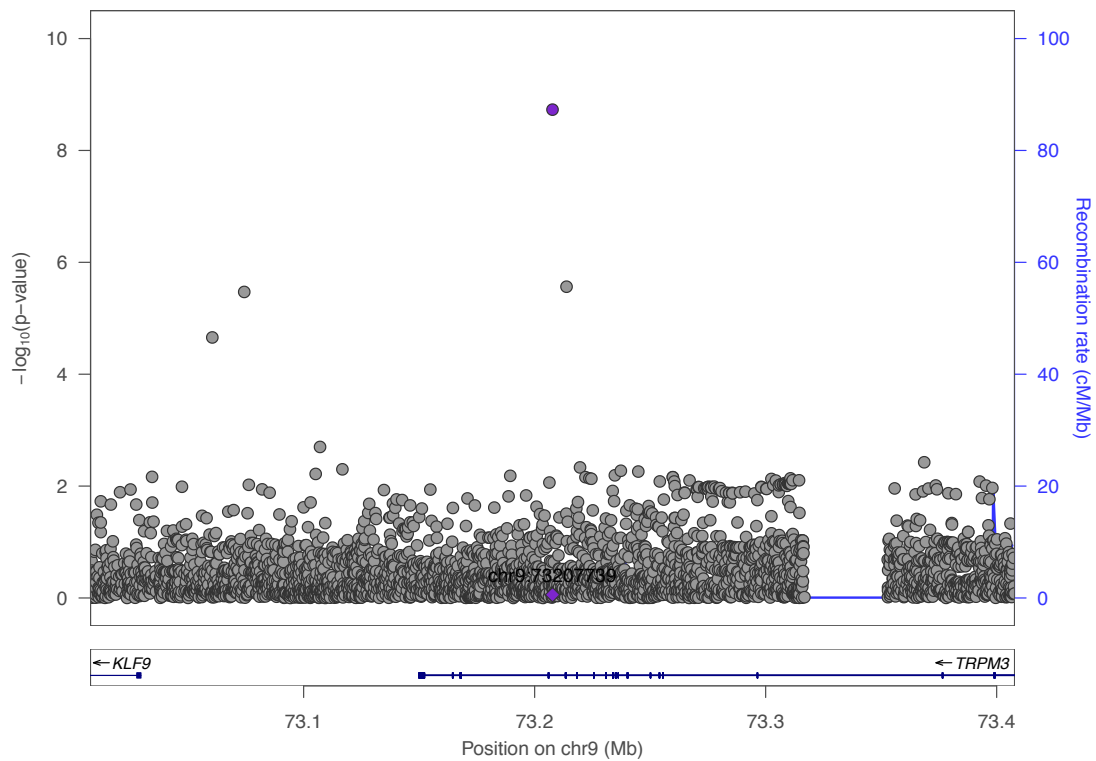
Supplementary Figure 3.6 LocusZoom plots on the rest four regions with $P < 5 \times 10^{-7}$

There lacks other significant signals in these four regions except for the lead SNP. No LD block was observed in these regions.

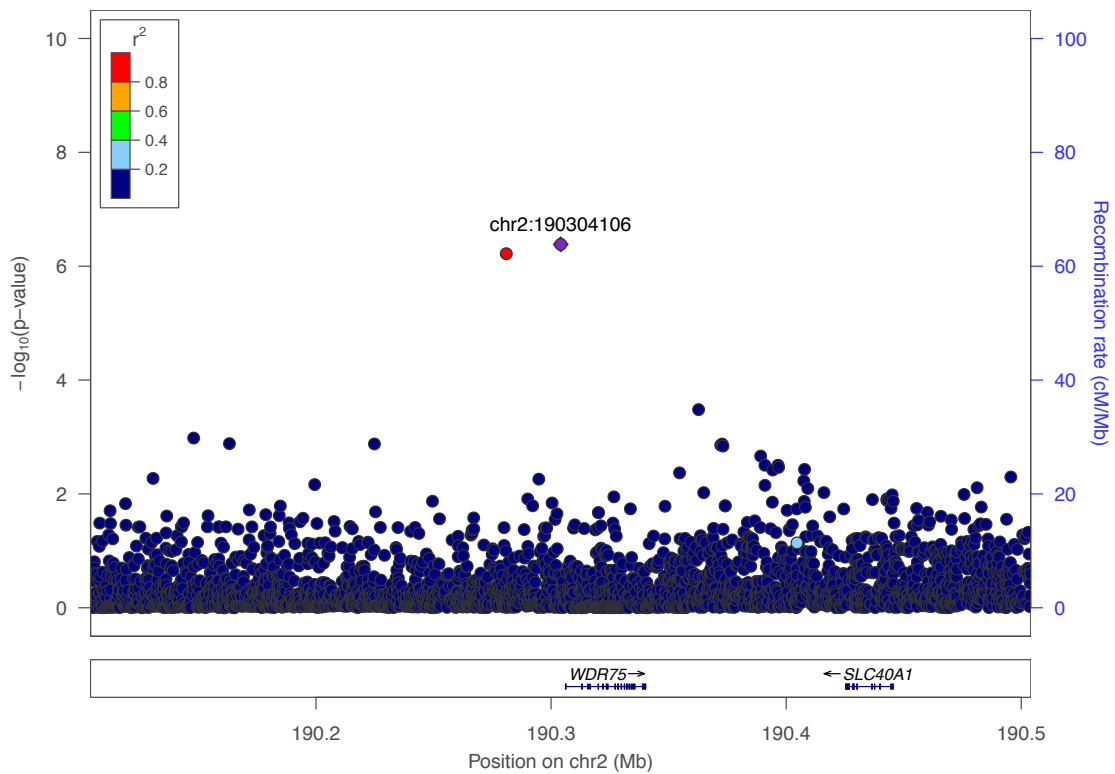
(A) The region at 7p12.1. The lead SNP is rs115561508 with $P = 3.6 \times 10^{-9}$ and MAF = 0.10. The lead SNP is in the intergenic region downstream of POM121L12.



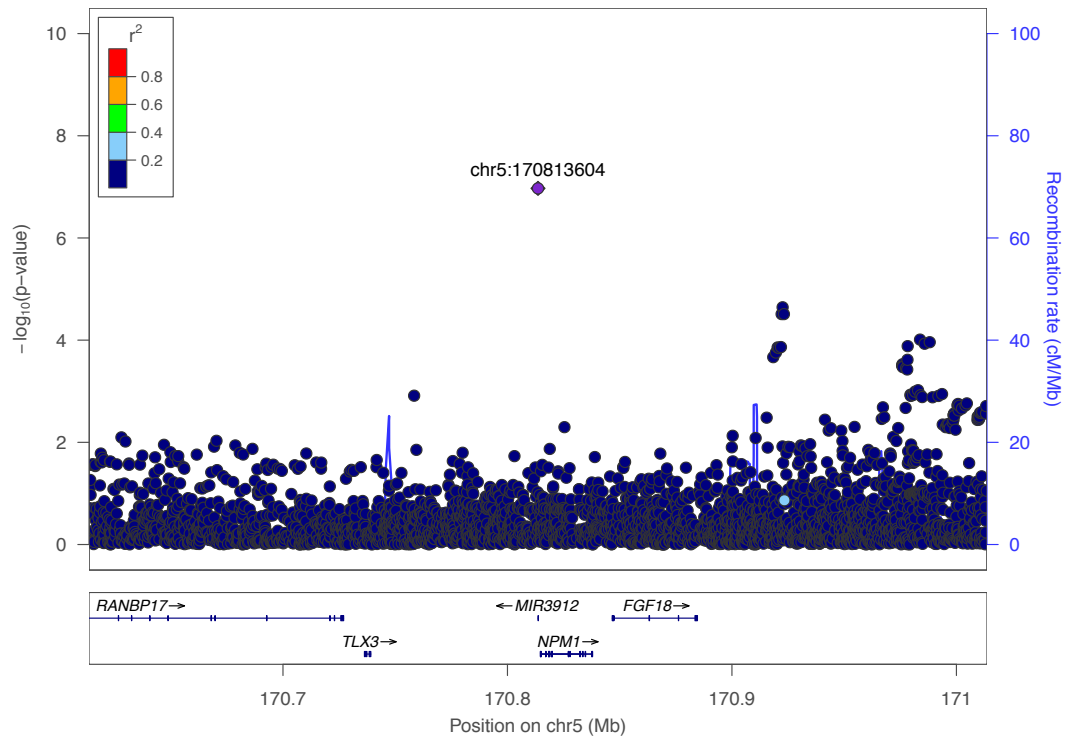
(B) The region at 9q21.12. The lead SNP is rs138140376 with $P = 1.9 \times 10^{-9}$ and MAF = 0.03%. The lead SNP, located at the coding region of TRPM3, is a multi-allelic SNP. Its other allele is not significant in meta-analysis, and is more rare in MAF.



(C) The region at 2p24.2. The lead SNP is rs78115417 with $P = 4.1 \times 10^{-7}$ and MAF = 0.7%. The lead SNP is intergenic and upstream of WDR75.

