이학박사 학위논문

Analysis strategy and method to improve accuracy of imputation on rare variants

희귀변이의 임퓨테이션 정확도 향상을 위한 분석
전략 및 방법 연구

2015년 8월

서울대학교 대학원

협동과정 생물정보학

김 영 진

Abstract

Analysis strategy and method to improve accuracy of imputation on rare variants

Young Jin Kim
Interdisciplinary Program in Bioinformatics
College of Natural Science
Seoul National University

Rare variants have gathered much attention as an alternative source of missing heritability. Rapid development in high throughput sequencing technology has enabled us to discover a large number of rare variants. Although next-generation sequencing technology is becoming a powerful tool in genomics, it is not yet feasible to perform a large-scale population based genome study due to its high cost and required high computing power. Alternatively, two approaches, imputation and customized chips such as exome array and Metabochip, have been widely used in large scale genome studies. Imputation is a cost-effective approach that imputes rare variants into existing genotype data. Generally imputation analysis requires two panels as input: reference panel is the template for predicting untyped markers and genotype panel is the target for imputation analysis. After imputation analysis, the information of genotype panel contains previously experimentally genotyped information and predicted genotypes based on reference panel information.

However, imputing rare variants is very challenging due to low accuracy of imputed rare variants. Moreover, low accuracy of imputed rare variants would mislead the results of region-based association tests. Customized chips are designed to contain rare variants yet those chips are designed only for the specific targets. Therefore, new analysis strategy and method for obtaining rare variants are urgently in need.

In this study, we developed two novel rare variant imputation approaches, combined approach and pre-collapsing imputation approach. We also applied two approaches to real data analysis. Imputation based association study was performed on liver enzyme traits.

First, we proposed combined approach that imputes genotype panel consists of combined data of GWAS chip and exome array. The effectiveness and performance of combined approach were demonstrated using reference panel comprising exome sequencing, exome array, and GWAS chip of 848 identical samples and 5,349 samples of genotype panel consisting of exome array and GWAS chip. As a result, the combined approach increased about 11% in imputation accuracy and enhanced about two times of genomic coverage for rare variants (MAF < 1%) compared to imputation results of genotype panel with GWAS chip alone. Regardless of samples size of reference panel, combined approach showed better imputation performance. Also combined approach outperformed previously reported two-step imputation approach.

Second, we developed new method, pre-collapsing based imputation approach (PreCimp), to increase imputation accuracy in forms of collapsed variables. Unlike with previously introduced imputation approaches, PreCimp only requires computational cost. PreCimp consists of two steps. In the first step, collapsed variables are generated using rare variants in the reference panel and new reference panel is constructed by inserting pre-collapsed variables (PCVs) into the reference panel. Next, typical imputation analysis with the new reference provides the imputed genotypes of collapsed variables. We demonstrated the performance of PreCimp on 5,349 genotyped samples using a Korean population specific reference panel including 848 samples of exome sequencing, Affymetrix 5.0, and exome chip. PreCimp outperformed a traditional post-collapsing method that collapses imputed variants after single rare variant imputation analysis. Although PreCimp poorly performed for genes sized larger than 200kb (about 3% of all genes), PreCimp approach by split large-sized genes into small sub-regions could control the poor performance issues. PreCimp approach was shown to increase imputation accuracy about 3.4 ~ 6.3% (dosage r^2 0.6 ~ 0.8), 10.9 ~ 16.1% (dosage r^2 0.4 ~ 0.6), and 21.4 ~ 129.4% (dosage r^2 below 0.4) compared with the results of post-collapsing method.

Two imputation approaches were applied to real data analysis. We performed imputation based association analysis on liver enzymes. Using whole-exome reference panel, imputation analysis was performed on 8,529 samples of combined data consisting of GWAS chip and exome chip. Subsequent association analysis on about half million imputed and genotyped variants revealed 20 associated loci responsible for the variation of liver enzymes ($P < 5x10^{-6}$). Among them, 7 novel loci including two missense variants were discovered.

Taken together, two novel rare variant imputation approach were developed and applied to real data analysis. Imputation based association analysis on liver enzyme discovered several novel findings. This study proposed efficient analysis approaches for enhancing imputation accuracy of rare variants.

Additionally, in application to real data analysis, discovered variants will be valuable resource for understanding rare variants and its association to various phenotypes.

Keywords: SNP, imputation, rare variant, genome-wide association study, association

Student Number: 2009-30105

Contents

Chapter 3 Pre-collapsing Imputation approach...............38

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background and Motivation

1.1.1 Genome-wide association study

The ultimate goal of human genetics is to discover associated variants for common complex diseases (Hirschhorn and Daly 2005; Bush and Moore 2012). At the end of human genome project, various genetic variants were discovered including single nucleotide polymorphism (SNP). Most of SNPs are biallelic and most abundant type of variant in the human genome. In dbSNP b141 (21/05/2014), there are 62,387,983 reference SNP ids (rsID) available. One property of SNP is linkage disequilibrium (LD) that is non-random association of alleles at different loci (Bush and Moore 2012). Using LD structure, SNPs can be used as an indirect marker that is in high LD with causal SNP or as direct association marker that has functional effect on diseases (Bush and Moore 2012). These characteristics of SNPs enabled us to conduct association mapping for diseases and traits (Figure 1.1). There are two commonly used approaches for association mapping. The first approach is candidate gene study. This approach is hypothesis based study that genes are selected for association mapping based on other evidence of affecting disease risk (Hirschhorn and Daly 2005). Genome-Wide Association Study (GWAS) is the second approach. GWAS is hypothesis free approach that there is no assumption about genomic location or genes affecting disease risk (Hirschhorn and Daly 2005). GWAS scans disease associations across whole genome (Hirschhorn and Daly 2005).

In 2005, GWAS successfully identified various genetic loci associated with age-related macular degeneration (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005). During the last decade, GWAS has become efficient tool for human genomics for identifying genetic variants responsible for diseases and traits. As of November 2014, there are 2,051 publications and 14836 SNPs (GWAS catalogue: http://www.genome.gov/gwastudies/) (Hindorff et al. 2009). These increasing amount of information would lead us to an understanding of genetics underlying diseases.

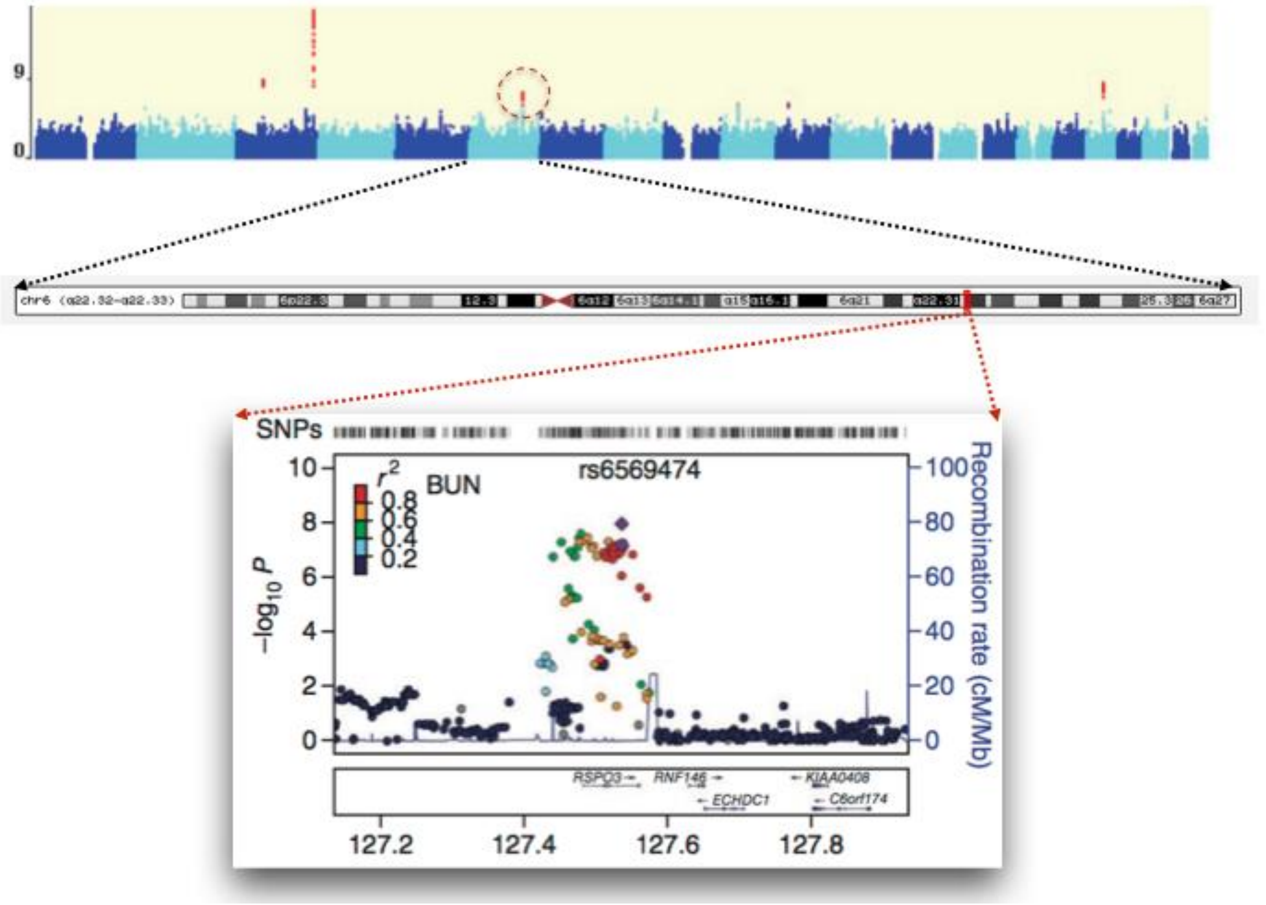


Figure 1.1 Example of genome-wide association mapping

1.1.2 Genotype imputation

Genotype imputation predicts untyped markers of genotyping chip using reference haplotypes with dense set of markers such as International HapMap project or 1,000 genomes project (Marchini and Howie 2010) (Figure 1.2). Imputation analysis has been widely used in GWAS to perform *in silico* fine mapping and genome-wide meta-analysis (Marchini and Howie 2010).

In GWAS, common SNP genotyping platform is SNP microarray. The chip, SNP microarray, contains more than hundreds of thousands of SNPs in a single chip. One of most widely used commercial chip contains approximately 1 million SNPs. SNP microarray contains only limited number of SNP markers that locate across human genome. Therefore, researchers would select associated regions of interests and perform fine mapping on the regions by resequencing or high density genotyping of SNPs in the region. These post-GWAS process requires high cost, time, and additional DNA samples of participants. By genotype imputation, researchers can perform in silico fine mapping with computational cost only. Example of in silico fine mapping is shown in Figure 1.3. High density imputed genotypes enhance association mapping power for the discovery of associated variants(Marchini and Howie 2010).