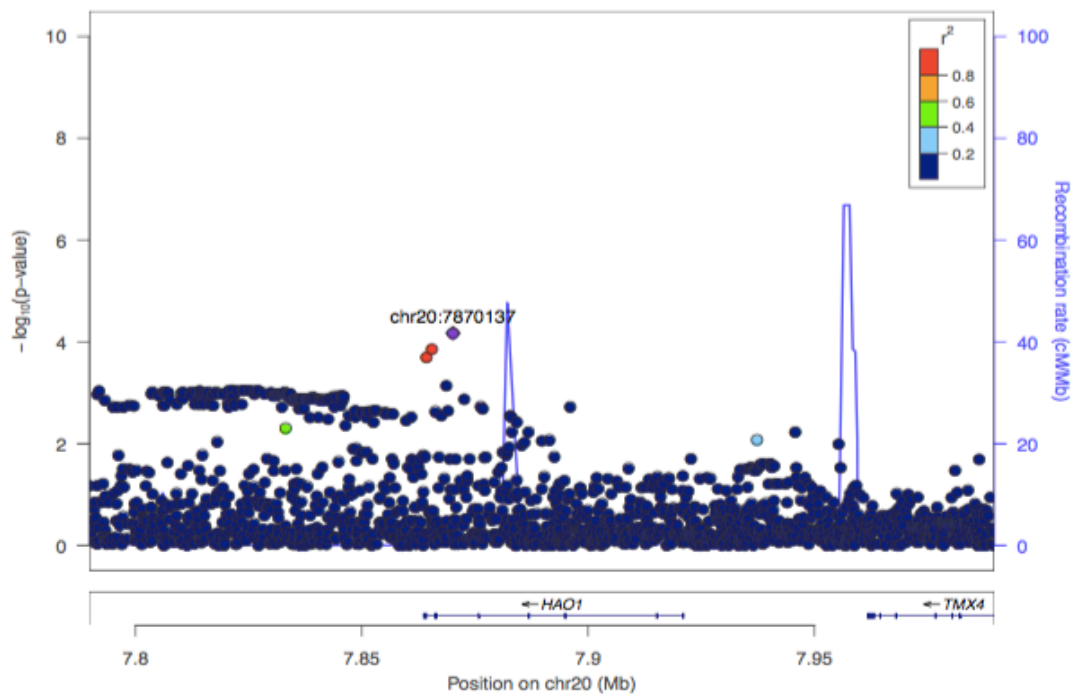


(D) The region at 5q35.1. The lead SNP is rs55504933 with $P = 1.7 \times 10^{-7}$ and MAF = 0.4%, at intergenic region. This region contains multiple genes.



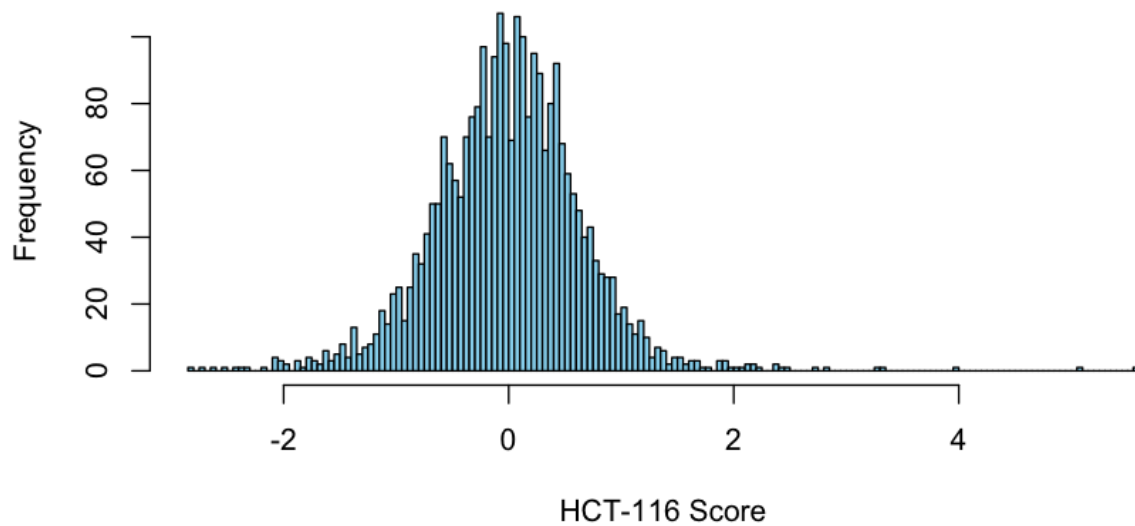
Supplementary Figure 3.7 LocusZoom plot of HAO1

The gene HAO1, with $P = 5.0 \times 10^{-6}$ includes 10 coding SNPs in gene-based test. None of the SNPs showed evidence of significance in meta-analysis. The significant P value of the gene is due to the concordant effect direction and relatively small P values from all 10 variants.



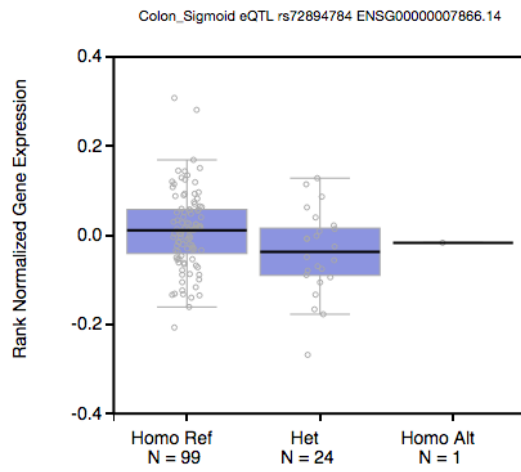
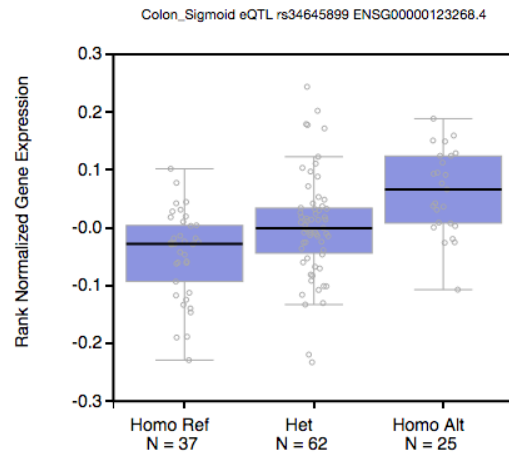
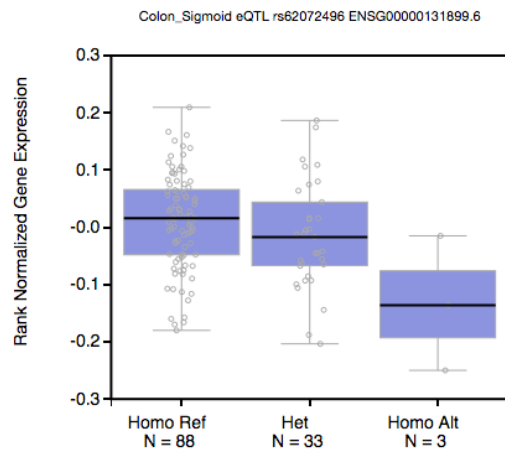
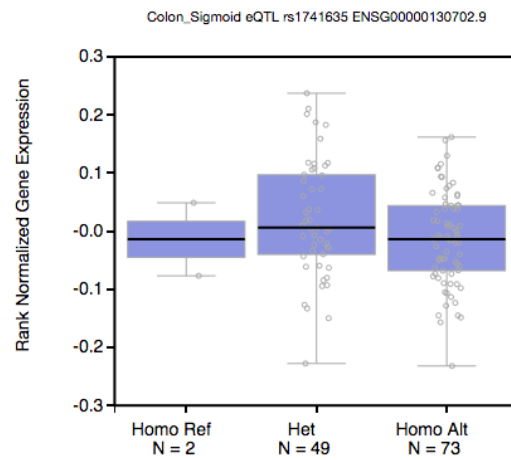
Supplementary Figure 3.8 Distribution of HCT-116 deltaSVM scores

With a model built from ENCODE colon cancer cell line HCT-116 DNaseI sensitivity peak data, we generated deltaSVM scores for 3,000 SNPs from the GECCO HRC imputed dataset. SNPs were selected based on LD pruning from meta-analysis P values.