For further identification of disease associations and increasing statistical power, genome-wide meta-analysis has been widely used (Thompson et al. 2011). However, large discrepancy in contents of commercial arrays used for GWAS is the major problem in genome-wide meta-analysis. For example, 1M chips of Affymetrix and Illumina only shares about 30% of their contents (http://www.affymetrix.com, http://www.illumina.com). The main reason of the difference is from the difference in design of chips. Affymetrix SNP genotyping 6.0 (1M chip) contains markers that evenly spaced across genomes and tagging SNPs. Illumina 1M chip contains most of markers as tagging SNPs. This problem also can be solved via genotype imputation. Since imputation estimates every SNPs of the reference panel, all study genotypes after imputation have the same contents for meta-analysis. Example of imputation application in genome-wide meta-analysis is shown in Figure 1.4.

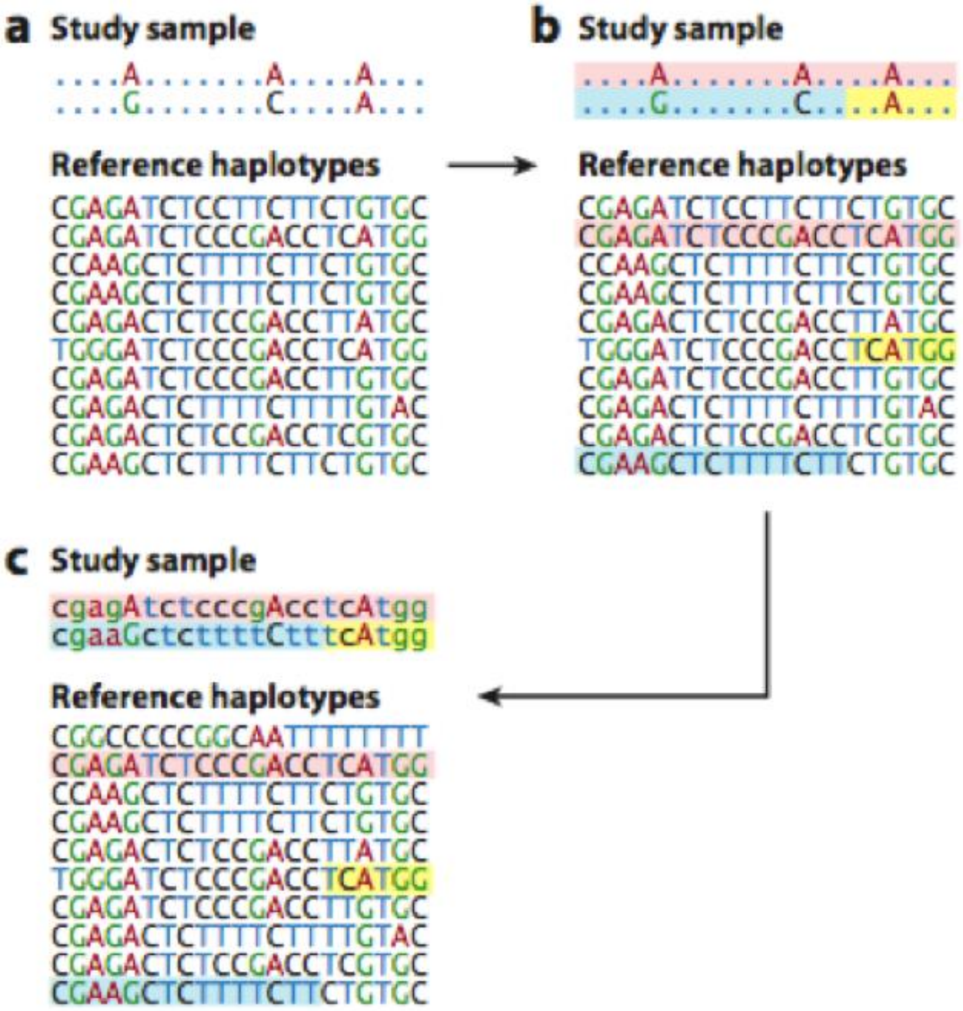


Figure 1.2 Schematic flow of imputation analysis (Li et al., Annu. Rev. Genomics Hum. Genet. 2009)

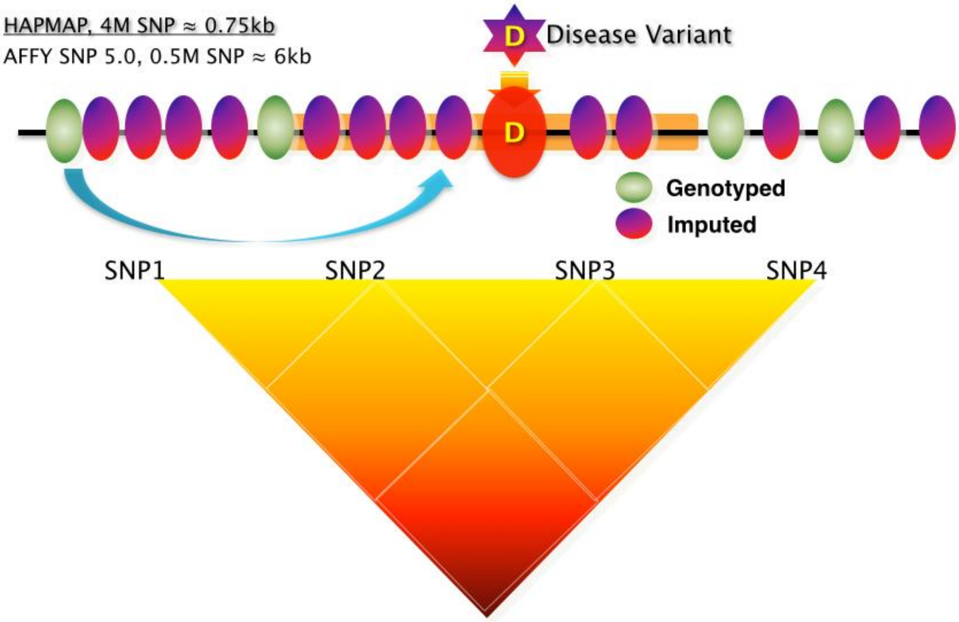
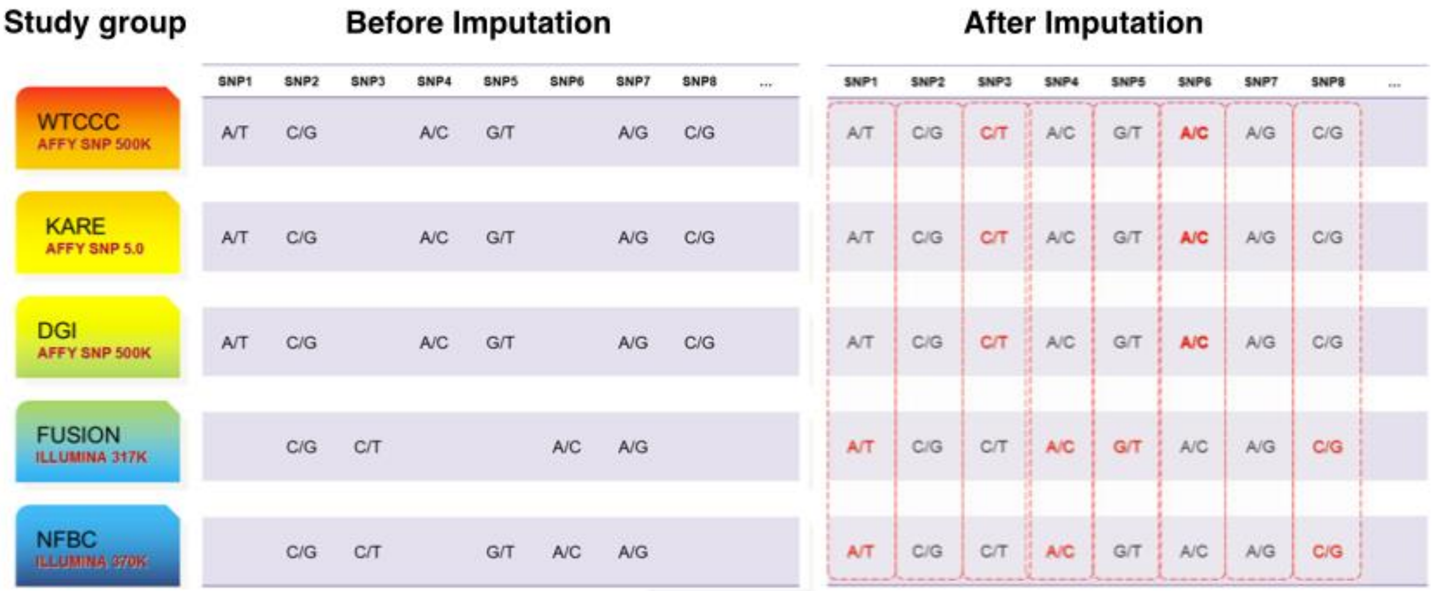


Figure 1.3 Example of *in silico* fine mapping



Study group	Before Imputation									After Imputation								
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	...	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	...
WTCCC AFFY SNP 500K	A/T	C/G		A/C	G/T		A/G	C/G		A/T	C/G	C/T	A/C	G/T	A/C	A/G	C/G	
KARE AFFY SNP 5.0	A/T	C/G		A/C	G/T		A/G	C/G		A/T	C/G	C/T	A/C	G/T	A/C	A/G	C/G	
DGI AFFY SNP 500K	A/T	C/G		A/C	G/T		A/G	C/G		A/T	C/G	C/T	A/C	G/T	A/C	A/G	C/G	
FUSION ILLUMINA 317K		C/G	C/T			A/C	A/G			A/T	C/G	C/T	A/C	G/T	A/C	A/G	C/G	
NFBC ILLUMINA 370K		C/G	C/T		G/T	A/C	A/G			A/T	C/G	C/T	A/C	G/T	A/C	A/G	C/G	

Figure 1.4 Example of imputation application in genome-wide meta-analysis

1.1.3 Missing heritability

Despite the success of GWAS, discovered variants from GWAS have only explained small proportion of phenotypic variance (Manolio et al. 2009). For example, previous height GWAS on 180,000 samples discovered 180 loci and those loci only explain about 12% of heritability (Lango Allen et al. 2010; Lander 2011). Since estimated heritability from siblings was about 80% (Visscher et al. 2006), several questions on “missing heritability” after GWASs on tens of thousands of samples only explain small phenotypic variance. One of questions is the source of missing heritability. GWAS has primarily focused on common variants (minor allele frequency; MAF > 5%). Therefore, alternative source of missing heritability would be as follows (Manolio et al. 2009; Zuk et al. 2014): 1) much large number of common variants with small effect, 2) rare variants (MAF < 1%), 3) structural variants, 4) Gene-Gene interaction, and 5) inadequate accounting for shared environment among relatives.