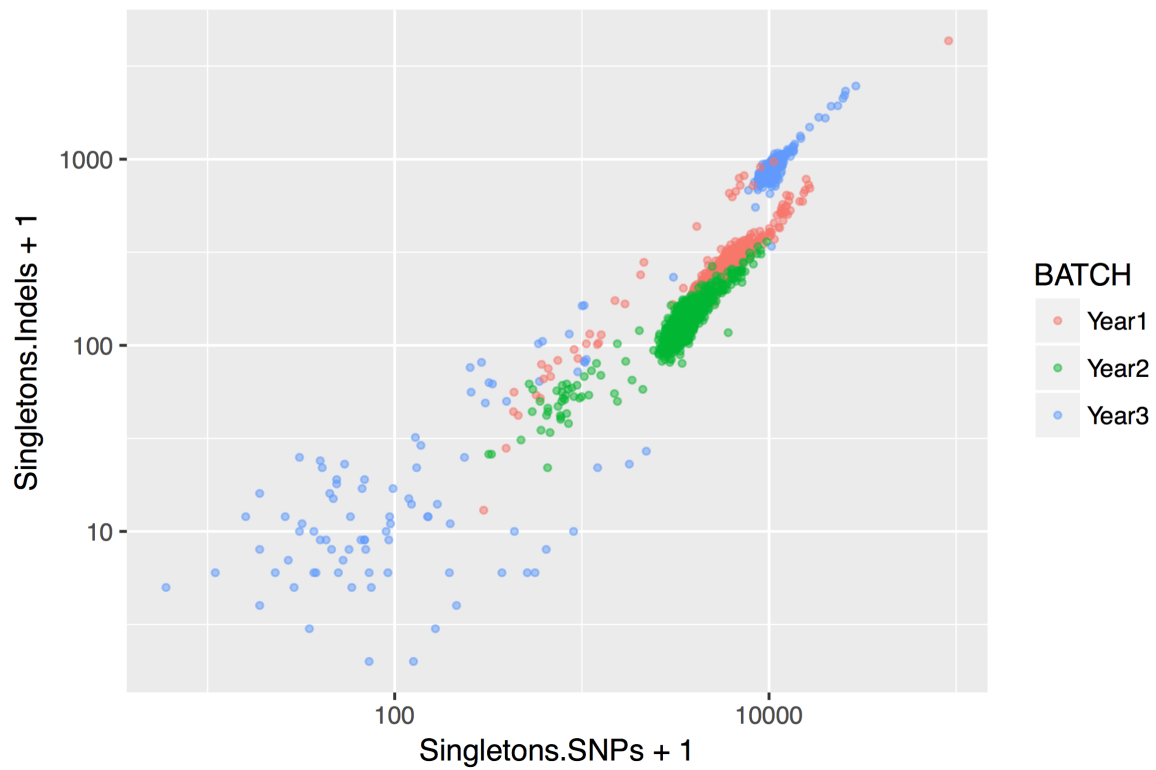
Supplementary Figure 3.9 eQTL expression of the nearby genes of the 4 variants with $P < 0.05$ at Colon Sigmoid

- (A) rs72894784 tested on *TEAD3*. Meta-analysis $P = 1.0 \times 10^{-5}$, eQTL $P = 3.3 \times 10^{-2}$. Heterozygotes have lower expression level, while data is insufficient to decide expression level of homozygotes.
- (B) rs34645899 tested on *ATF1*. Meta-analysis $P = 8.7 \times 10^{-5}$, eQTL $P = 3.3 \times 10^{-2}$. Significant expression level differences were observed across the three genotypes, with the alternative allele positively associated with expression level.
- (C) rs62072496 tested on *LLGL1*. Meta-analysis $P = 7.2 \times 10^{-5}$, eQTL $P = 1.1 \times 10^{-2}$. Substantial lower expression was observed in heterozygotes compared with the reference genotype.
- (D) rs1741635 tested on *LAMA5*. Meta-analysis $P = 2.4 \times 10^{-6}$, eQTL $P = 4.5 \times 10^{-2}$. Homozygous alternative genotypes have substantial lower expression level than heterozygotes, but data is insufficient for the reference genotype.

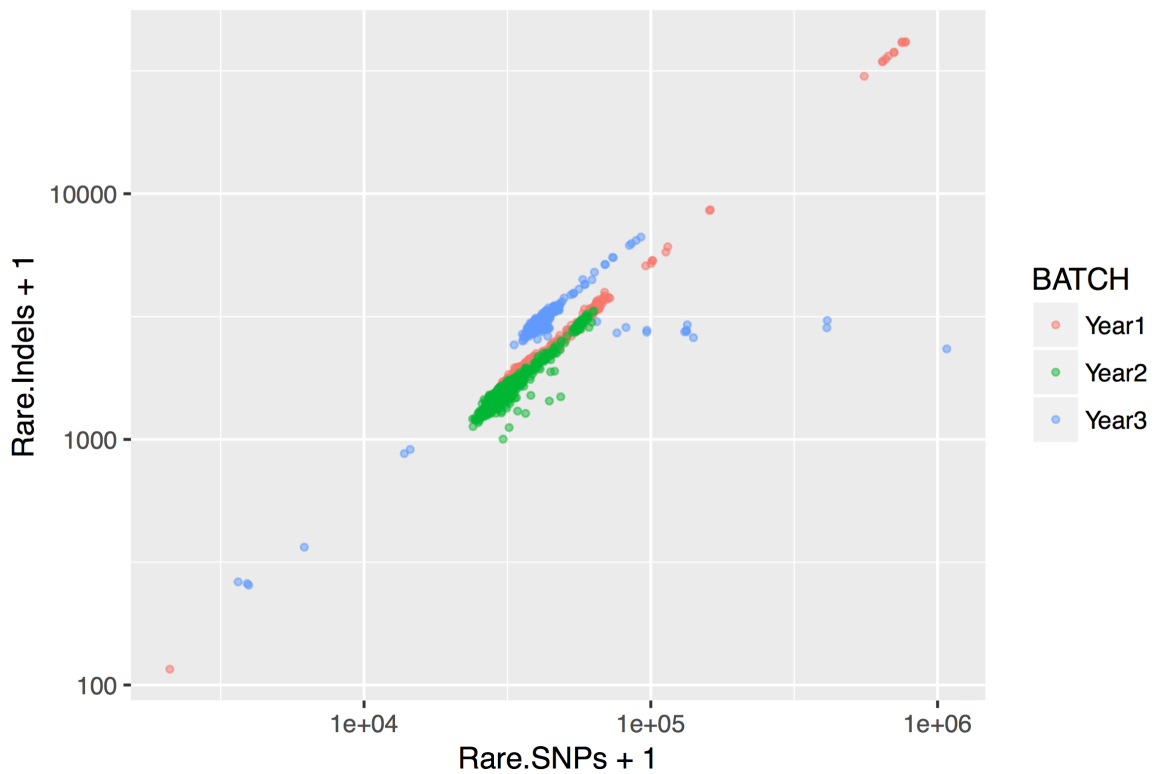
A**B****C****D**

Supplementary Figure 3.10 Singleton and rare variant counts per individual

(A) Singletons of SNPs and Indels per individual. Year 3 (deeply sequenced) has more singletons discovered than Year 1 and Year2 (low-coverage sequenced). Population outliers were excluded.



(B) Rare variant (MAF < 0.5%) counts per individual. Year 3 (deeply sequenced) has more rare variants discovered. African Americans in Year 1 have high counts of rare variants. A few samples have been truncated during BWA MEM remapping, and as a result, show extremely low counts of rare variants.



Supplementary Tables

Supplementary Table 3.1 Study information

Study	Abbreviation	Design	Country	#Cases	#Controls	Mean Age	%Female	Genotyping platform
Colon Cancer Family Registry	CCFR subset3	Case-control	United States, Canada,	398	413	52	54	Human1Md
USC	CCFR USC	Sib-pair	Australia	1171	980	54	50	Unknown
Assessment of Risk in Colorectal Tumors In Canada	ARCTIC	Case-control	Canada	769	665	65	52	Affy Chips
Diet, Activity and	DALS1	Case-control	United States	410	464	65	45	CytoSNP

Darmkrebs: Chancen der Verhütung durch Screening	DACHS	Case-control	Germany	2376	2206	69	40	
Colorectal Cancer Studies 2&3	Hawaiian Colo2&3	Case-control	United States	87	125	65	45	
VITamins And Lifestyle	VITAL	Cohort	United States	285	288	67	48	
Postmenopau sal Hormone study	PMH	Case-control	United States	280	122	65	100	
Association STudy Evaluating RISK for sporadic	ASTERISK	Case-control	France	948	947	65	41	

colorectal cancer								
Multiethnic Cohort Study	MEC	Cohort	United States	328	346	63	46	
Nurses' Health Study	NHS	Cohort	United States	553	955	60	100	OminiExpress
Nurses' Health Study, Adenoma Set	NHS Ad	Cohort	United States	513	578	57	100	
Physicians' Health Study	PHS	Cohort	United States	382	389	58	0	
Health Professionals Follow-up Study	HPFS	Cohort	United States	403	402	65	0	
Health Professionals' Follow-up	HPFS Ad	Cohort	United States	313	345	61	0	

Study, Adenoma Set								
Women's Health Initiative	WHI	Cohort	United States	1476	2538	67	100	Initial
Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO	Cohort	United States	1019	2391	64	31	
Diet, Activity and Lifestyle Study	DALS2	Case-control	United States	706	710	65	45	

Supplementary Table 3.2 Meta-analysis results on previously identified CRC risk loci