1.1.4 Rare variant imputation

By the advent of Next Generation Sequencing (NGS) technology, rare variants have increasing attention among alternative source of missing heritability (Bansal et al. 2010; Zuk et al. 2014). However, NGS requires high cost and compute intensive process. Therefore, NGS is not yet applicable to a large scale population based genomic study such as GWAS.

Alternatively, imputation analysis has been used for studying less common or rare variants (Auer et al. 2012). Imputation analysis is efficient way to obtain rare variants since it only requires computational cost. However, imputation has major limitation in imputing rare variants. Li et al. reported that extremely rare variants are unlikely to impute even with thousands of reference samples (Li et al. 2011).

To enhance imputation accuracy of rare variants, previous studies have reported various approaches (Joshi et al. 2013; Kreiner-Moller et al. 2014; Li et al. 2011; Duan et al. 2013; Deelen et al. 2014). Previous strategies can be categorized into four types: 1) construct the reference panel with sequenced samples (Duan et al. 2013; Deelen et al. 2014), 2) increase samples size of reference panel (Li et al. 2011), 3) use complementary information retrieved from local sequencing (Joshi et al. 2013), and 4) local ultra-high-density genotyping arrays (Kreiner-Moller et al. 2014). However, previous studies have mainly focused on utilization of reference panel. Therefore, different aspects of imputation strategy and methodological approach are warranted to more efficiently improve imputation accuracy of rare variants.

1.2 Objective of the research

Previous studies on improving imputation accuracy of rare variants suggested strategies based on construction or complementing information of reference panels (Joshi et al. ; Kreiner-Moller et al. ; Li et al. ; Duan et al. 2013; Deelen et al.). Since sequencing thousands of samples for constructing reference panel is not feasible and genotyping or sequencing a subset of samples require additional round of experiments, different aspect of rare imputation strategy and methodological approach is urgently in need. In this context, the primary purpose of this study is to develop rare variant imputation methods. First, combined approach was proposed. Combined approach uses combined data comprising GWAS chip and exome array for constructing genotype panel and following imputation analysis enhanced imputation accuracy and genomic coverage of rare variants. Second, a novel rare variant imputation method, pre-collapsing imputation approach, was proposed. Pre-collapsing imputation approach was developed to increase imputation accuracy of rare variants in terms of collapsed variables. In addition, we applied two approaches to real data analysis. Imputation based association analysis was performed on liver enzyme traits.

The dissertation is organized as follows: Chapter 1 introduces the background of this study. Chapter 2 contains the study of analysis strategy of combined approach to enhance imputation accuracy of rare variants. In Chapter 3, pre-collapsing imputation approach was developed to increase imputation accuracy of rare variants in terms of collapsed variables. In following Chapter 4, developed approaches of previous chapters were used in imputation based association analysis on liver enzyme traits. Finally, Chapter 5 summarizes the paper and conclusion.

Chapter 2. Imputation approach using combined data

2.1 Introduction

Genome-wide association studies (GWAS) have revealed unprecedented amount of disease associated loci (Zuk et al. 2014; Hindorff et al. 2009). However, previously reported loci only explained small proportion of heritability (Zuk et al. 2014; Gorlov et al. 2008; Bansal et al. 2010). Since previous GWAS mainly focused on common variants (minor allele frequency (MAF) > 5%), rare variants have gathered an increasing attention as an alternative source of missing heritability (Zuk et al. 2014; Gorlov et al. 2008; Bansal et al. 2010). By the advent of recent advancement in high-throughput sequencing technology, genome-wide assessment of rare variants has become possible (Zuk et al. 2014). For a large scale population based genome studies, however, sequencing technology is not yet feasible because of high cost and its computing intensive analysis process (Magi et al. 2012; Auer et al. 2012). Alternatively, two cost effective approaches have been widely used for studying rare variants. One approach is the genotype imputation analysis that estimates untyped rare markers using thousands of sequenced samples as a reference panel such as 1,000 genomes project data (Howie et al. 2012; Marchini and Howie 2010). The second approach is using genotyping chips such as Metabochip and exome array that are customized to contain rare variants (Huyghe et al. 2013 ; Lango Allen et al. 2010). These chips can genotype at less cost than commercial genome-wide single nucleotide polymorphism (SNP) arrays, and contain about quarter millions of variants optimized for specific targets. For example, Metabochip includes SNPs for replication and fine mapping aiming to study metabolic, cardiovascular, and anthropometric traits (Lango Allen et al. 2010). Exome array contains mainly functional coding variants selected from ~ 12,000 sequenced samples (Huyghe et al. 2013).

Indeed, the two approaches have been cost effective methods to access rare variants. Recent imputation based association studies have discovered numerous less common or rare variants associated with coronary artery disease, blood cell traits, serum creatinine, chronic kidney disease, and adult body height (Du et al. 2014; Sveinbjornsson et al. 2014; Auer et al. 2012). Customized chips designed to contain rare variants have successfully identified novel associations for hematological traits, blood lipid traits, coronary heart disease, and glycemic traits (Auer et al. 2012; Holmen et al. 2014; Peloso et al. 2014; Scott et al. 2012).

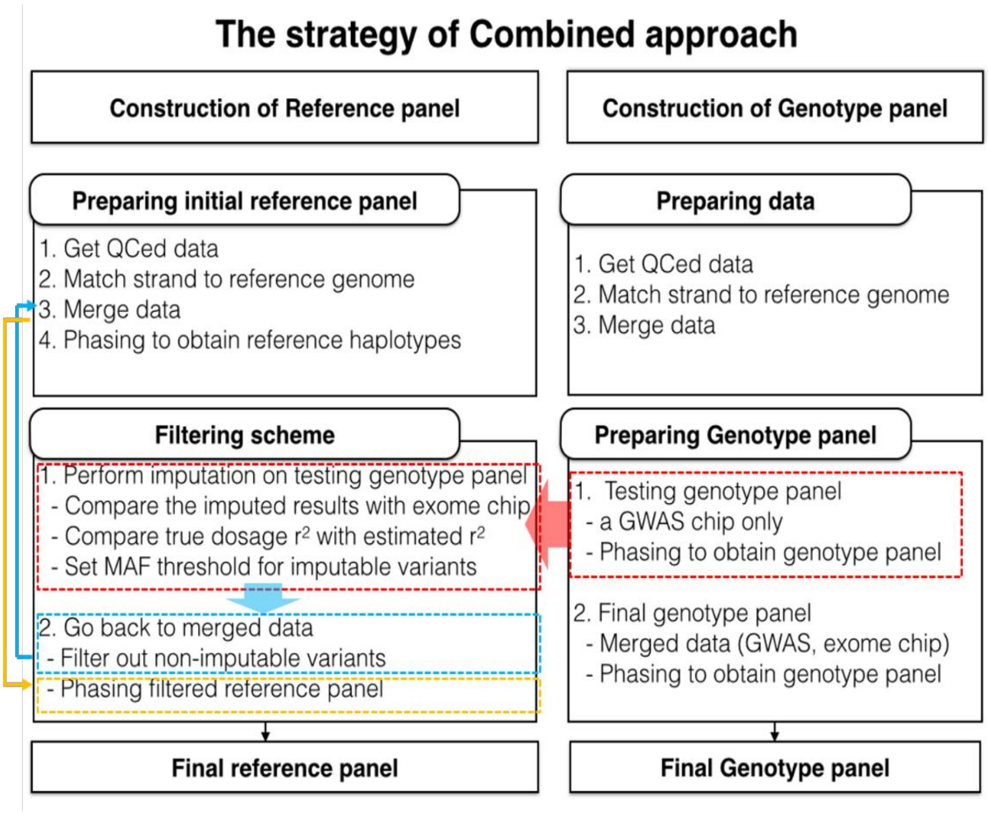
Despite noticeable successes, the two approaches have limitations. Imputing rare variants has been challenging due to low accuracy of imputed genotypes of rare variants (Li et al. 2011; Auer et al. 2012). Poorly imputed rare variants would result in misleading results in the following association study. Generally, imputation estimates untyped markers using haplotype patterns of common markers between reference panel and genotype panel (Howie et al. 2012; Marchini and Howie 2010). Therefore, the main reason for poor performance would be due to low correlation between rare variants and common tagging markers genotyped by GWAS chips. Accuracy of imputed rare variants would be improved if a chip used for genotype panel is designed to contain rare variants or markers tagging nearby rare variants (Joshi et al. 2013; Li et al. 2011). Customized chips are limited in that they are designed for specific purposes. Those chips do not contain markers for genome-wide scan. However, it can be a source of rare variants as a part of genotype panel for imputation analysis. In this context, the combined approach taking advantages of two approaches would be more powerful to obtain the genotypes of rare variants. If custom arrays can be genotyped on identical samples that were previously genotyped with genome-wide scan arrays, the combined approach would enhance imputation performance and association mapping power. Although general analysis strategy of imputation and custom arrays have been introduced (Howie et al. 2012; Marchini and Howie 2010; Lango Allen et al. 2010; Duan et al. 2013), analysis strategy and its effectiveness of combined approach have not been reported.

In this study, we describe the analysis strategy and its effectiveness of combined approach that performs imputation analysis on merged data including exome array and existing GWAS chip data. To demonstrate the effectiveness of our established strategy, we built a reference panel from 848 samples who have exome sequencing data, GWAS chip data, and exome chip data and then performed imputation analysis on genotype panels with 5,349 identical samples of an exome chip, a GWAS chip only, and merged data comprising a GWAS chip and an exome chip. Additionally, we studied sample size effect of reference panel on imputation performance of GWAS chip only and combined data. Also previously suggest two-step approach was compared with imputation results of GWAS chip only and combined data. To compare performance of results, we accessed imputation quality score and genomic coverage.

2.2 Materials and Methods

2.2.1 Overview of combined approach

The overview of strategy of combined approach is described in Figure 2.1. Dataset used for constructing panels is summarized in Table 2.1. For reference panel, we built an initial reference panel and a final reference panel without non-imputable extremely rare variants that possibly mislead in interpreting imputation quality score, estimated r^2. For a genotype panel, testing genotype panel consists of a GWAS chip only of 5,349 samples and final genotype panel comprised of merged data of an exome chip and a GWAS chip of 5,349 identical samples. To set a threshold for excluding non-imputable variants, initial reference panel was constructed by merging exome sequencing, exome chip, and GWAS chip data of 848 identical samples. Testing genotype panel was imputed using initial reference panel and imputation results were compared with true genotypes. Lower bound MAF showing concordance between dosage r^2 and estimated r^2 was used as a MAF threshold for excluding non-imputable variants. Finally, the final reference panel removing non-imputable variants were used to impute final genotype panel.