Position/Gene	SNP	Position^a	Allele^b	MAF	P value	R2^c	Population
1q25.3/LAMC1	rs10911251	183081194	A/C	0.43	5.5E-06	0.92	European
1q41	rs6691170	222045446	T/G	0.03	3.6E-01	0.91	European
	rs6687758	222164948	G/A	0.20	2.4E-02	1.00	
2q32.2	rs11903757	192587204	C/T	0.16	2.6E-03	0.85	East Asian
3p14.1/LRIG1	rs812481	66442435	G/C	0.48	5.1E-04	0.99	Asian, European
3p22.1/CTNNB1	rs35360328	40924962	A/T	0.15	1.4E-06	0.94	European
3q26.2/TERC	rs10936599	169492101	C/T	0.24	5.2E-01	0.99	European
6p21	rs1321311	36622900	A/C	0.25	1.9E-01	1.00	African American
6q26.27/SLC22A3	rs7758229	160840252	T/G	0.32	4.5E-01	1.00	Asian
8q23.3/EIF3H	rs16892766	117630683	C/A	0.08	5.6E-11	0.99	European
8q24/MYC	rs10505477	128407443	A/G	0.49	7.9E-12	0.97	European

	rs6983267	128413305	G/T	0.48	2.8E-12	0.98	
	rs7014346	128424792	A/G	0.38	1.1E-11	1.00	
9p24	rs719725	6365683	A/C	0.38	8.4E-05	0.99	European
10p14	rs10795668	8701219	G/A	0.31	4.4E-02	1.00	European
10q22.3/ZMIZ1- AS1	rs704017	80819132	G/A	0.42	8.7E-05	0.92	East Asian, European
10q24.2/SLC25A2 8/ENTPD7/COX1 5/CUTC	rs11190164	101351704	G/A	0.26	2.8E-04	0.97	European
10q25.2/VTI1A/T CF7L2	rs12241008	114280702	C/T	0.09	4.2E-03	0.95	East Asian, African American
	rs11196172	114726843	A/G	0.14	8.2E-02	0.95	East Asian
11q12.2/MYRF/F EN1/FADS1/FAD S2	rs174537	61552680	G/T	0.32	1.4E-04	0.95	
	rs4246215	61564299	G/T	0.34	5.4E-04	0.95	
	rs174550	61571478	T/C	0.32	8.6E-05	0.97	

	rs1535	61597972	A/G	0.33	9.3E-05	0.99	
11q13.4/POLD3	rs3824999	74345550	G/T	0.48	1.6E-05	1.00	Asian, European
11q23	rs3802842	111171709	G/T	0.29	1.7E-07	0.97	East Asian, West Asian, European
12p13.32	rs10774214	4368352	T/C	0.38	9.1E-03	0.96	East Asian
	rs3217810	4388271	T/C	0.13	3.8E-09	0.79	
	rs3217901	4405389	G/A	0.42	2.4E-06	0.90	
	rs10849432	6385727	T/C	0.11	2.5E-02	0.97	
12q13.13	rs7136702	50880216	T/C	0.35	1.5E-03	0.87	European
	rs11169552	51155663	C/T	0.26	1.3E-02	1.00	
	rs3184504	111884608	C/T	0.49	1.1E-06	0.99	
	rs59336	115116352	T/A	0.49	3.4E-06	0.96	
	rs73208120	117747590	G/T	0.09	1.3E-05	0.94	
14q22.2/BMP4	rs4444235	54410919	C/T	0.48	4.4E-05	0.99	European,

							East Asian
	rs1957636	54560018	T/C	0.40	1.4E-02	1.00	
15q13/CRAC1/H MPS/GREM1	rs16969681	32993111	T/C	0.19	1.9E-02	0.98	European
	rs4779584	32994756	T/C	0.20	4.5E-09	0.97	
	rs11632715	33004247	A/G	0.48	3.3E-03	0.99	
16q22.1/ <i>CDH1</i>	rs9929218	68820946	G/A	0.29	5.9E-03	1.00	European
17p13.3	rs12603526	800593	C/T	0.02	1.4E-02	0.86	East Asian
18q21/ <i>SMAD7</i>	rs4939827	46453463	T/C	0.47	8.1E-11	0.98	European
19q13.1/ <i>RHPN2</i>	rs10411210	33532300	C/T	0.10	2.4E-02	0.99	African American, Asian
19q13.2/TGFB1	rs1800469	41860296	G/A	0.31	1.4E-01	0.99	East Asian
	rs2241714	41869392	C/T	0.32	1.5E-01	0.97	
20p12.3/BMP2	rs961253	6404281	A/C	0.36	1.1E-05	0.99	East Asian, European
	rs4813802	6699595	G/T	0.35	1.0E-05	0.94	

20q13.3	rs6066825	47340117	A/G	0.36	1.0E-02	0.98	European
20q13.33/ <i>LAMA5</i>	rs4925386	60921044	C/T	0.30	4.6E-03	1.00	European

^aPosition is based on assembly GRCh37

^bAnnotated as reference allele / alternative allele

^cImputation R-square from Minimac3

Supplementary Table 3.3 SNPs included in the Burden test of HAO1

Ten coding SNPs were included in the Burden test of gene HAO1. None of them is genome-wide significant, but over half of them have low p values in single-variant test.

SNP	Position	Allele	MAF	Single-variant p value
rs34249643	7864284	T/C	0.024	2.0E-03
rs138358725	7866188	C/A	8.0E-04	0.72
rs573481520	7866189	G/A	0.033	2.3E-03
rs201216901	7866210	C/T	0.033	2.3E-03
rs150881591	7866376	T/C	4.0E-03	0.66
rs139675589	7875820	G/A	4.0E-03	0.34
rs146825169	7886834	A/G	2.4E-03	6.0E-03
rs142998832	7886869	C/T	2.0E-03	0.80
rs201859601	7915210	C/A	0.030	0.044
rs147089441	7915212	T/C	0.030	0.044

Supplementary Table 3.4 Parameter estimates from fGWAS model

(A) Ridge parameter estimates from fGWAS model.

fGWAS model was first built with six annotations separately: exon, intron, nonsynonymous, synonymous, stop gain, stop loss. Then annotations were added and dropped under a cross-validation model, until the model likelihood was maximized. Penalty was adjusted to maximize an ridge likelihood. The final model included three annotations: exon, synonymous and stop gain, with exon showing the highest enrichment, implying its stronger effect.

Parameter	Enrichment
Penalty	0.10
Exon	1.44
Synonymous	-0.52
Stop Gain	-0.011

(B) Tissue-specific enrichment in DNaseI sensitive regions, estimated by fitting fGWAS models using peak information from each cell line separately. For HCT-116 and CACO-2, we only used consensus peaks between the two replicates. From the fGWAS models, the colon cancer cell line HCT-116 has higher enrichment than the control cell line GM12878. No significant enrichment was observed in another colon cancer cell line CACO-2, possibly because of the low data quality and inconsistency of peaks between its two replicates

Cell line	# Peaks	Enrichment (95%CI)
GM12878 (Duke)	121127	3.1E-12 (6.3E-21~67)
HCT-116	93332	0.54 (2.1E-09~99)
CACO-2	55623	3.1E-11 (6.4E-20~213)

Supplementary Table 3.5 SNPs and genome blocks with causality predicted by fGWAS

PPA > 0.2 was used as the threshold for screening potentially causal genome blocks. If the chunk doesn't have such SNPs with PPA > 0.2, we show the SNP with highest PPA in this chunk.

Region	Chunk	Chunk PPA	SNP	SNP PPA	SNP P value
1p22.1-1p21.3	94644828-95289080	0.36	rs841684	0.16	1.5E-07
4q22.2	94491587-95040993	0.20	rs2618731	0.01	1.1E-06
6p21.31	35219926-35785860	0.36	rs11545925	0.04	9.6E-07
7p12.1	53086094-53536962	0.74	rs115561508	0.73	3.6E-09
8q23.3-8q24.11	117072454-117720230	1.00	rs16892766	0.65	5.6E-11
			rs16888589	0.28	1.3E-12
8q24.11	117720250-118306941	0.32	rs139444083	0.13	6.7E-08
8q24.21	127948781-128453248	1.00	rs12682374	0.37	1.2E-12
8q24.21	130622915-131195299	0.76	rs62525036	0.02	1.2E-07
9q21.12	72785759-73425880	0.84	rs138140376	0.83	1.9E-09

15q13.3	32378406-33312613	1.00	rs1406389	0.28	2.0E-10
			rs2293581	0.23	2.5E-10
18q21.1	46285923-46858626	1.00	rs4939567	0.21	5.5E-12
			rs2337113	0.30	3.9E-12
			rs11874392	0.28	4.1E-12

Supplementary Table 3.6 eQTL test on predicted DNaseI sensitive sites

11 SNPs were tested in Colon Sigmoid at GTEx Portal Website. Nearest genes were decided through HaploReg v4.1. Four sites have eQTL $P < 0.05$.