The strategy of Combined approach

Construction of Reference panel

Preparing initial reference panel

1. Get QCed data
2. Match strand to reference genome
3. Merge data
4. Phasing to obtain reference haplotypes

Filtering scheme

1. Perform imputation on testing genotype panel
 - Compare the imputed results with exome chip
 - Compare true dosage r^2 with estimated r^2
 - Set MAF threshold for imputable variants

2. Go back to merged data
 - Filter out non-imputable variants
 - Phasing filtered reference panel

Final reference panel

Construction of Genotype panel

Preparing data

1. Get QCed data
2. Match strand to reference genome
3. Merge data

Preparing Genotype panel

1. Testing genotype panel
 - a GWAS chip only
 - Phasing to obtain genotype panel

2. Final genotype panel
 - Merged data (GWAS, exome chip)
 - Phasing to obtain genotype panel

Final Genotype panel

Figure 2.1 Overview of combined approach

Table 2.1 Datasets used in this study

Category (# of samples)	Type	Exome sequencing	GWAS chip	Exome chip
# of variants		500,821	344,366	66,196
Reference panel (848)	Initial reference panel	O	O	O
	Final reference panel	O	O	O
Genotype panel (5,349)	Exome chip genotype panel	X	X	O
	Testing genotype panel	X	O	X
	Final genotype panel	X	O	O
True data (5,349)	For imputed variants using testing genotype panel	X	X	O

2.2.2 Exome sequencing

By the Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Ethnic Samples (T2D-GENES) Consortium at the Broad Sequencing Center, about 10,000 exomes from five ethnic groups have been sequenced using Agilent Human Exon v2 capture (~18,000 genes). Among them, part of samples from Korea Association REsource (KARE) project (Cho et al. 2009), including 538 samples from type 2 diabetes cases and 579 samples from controls, were included and 1,087 samples were used for further analysis after quality control on samples. The reference genome hg19 was used for alignment and variant calling process that was performed using the Genome Analysis Toolkit v2 (McKenna et al. 2010). As a result, 500,821 autosomal variants from 848 Korean samples were used for constructing reference panels. Accuracy of called variants was calculated by comparing genotypes from sequencing data with genotypes of genotyping chip data. Overall concordance was 99.76% and 99.96% for Affymetrix 5.0 and exome array, respectively.

2.2.3 GWAS and exome chip genotyping

Previously, 8,842 samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix Inc., San Diego, CA, USA) (Cho et al. 2009). Among them, 6,197 identical samples were genotyped using the Illumina HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA, USA) exome array. For the two platform, standard quality control on samples were conducted excluding samples with a high missing rate (>4%), gender discrepancy, excessive heterozygosity, or cryptic first degree relatives. Exclusion criteria for SNPs of Affymetrix GWAS chip was as follows: Hardy-Weinberg equilibrium p-values < 10^{-}

^{6}, genotype call rates < 95%, and MAF < 0.01. All chromosomal position of SNPs were updated to hg19 using the Affymetrix annotation file. Quality control on variants of exome array were similar to those of GWAS chip except threshold for filtering out variants with low allele frequency. Only monomorphic markers were excluded for further analysis. From quality controlled data, we used 6,197 samples that were common between sets of samples of Affymetrix GWAS chip and exome array. Variants included in the analysis were 344,366 and 66,196 for GWAS chip and exome array, respectively. Among 6,197 samples, 848 samples were used for constructing reference panel and remaining 5,349 samples were used for genotype panels.

2.2.4 Building reference panel

We constructed the reference panel by merging exome sequencing, exome array, and GWAS chip of 848 identical samples. The description of each data is summarized in Table 2.1. Prior to merging process, overlapped variants between sequencing data and chip data were removed from chip datasets. For overlapped variants between GWAS chip exome array, variants from exome array were used and overlapped variants were removed from GWAS chip. Number of overlapped and unique variants are shown as a Venn diagram in Figure 2.2. After merging all data, initial reference panel contained 856,690 variants and phased using the ShapeIT v2 program (Delaneau et al. 2012). Initial reference panel was used to impute testing genotype panel for selecting MAF threshold to exclude non-imputable variants. After extremely rare non-imputable variants (MAF < 0.3%) were excluded, the final reference panel contained 487,381 variants and phased using the ShapeIT v2 program.

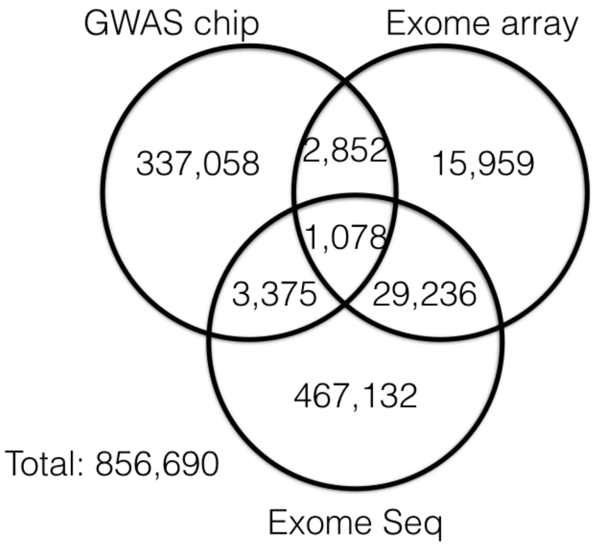


Figure 2.2 A Venn diagram of variants of three platforms used in the reference panel

2.2.5 Building genotype panel

Among 6,197 samples, 5,349 samples were remained after excluding 848 samples used for constructing reference panel. Genotype panel consists of exome array of 5,349 samples were phased using the ShapeIT v2 progam. As the testing genotype panel, GWAS chip data of 5,349 samples were phased using the ShapeIT v2 program. For the final genotype panel, GWAS chip and exome array of 5,349 identical samples were merged and phased using the ShapeIT v2 program.