Region	SNP	Position	MAF	P value	Score	eQTL P value	Nearest gene
3p22.2	rs11129737	36890461	0.26	8.97E-05	2.18	0.37	<i>TRANK1</i>
6p21.31	rs72894784	35462305	0.08	9.99E-06	2.45	0.033	<i>TEAD3</i>
	rs146718198	35511735		2.74E-07	-2.32	0.59	<i>FKBP5</i>
8q24.11	rs117982378	117707559	0.04	1.45E-05	-2.24	0.34	<i>EIF3H</i>
8q24.21	rs62524989	130817869	0.20	1.23E-06	2.64	0.73	<i>GSDMC</i>
	rs62525041	130830724	0.18	2.09E-07	3.32	0.31	<i>RP11-473O4.5</i>
12p13.2	rs145997566	12785302	0.03	9.28E-05	-2.22	0.32	<i>CREBL2</i>
12q13.12	rs34645899	51201749	0.40	8.68E-05	2.30	2.6E-06	<i>ATF1</i>
17p11.2	rs62072496	18124743	0.20	7.18E-05	-2.19	0.011	<i>LLGL1</i>

17q22	rs7226124	55030689	0.18	5.44E-05	2.21	0.37	<i>COIL</i>
20q13.33	rs1741635	60938197	0.21	2.35E-06	2.17	0.045	<i>LAMA5</i>

References

1. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival. *JNCI: Journal of the National Cancer Institute* **109**, (2017).
2. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**, 984–988 (2007).
3. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
4. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317 (2007).
5. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40**, 631–637 (2008).
6. Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.* **40**, 26–28 (2008).
7. Tomlinson, I. P. M. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008).
8. COGENT Study *et al.* Meta-analysis of genome-wide association data

- identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435 (2008).
9. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42**, 973–977 (2010).
 10. Tomlinson, I. P. M. *et al.* Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet.* **7**, e1002105 (2011).
 11. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat. Genet.* **44**, 770–776 (2012).
 12. Savinainen, K. J. *et al.* Expression and copy number analysis of TRPS1, EIF3S3 and MYC genes in breast and prostate cancer. *Br. J. Cancer* **90**, 1041–1046 (2004).
 13. Okamoto, H., Yasui, K., Zhao, C., Arii, S. & Inazawa, J. PTK2 and EIF3S3 genes may be amplification targets at 8q23-q24 and are associated with large hepatocellular carcinomas. *Hepatology* **38**, 1242–1249 (2003).
 14. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
 15. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* **99**, 260–266 (2002).

16. Jiao, S. *et al.* Estimating the heritability of colorectal cancer. *Hum. Mol. Genet.* **23**, 3898–3905 (2014).
17. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
18. Sherry, S. T., Ward, M. H., Kholodov, M. & Baker, J. dbSNP: the NCBI database of genetic variation. *Nucleic acids ...* (2001).
19. Taliun, D. *et al.* LASER server: ancestry tracing with genotypes or sequence reads. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx075
20. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat. Genet.* **46**, 409–415 (2014).
21. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
22. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
23. Jiao, S., Hsu, L., Hutter, C. M. & Peters, U. The use of imputed values in the meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **35**, 597–605 (2011).
24. Lin, D.-Y. & Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).
25. EPACTS. *genome.sph.umich.edu* Available at: <http://genome.sph.umich.edu/wiki/EPACTS>.

26. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
27. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
28. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
29. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829 (2014).
30. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
31. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
32. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
33. Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179–185 (2008).
34. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

35. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
36. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799–807.e24 (2013).
37. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
38. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
39. Moyerbrailean, G. A. *et al.* Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLoS Genet.* **12**, e1005875 (2016).
40. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
41. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
42. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinformatics* **14**, 144–161 (2013).
43. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45**, D626–D634 (2017).
44. Bhattacharya, G., Abraham, C. A., Wildrick, D. M. & Boman, B. M. Purification of the adenomatous polyposis coli (APC) gene product. *Arch.*

- Biochem. Biophys.* **323**, 233–236 (1995).
45. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–4 (2012).
 46. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
 47. GRCh37 - Assembly - NCBI. doi:10.1093/jnci/djx030
 48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 49. BamUtil: _bam2FastQ. *genome.sph.umich.edu* Available at: http://genome.sph.umich.edu/wiki/BamUtil:_bam2FastQ.
 50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
 51. BamUtil. *genome.sph.umich.edu* Available at: <http://genome.sph.umich.edu/wiki/BamUtil>.
 52. Li, B. *et al.* QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data. *BioMed Research International* **2013**, 1–4 (2013).
 53. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat. Genet.* **46**, 409–415 (2014).
 54. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics* **91**, 839–848 (2012).

55. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. *An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome research* **25**, 918–925 (2015).
56. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
57. Vt. *genome.sph.umich.edu* Available at:
<http://genome.sph.umich.edu/wiki/Vt>.
58. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).
59. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
60. Alakus, H. *et al.* Genome-wide mutational landscape of mucinous carcinomatosis peritonei of appendiceal origin. *Genome Med* **6**, 43 (2014).
61. Blandino, G. *et al.* Tumor suppressor microRNAs: a novel non-coding alliance against cancer. *FEBS Lett.* **588**, 2639–2652 (2014).
62. Brown, A. R. *et al.* Krüppel-like factor 9 (KLF9) prevents colorectal cancer through inhibition of interferon-related signaling. *Carcinogenesis* **36**, 946–

955 (2015).

63. Kang, L., Lü, B., Xu, J., Hu, H. & Lai, M. Downregulation of Krüppel-like factor 9 in human colorectal cancer. *Pathol. Int.* **58**, 334–338 (2008).
64. Wang, X., Wang, Q., Hu, W. & Evers, B. M. Regulation of phorbol ester-mediated TRAF1 induction in human colon cancer cells through a PKC/RAF/ERK/NF-kappaB-dependent pathway. *Oncogene* **23**, 1885–1895 (2004).
65. Li, L., Lou, Z. & Wang, L. The role of FKBP5 in cancer aetiology and chemoresistance. *Br. J. Cancer* **104**, 19–23 (2011).
66. Mukaide, H. *et al.* FKBP51 Expressed by Both Normal Epithelial Cells and Adenocarcinoma of Colon Suppresses Proliferation of Colorectal Adenocarcinoma. *Cancer Investigation* **26**, 385–390 (2009).
67. Han, S. *et al.* Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. *Leuk. Res.* **34**, 1271–1274 (2010).
68. Jia, W.-H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.* **45**, 191–196 (2013).
69. Whiffin, N., Dobbins, S. E. & Hosking, F. J. Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Human molecular ...* (2013).
70. Du, M. *et al.* Fine-Mapping of Common Genetic Variants Associated with Colorectal Tumor Risk Identified Potential Functional Variants. *PLOS ONE* **11**, e0157521 (2016).

71. Clark, B. S. *et al.* Loss of Llg1 in retinal neuroepithelia reveals links between apical domain size, Notch activity and neurogenesis. *Development* **139**, 1599–1610 (2012).
72. Yao, L., Tak, Y. G., Berman, B. P. & Farnham, P. J. Functional annotation of colon cancer risk SNPs. *Nat Commun* **5**, 5114 (2014).
73. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

CHAPTER IV

Improvements for the meta-analysis software

RAREMETAL

Introduction

Advances in array genotyping and next-generation sequencing technology have greatly reduced the cost of variant detection, and as a result, generated unprecedented amounts of genomic variants from population-scale studies. To search for causal variants, the genome-wide association approach that surveys the whole genome without pre-requisite knowledge has been the major trend¹⁻³.

Recently, meta-analysis in GWAS analysis has been successfully applied in many large-scale human genetics studies²⁻⁵. Meta-analysis is the statistical procedure for combining data from multiple studies⁶. In ideal situations, the performance of modern meta-analysis methods provides equivalent power to that

This work in Chapter 4 will be submitted as:
Chen S, ..., Abecasis G. "RAREMETAL2: a more efficient and flexible tool for meta-analysis".
Yang J, Chen S, Abecasis G. "Improved score statistics for meta-analysis in gene-level association studies".
Part of this work in Chapter 4 has been submitted as:
Zhan X*, Chen S*, Jiang Y, Liu M, ..., Vrieze S, Abecasis G, Liu D. "Meta-analysis of Sequence-based association studies in the presence of multi-allelic sites". (* equal contribution)

of joint-analyses that require sharing of individual level data (and which are often much more cumbersome to execute)⁷.

Discovery of these association signals has been enabled by rapid improvements in meta-analysis methods and tools^{5,7-10}. Among these tools, RAREMETAL¹⁰ is one of the most widely used ones, with over hundreds of downloads and several successful applications to large consortium studies. However, during the real practice, we found that in a few non-ideal situations, the underlying assumptions of the standard meta-analysis method may be violated, resulting in biased approximation. Below are three common scenarios with violation of meta-analysis assumptions in real datasets:

1. In standard meta-analysis, one assumption is that within-study phenotypic mean and variances are equal to those in joint studies^{7,11}. Yet this assumption is not true for some situations, for example, in meta-analysis that combines traditional case-control. Studies (with case-control ratio typically close to 1) with biobank and population-based studies (with case-control ratio typically much larger). Although weighing summary statistics from each study by their effective sample size may provide some adjustment and reduce the power loss, this weighing strategy will fail for gene-based tests, because the within-study score statistics are related to sample size.

2. Compared to the traditional linear mixed models (LMMs), general linear mixed model (GLMMs) provides more calibrated results in some situations^{12,13} and has been used more often in recent years. A violation of the meta-analysis assumption occurs when combining score statistics generated from those old LMM models and those from the newly developed GLMM models. A few studies have indicated the systematic bias in the meta-analysis method of summing within-study score statistics from these two models, but little research has been conducted to address this challenge¹⁴.

3. The standard method for generating summary statistics suffers substantial power loss in sites with more than one alternative allele (typically named as multi-allelic sites). Exome Aggregation Consortium (ExAC) shows that 8% of the variants in the human exome are multi-allelic¹⁵. However, most analysis method can only model effects of one reference and one alternative allele. As a result, these multi-allelic SNPs are usually analyzed as separated sites, with the reference allele and one alternative allele on each site. As the multi-allelic information is ignored, this approach may lead to substantial power loss in both single-variant meta-analysis and gene-based test, especially for datasets with some individuals having heterozygous genotypes of two different alternative alleles.

Here, we implemented several new methods into RAREMETAL to better address these challenges: First, we implemented a new method that, rather than simply

summing summary statistics across studies, also accounts for variation in phenotypic means across studies, also accounts for variation in phenotypic means across studies. Second, we implemented a new method that transforms LMM generated summary statistics to be numerically equivalent to those from GLMMs. This method can additionally help to analyze family-based datasets where GLMM parameters are difficult to estimate. Third, we re-designed RAREMETAL's data structure, making it accommodate multi-allelic sites correctly. At the same time, we implemented a new method that jointly analyzes the effects of multiple alleles rather than the traditional method that analyze multiple alleles separately.

Additionally, as the genetics studies datasets becomes larger and larger, there is a growing need of optimizations on the speed and disk space usage of the software. To match this need, we also improved RAREMETAL, and its companion package RAREMETALWORKER (which is used to generate summary statistics as RAREMETAL's input) in the software engineering level. With the improvements, the new version taking 7 times less disk space to store covariance matrix files, and uses 30% less time to run meta-analysis.

Methods

Score statistics for individual studies

Consider a meta-analysis with K studies and N samples in total. Each study has n_k samples genotyped in m_k variants. Let Y_k denote the $n_k \times 1$ phenotype vector;

G_k denote the $n_k \times m_k$ genotype matrix (centered and normalized); X_k denote the $n_k \times (q_k + 1)$ augmented covariate matrix with first column set to 1 and the others encoding q_k covariates.

In some practical uses, covariates may need to be regressed out before fitting the model. To do this, we first fit covariates against the phenotypes under a linear model:

$$Y_k = Q_k X_k + \varepsilon_k$$

Then we denote

$$\tilde{Y}_k = Y_k - Q_k X_k$$

When covariates need to be regressed out, Y_k is to be replaced by the residual, \tilde{Y}_k .

For a specific SNP i , we denote its score statistics in study k as $(U_{i,k}, V_{i,k})$. We denote the meta-analysis score statistics as $(U_{i,meta}, V_{i,meta})$. Using the standard meta-analysis method, these meta-analysis score statistics were calculated as:

$$U_{i,meta} = \sum_{k=1}^K U_{i,k}$$

$$V_{i,meta} = \sum_{k=1}^K V_{i,k}$$

Improved score statistics for unbalanced studies

Consider the situation with unbalanced studies. An exact meta-analysis score statistics can be derived as:

$$U_{i,meta} = \sum_{k=1}^K U_{i,k} - 2n_k \delta_k (f_i - f_{i,k})$$

$$V_{i,meta} = \sum_{k=1}^K \tilde{\sigma}^2 \left[\frac{V_{i,k}}{\hat{\sigma}_k^2} - 4n_k (f_i f_{i'} - f_{i,k} f_{i,k}') \right]$$

Here δ_k represents the deviation between phenotype mean in study k and the overall phenotype mean. $f_{i,k}$ represents minor allele frequency (MAF) of this site i in study k , and f_i represents its overall MAF. $\hat{\sigma}_k^2$ is the residual variance of phenotypes in study k , and the joint residual variance is estimated as:

$$\tilde{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^K [(n_k - 1) \hat{\sigma}_k^2 + n_k \delta_k^2]$$

Combine score statistics from LMMs and GLMMs

For association analysis in study k , we assume the underlying model is a generalized linear model:

$$\text{logit}(Pr(Y_k = 1)) = G_{i,k} \beta_{GLMM,k} + Q_k X_k + g_{GLMM,k} + \epsilon_{GLMM,k}$$

However, a standard linear mixed model is use for association analysis:

$$Y_k = G_{i,k} \beta_{LMM,k} + Q_k X_k + g_{LMM,k} + \epsilon_{LMM,k}$$

Where the coefficient can be estimated from the score statistics:

$$\hat{\beta}_{LMM,k} = V_{LMM,k}^{-1} U_{LMM,k}$$

Through Taylor expansion and approximation on variance components for rare variants, we can estimate $\hat{\beta}_{GLMM,k}$ simply from $\hat{\beta}_{LMM,k}$:

$$\hat{\beta}_{GLMM,k} = \frac{\hat{\beta}_{LMM,k}}{C_k}$$

Where

$$C_k = \int \frac{e^{\alpha+e_k}}{(1 + e^{\alpha+e_k})^2} dF(e_k)$$

$$e_k \sim N(0, \text{var}(\hat{g}_{LMM,k}) + \text{var}(\hat{\epsilon}_{LMM,k}))$$

With this estimated correction term C_k , the score statistics from the LMM can be transformed to those from GLMMs:

$$\hat{U}_{GLMM,k} = C_k U_{LMM,k}$$

$$\hat{V}_{GLMM,k} = C_k^2 V_{LMM,k}$$

Using a standard meta-analysis method, these transformed score statistics can be summed with score statistics from other studies analyzed using GLMMs.

Jointly modeling allele effects of multi-allelic sites

Traditionally, for a variant with L alternative alleles, the reference allele is coded as 0, while the alternative alleles are consequently coded as 1 to L . Here, instead, for an individual j in study k at site i , we encode the genotype as a L -length vector $G_{i,j,k} = (G_{i,j,k,1}, G_{i,j,k,2}, \dots, G_{i,j,k,L})$, where the l^{th} entry is the number of the l^{th} alternative alleles.

Naturally, a linear mixed model to analyze the association between site i and the phenotype is

$$Y_k = \sum_{l=1}^L G_{i,k,l} \beta_{i,k,l} + Q_k X_k + g_{LMM,k} + \epsilon_{LMM,k}$$

And score statistics can be derived as:

$$U_{i,k,l} = \sum_{j=1}^{n_k} G_{i,j,k,l} (Y_{j,k} - Q_k X_{j,k})$$

Having the phenotype variance denoted as σ_k^2 , we have the variance-covariance matrix for multi-allelic sites represented as:

$$V_{i,k \sim i,k} = \sigma_k^2 (G'_{i,k} G_{i,k} - G'_{i,k} X_k (X'_k X_k)^{-1} X'_k G_{i,k})$$

Testing of l^{th} allele can be controlled on the effect of the remaining $L-1$ alternative alleles, using a method similar to conditional meta-analysis. We can then derive the score statistics for testing the effect of allele l as:

$$U_{i,k,l|-l} = \sum_{j=1}^{n_k} G_{i,j,k,l} (Y_{j,k} - \hat{Y}_{j,k})$$

$$\hat{Y}_{j,k} = \sum_{l'=1, l' \neq l}^L G_{i,j,k,l'} \beta_{i,k,l'}$$

Similarly, $V_{i,k,l|-l}$ can be estimated as the covariance between the score statistics of the l^{th} allele and the remaining $L-1$ alternative alleles. To test the effect of the l^{th} allele, a standard meta-analysis can be performed by summarizing $U_{i,k,l|-l}$, $V_{i,k,l|-l}$.

In gene-based tests, $U_{i,l|-l}$ can be used together with other score statistics from bi-allelic sites. In addition, to calculate covariance between a multi-allelic site i and a bi-allelic site i' , we have:

$$\text{cov}(U_{i,k,l|-l}, U_{i',k}) = \text{cov}(U_{i,k}, U_{i',k}) - (V_{i,k,l|-l} V_{i,k,-l \sim i,k,-l}^{-1}) \text{cov}(U_{i,k}, U_{i',k})$$

Here $U_{i',k}$ is calculated under the traditional bi-allelic model.

Optimization on covariance matrix files

The original RAREMETALWORKER stored a covariance matrix as a gzipped text file, one variant per line, using the format of chromosome, position, variant in the same LD block, and variant covariance in the same LD block. In the field of variant names and variant covariance, values were comma separated. With such method, the disk usage is scaled as $O(n^2)$, becoming extremely huge for large datasets.

To reduce the disk space usage, we changed the storing method for covariance matrix files in the new version. First, for a specific variant, we search for its nearby variants by tracing back the hash table of loaded variants, instead of directly recording nearby variants as a field in the file. Second, for a covariance with an extremely small value comparing to covariance from other nearby variants, it would be approximated to zero. Third, we only store index and covariance of variants with non-zero covariance.

It is worth noting that with this approximation a covariance matrix may not be positive semi-definite, or even become a zero matrix. To make downstream analysis possible, we add an identity matrix to the covariance matrix so that the following analysis can be normally performed.

Simulation studies

To test the performance of our new method in unbalanced studies, we simulated 20,000 haplotypes of 5 KB length for 5 European populations using COSI¹⁶. Then we sampled genotypes of 339 variants in 100,000 individuals. We simulated dichotomous phenotypes according to the standard logistic regression model with half randomly selected causal variants in a randomly selected region of 100 variants. We set the intercept term subject to 1% disease prevalence and the log odds ratio as $\frac{C}{f_k(1-f_k)}$, where C is a given constant. Here we simulated phenotypes with the constant C as zero. For balanced studies, each study had 300 cases and 300 controls. For unbalanced studies, the number of cases for the five studies were (60, 180, 300, 420, 540), and the numbers of controls for the five studies were (540, 420, 300, 180, 60). Two covariates (C_1 , C_2) were simulated for each study. C_1 was a binary covariate subjected to a Bernoulli distribution of $P = 0.5$; C_2 was a continuous covariates subjected to $N(0, 1)$. Covariate coefficients were taken as 0.1 in the logistic model.

To test performance of our new method in combining LMMs and GLMMs, we simulated another dataset of 1000 variants based on the standard logistic regression model:

$$Y = \text{Bernoulli} \left(\frac{e^{\beta G}}{1 + e^{\beta G}} \right)$$

Each element of the genetic effect vector β followed a normal distribution:

$$\beta_i \sim \pi_0 I(0) + (1 - \pi_0) N(0, \tau_G^2)$$

Here we set τ_G^2 is as 0.033, and set half of the variants as causal variants ($\pi_0=0.5$).

To simulate multi-allelic sites, we randomly sampled from the 219,680 multi-allelic sites in ExAC¹⁵ project. With the MAFs from these sampled sites, we simulated 1000 samples, with genotypes sampled from a multinomial distribution. Similar to previous datasets, we simulated genetic effect of each allele from the distribution of $N(0, \tau_G^2)$.

Results

Optimized methods in unbalanced studies

We simulated a test dataset of both balanced and unbalanced studies, with 3,000 individuals and 339 variants. We tested the original and the new versions of RAREMETAL on this dataset. Joint analysis was also performed as the golden standard for evaluation.

In the meta-analysis of balanced studies, the original and the new version show similar performance as the joint analysis. As we expected, in this situation, with the correction term for between-study phenotypic variation of zero, the new method shrinks to the original method. In the meta-analysis of balanced and unbalanced studies, the new method shows similar performance as the joint analysis, while the original method suffers substantial power loss (Figure 4.1).

Combining statistics from LMMs and GLMMs

We generated summary statistics using LMM on 4 simulated studies, and meta-analyzed the summary statistics using the new version of RAREMETAL, the method proposed by Pirinen et al¹⁴, and the original version. Joint analysis was performed on these simulated studies using GLMM as a golden standard. The new version achieves a equivalent association analysis power as the joint analysis (Table 4.1). The original method shows substantial power loss, compared to the joint analysis. Association detection power of Pirinen et al's method is higher than the original method but lower than our new method. These results indicate a correction of summary statistics in such scenario is necessary.

Jointly modeling allele effects of multi-allelic sites

Sampled from the real data in ExAC Project, we simulated a multi-allelic dataset to test the performance of our new method and the traditional method that analyzes each allele separately. In different levels of genetic effects, our new method shows higher power than the standard method (Table 4.2). For example, at the genetic effect of 0.25, the power from the traditional method is 0.62, 11% lower than the power from our new method of 0.66.

Software engineering improvements

We performed single-variant meta-analysis for 69,362 markers across 11,998 unrelated individuals (European ancestry) in 20 studies using RAREMETAL. The new version takes 11.5 seconds while the original version takes 17.2 seconds.

To test performance of optimizing covariance matrix files, we randomly sampled the Haplotype Reference Consortium (HRC)¹⁷ Panel imputed variants from the GECCO¹⁸ “Omni chip” (detailed description of the dataset is in Chapter 3). For a 10 Mb randomly sampled region from chromosome 9 with 145,017 variants, we generated covariance matrix files using the original version and new the version of RAREMETALWORKER. The file size of the original and the new version are 4.9 GB and 0.73 GB, respectively. Using the two covariance matrix files and 100 randomly generated variant groups, we performed burden test¹⁹ respectively. The resulting p values from the two covariance matrix files were highly concordant (Figure 4.2).

Discussion

In this chapter, we have described a major update to our software RAREMETAL that brings in software engineering improvements and several useful new methods for rare variant analysis. Using simulated datasets, we show the new update in addition preserve the software’s ability to meta-analysis in unbalanced studies, multi-allelic sites and GLMMs.

Our new method for unbalanced studies, which incorporates cross-sample phenotypic variation into calculating score statistics, greatly rescued the power loss of the standard method. For now, the method only applies to unrelated samples. It remains as a future direction of adjusting this method to datasets with family structures.

Our new method of combining score statistics generated by LMMs and GLMMs has better power than the standard method that naively combines score statistics from these different models altogether. From our simulation, we show that this method provides a feasible way of analyzing datasets with dichotomous traits, especially for family-based data where GLMM model parameter is difficult to estimate. Alternatively, the dataset can be fit with LMMs first, and later in meta-analysis RAREMETAL can transform these summary statistics to those from GLMMs.

Our new method for multi-allelic sites jointly model effects of multiple alleles, rather than the traditional method that analyzes multi-allelic sites separately. From the simulated dataset, we observed substantial power loss for the traditional method, which may indicate an under-estimation of the phenotyp-genotype association of multi-allelic sites in previous studies. Considering the huge amount of multi-allelic variants in large and deeply sequenced datasets, it will be very promising to apply our new version of RAREMETAL to these real datasets to re-evaluate the effects of multi-allelic sites.

Additionally, we made software engineering improvements to RAREMETAL and its companion package RAREMETALWORKER, making meta-analysis and summary statistics storage more efficient. The reduced covariance matrix file size will enable us to do gene-based tests in a subset of non-coding regions. However, to perform gene-based tests genome-wide, the optimized covariance matrix file is still a huge disk space cost. Considering the increasing number of variants in sequencing studies and imputation reference panels, a further reduction of covariance matrix file size is necessary. Possible solutions include binary coding and customized compression, but more experiments need to be conducted to evaluate their compression rate and real time cost. Nevertheless, our current improvement in RAREMETALWORKER is a promising step towards this direction.

In conclusion, we updated our meta-analysis software RAREMETAL with software engineering improvements and several new methods. With these improvements, we believe RAREMETAL will be even more useful for meta-analysis in future genetics studies.

Figures and Tables

Figure 4.1 P values from the standard and the optimized method for unbalanced studies in simulated dataset

Meta-analysis power the standard and the optimized method is compared against the joint analysis, which is used as golden standard for evaluating association detection power.

- (A) In balanced studies, the three methods are equivalent. Data points of the standard and the optimized method perfectly overlap.
- (B) In unbalanced studies, the standard method suffers substantial power loss while the optimized method is unaffected (still showing similar power as the joint analysis).

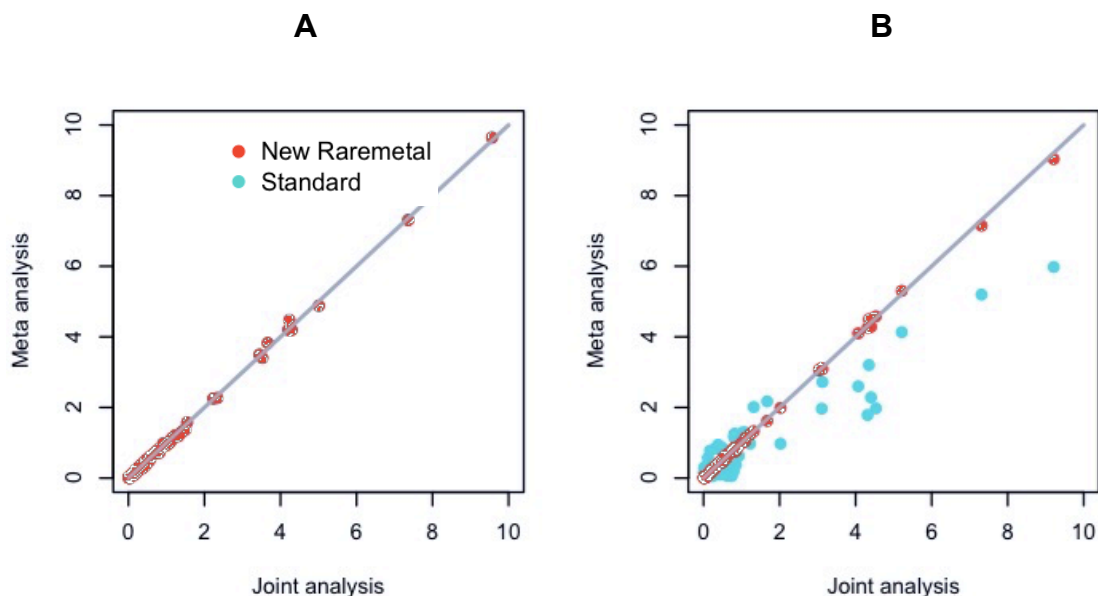


Figure 4.2 Burden test p values from the standard and the optimized covariance matrix

We sampled a 10 Mb genomic region and simulated genes for Burden test. The two methods showed almost the same P values for Burden tests, indicating the reduction on covariance matrix file size does not affect the performance of association test.

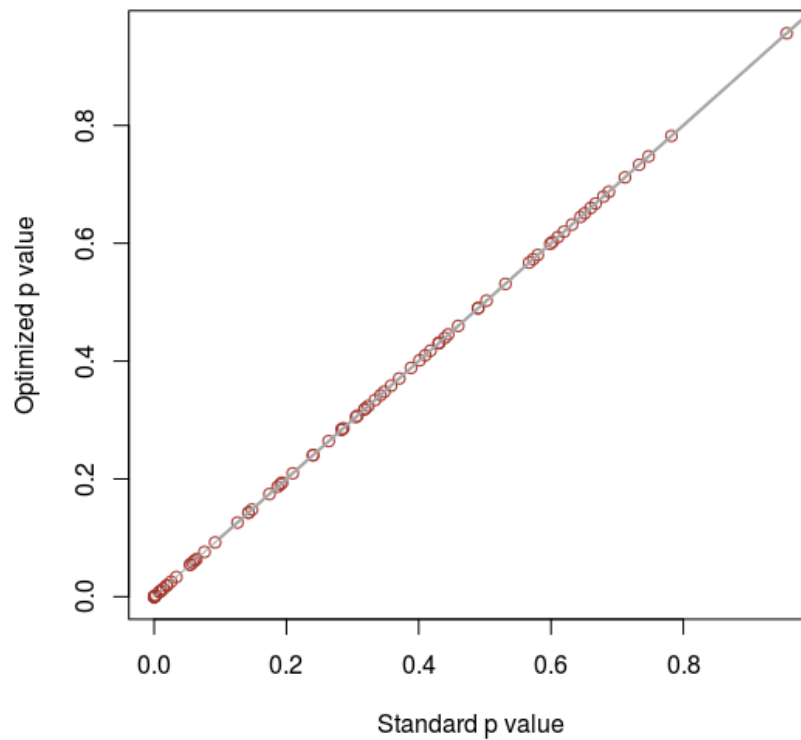


Table 4.0-1 Power of VT test for LMM analyzed studies

Four methods were compared. The original method naively combines summary statistics from LMMs together. Our new method and Pirinen et al's method scale LMM-derived summary statistics before combining. The joint analysis analyzes the pooled samples under GLMM. The power was evaluated under significance threshold $\alpha = 2.5 \times 10^{-6}$ in 1 million replicates. The optimal method has similar power as the joint analysis, while the standard method suffers substantial power loss.

Fraction of causal variants	Heritability	Power			
		Optimal	Pirinen et al	Standard	Joint analysis
0.5	0.3	0.35	0.32	0.28	0.35
	0.5	0.35	0.33	0.28	0.35
	0.7	0.36	0.31	0.28	0.36
0	0.5	2.4×10^{-6}	2.8×10^{-6}	2.6×10^{-6}	2.6×10^{-6}

Table 4.0-2 Power from the standard bi-allelic analysis and the new multi-allelic analysis method on multi-allelic sites

Genotypes were simulated based on MAF of multi-allelic sites in 7 sub-populations from the ExAC project from a multinomial distribution. Alternative allele effects were simulated from a normal distribution with the variance equal to the given genetic effect. The total sample size is 10,000. Association power was evaluated under significance threshold $\alpha = 5 \times 10^{-8}$ in 1 million replicates. The new multi-allelic method shows better power than the standard method under different levels of genetic effects.

Genetic effect	Association power	
	Standard bi-allelic method	New multi-allelic method
0	4.9×10^{-8}	4.8×10^{-8}
0.1	0.32	0.37
0.25	0.46	0.51
0.5	0.62	0.66

References

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
3. Barrett, J. C., Dunham, I. & Birney, E. Using human genetics to make new medicines. *Nat. Rev. Genet.* (2015).
4. Begum, F., Ghosh, D., Tseng, G. C. & Feingold, E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* **40**, 3777–3784 (2012).
5. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
6. Haidich, A. B. Meta-analysis in medical research. *Hippokratia* **14**, 29–37 (2010).
7. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
8. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
9. Lin, D.-Y. & Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.*

- 89**, 354–367 (2011).
10. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829 (2014).
 11. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101 (1954).
 12. Jiang, D., Zhong, S. & McPeck, M. S. Retrospective Binary-Trait Association Test Elucidates Genetic Architecture of Crohn Disease. *The American Journal of Human Genetics* **98**, 243–255 (2016).
 13. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for Heritable Covariates Can Bias Effect Estimates in Genome-Wide Association Studies. *The American Journal of Human Genetics* **96**, 329–339 (2015).
 14. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics* **7**, 369–390 (2013).
 15. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
 16. Liu, Y., Athanasiadis, G. & Weale, M. E. A survey of genetic simulation software for population and epidemiological studies. *Human Genomics* 2008 3:1 **3**, 79 (2008).
 17. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

18. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799–807.e24 (2013).
19. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

CHAPTER V

Summary and Discussion

Summary

Advances in genotyping technology have led to the discovery of an unprecedented amount of variants, while association analysis has become the major approach to finding disease-related genes and loci out of these millions or even billions of variants. The findings from these revolutionized technology and methodology have been evolving our understanding of human genetics disorders. In this dissertation, I have contributed to discovering disease-related variants and genes, both in terms of method development and in terms of real data analysis.

In chapter 2, I described a likelihood-based method, LIME, to detect MEIs (a specific type of novel insertion) from sequencing data. The method naturally accommodates cross-sample heterogeneity, and generates genotype likelihood to measure the probability of each MEI event. Tested by both simulated data and real data, LIME shows better performance than existing methods, especially in low-coverage data. In addition, by applying LIME to samples from the Sardinia

Whole Genome Sequencing Project, we generated a MEI dataset with functional impact prediction on each MEI. We believe LIME will help to generate better-quality MEI dataset from other sequencing studies as well.

In chapter 3, I presented a genome-wide association analysis on colorectal cancer in 26,903 individuals imputed onto the Haplotype Reference Consortium (HRC) reference panel. We identified 6 regions of potential associations with colorectal cancer. We replicated many previous findings on common and high penetrant variants related to colorectal cancer as well, showing the reliability of our method. By incorporating functional annotation, sequence function prediction and online eQTL data, we highlighted additional loci for potential causality and impact on nearby gene expression. Although it is difficult to obtain new significant variants in absence of extremely large dataset, our analysis provides new insights into association analysis under limited sample size.

In chapter 4, I developed several improvements to the meta-analysis software, Raremetal. We implemented a new meta-analysis method that incorporates cross-study phenotypic variation, and thus solves the problem of substantial power loss in unbalanced studies. We improved the method of meta-analyzing statistics from LMMs and GLMMs. This improvement also allows single studies to be analyzed with LMMs, since the score statistics can be later transformed as equivalent to those from GLMMs. We improved the coding scheme and analysis method of multi-allelic variants. With the new method, we got better power for

multi-allelic variants than the previous method. Finally, we optimized the storage of covariance matrix files from the software, making the file size much smaller than previous versions. With these improvements, the software Raremetal now more precisely captures the association between variants and diseases in complicated situations, and become more flexible and efficient on real datasets.

Future Directions

There still remain many questions in sequencing and association analysis, which could be statistically and computationally challenging. In the future, solutions or improvements in the following areas will lead to new genetic discoveries.

First, software LIME can be improved to incorporate the information from longer reads. In the future, read length of the short read approach may continue to increase, and it will become crucial to incorporate the information conveyed by those longer reads. One approach will be to incorporate the re-mapping score from MEI consensus sequence into likelihood calculation, replacing the current approach that only uses a binary standard. On the other hand, long read approaches, such as PacBio, and short read approaches may be combined to generate better calling of MEIs. To incorporate the information from both approaches, we need to have a better estimation of sequencing error rates from long read approaches, and then improve the likelihood calculation method so that information from both approaches will be incorporated.

Second, software LIME can be applied to more datasets to perform association tests for detected MEIs. Although a few studies have more or less implied the association between MEIs and certain diseases, there is very little research focused on MEIs' disease causality on a genome-wide scale. Since LIME generates genotype likelihood for each MEI, we can apply LIME to large datasets and perform LD-aware refinement the detected MEIs together with SNPs and Indels in order to improve MEI detection quality. With the combined variant dataset, we could perform association tests and obtain a comprehensive view of different variants' roles in disease causality. At the same time, special considerations are required for such analysis. Variant calling is always computationally intensive. In variant calling stage, we need to optimize the pipeline for both the short variant calling and MEI calling. In the phasing stage, we need a systematic review of the scale of likelihood between different types of variants; otherwise the phasing will be driven by those variants with extreme scale of genotype likelihood. In association analysis, since the larger variants may have higher impact on gene functions, we may need an optimized weighing scheme for short variants and MEIs.

Third, a replication analysis for the imputed GECCO data should be performed. Replication analysis is always the golden standard to test if an association signal is true. Moreover, with increased sample size, we may obtain significant signals for those associated rare variants, which previously were insignificant simply because of the limitation of sample sizes. Additionally, with more collaboration

between consortiums, we could incorporate samples from other studies as additional controls; as a result, this approach will increase the effective sample size as well as the association detection power.

Conclusion

New technology always brings up new challenges. With the increasing size of genetics data, appropriate analysis methods and efficient statistical tools will be in great need. In this dissertation, I have proposed improved methods for analyzing genetics data, and performed an analysis on a real dataset to discover associated variants for colorectal cancer. I believe these newly developed methods and approaches will facilitate analysis of genetics data, and provides insight to future genetics researches.