2.2.6 Two-step imputation approach

Recently, Kreiner-Moller et al. reported a two-step imputation approach for improving imputation accuracy of rare variants (Kreiner-Moller et al. 2014). Two-step imputation approach uses local reference panel constructed using ultra high density SNP array with many low frequency markers. This approach is implemented as follows: 1) Construct local reference panel by genotyping only a subset of samples using an array with many low frequency markers 2) Impute study genotype panel using local reference panel 3) Impute the study genotype panel imputed in 2) by using 1,000 genomes project reference panel. To compare two-step imputation approach with combined approach, we modified the strategy of two-step imputation approach to our dataset as follows: 1) construct local reference panel using only a subset of samples of combined GWAS chip and exome chip data 2) impute study genotype panel (GWAS chip only) using local reference panel 3) imputed genotype panel is then imputed using exome reference panel.

2.2.7 Statistical analysis

In this study, we performed typical pre-phasing based imputation analysis on genotype panels (Howie et al. 2012). For imputation analysis, we used minimac software, a low memory and computationally efficient implementation of the MaCH algorithm (Li et al. 2011). To select MAF threshold for imputable variants, we compared dosage r^2 and estimated r^2 of imputed variants by observing scatter plots of each MAF bins. Dosage r^2 was accessed by calculating squared Pearson correlation (dosage r^2) between imputed dosages and true genotypes from exome array. For comparison analysis of imputation performance of genotype panels, we used estimated r^2 provided by minimac as an imputation quality measure. To test the difference between estimated r^2 values of imputation results of genotype panels, the Wilcoxon signed-rank test was performed. Statistical analyses and visualization of the results were performed using the R program.

2.3 Results

2.3.1 Selecting MAF threshold for non-imputable variants

In this study, we excluded non-imputable variants to construct the final reference panel. Previously, Li et al. reported that estimated r^2 would not be a good estimator for extremely rare variants (MAF≤0.5%) (Li et al. 2011). Here, we defined non-imputable variants as ones having a large difference between estimated r^2 and dosage r^2. If estimated r^2 would not reflect true accuracy, one cannot filter out low quality imputed variants based on estimated r^2 and comparison analysis of imputation performance using estimated r^2 would be misleading. In this context, we compared estimated r^2 with dosage r^2 for several MAF bins. We first performed imputation analysis on testing genotype panel, containing GWAS chip only, by using unfiltered initial reference panel. Among imputed variants, 45,802 imputed variants from 5,349 samples were compared to the corresponding variants obtained from an exome array constructed using identical samples. Figure 2.3 shows the imputation results of variants by MAF bins. As Li et al. reported, the estimated r^2 did not reflect their true value, dosage r^2, for extremely rare variants (MAF < 0.3%, Figure 2.3D). However, the estimated r^2 worked well for variants with MAF ≥ 0.3% (Figure 2.3A-C). Therefore, variants with MAF ≥ 0.3% were regarded as imputable in the current study.

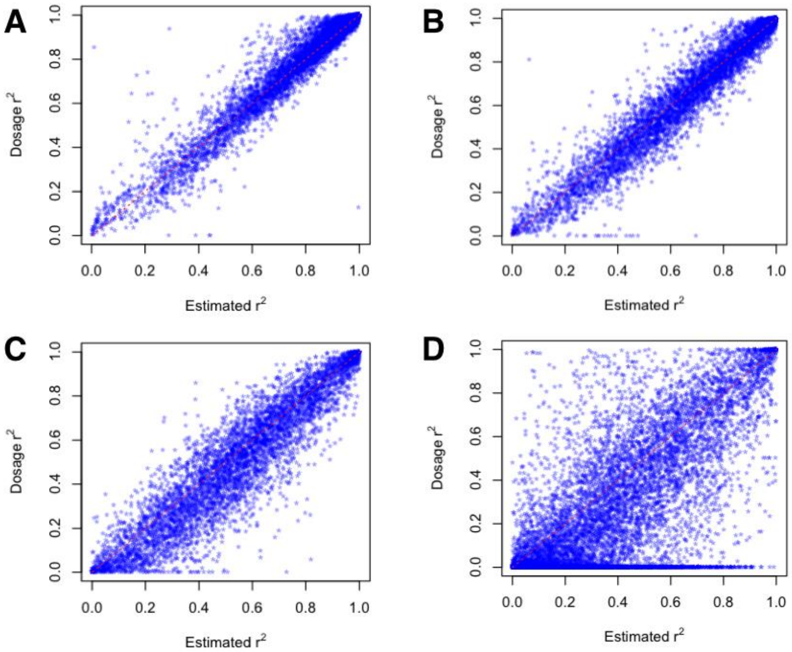


Figure 2.3 Scatter plot of estimated r^2 against dosage r^2 by MAF bins Estimated r^2 was plotted against dosage r^2 by MAF bins (A) MAF ≥ 5%, (B) MAF = 1 - 5%, (C) MAF = 0.3 - 1%, and (D) MAF < 0.3%. The red dotted line represents the diagonal.

2.3.2 Comparison of imputation accuracy among genotype panels

Using the final reference panel without non-imputable variants, we performed imputation analysis on genotype panels including exome array only, GWAS chip only, and merged data of two platforms. For comparison analysis, we used imputed 108,682 variants in overlap among three genotype panels and estimated r^2 was a measure for imputation accuracy. Number of variants were 35,443 (32.6%), 21,191 (19.5%), 19,527 (18.0%), and 32,547 (29.9%) for variant with MAF ≥ 5%, 1-5%, 0.5-1%, and < 0.5%, respectively. Figure 2.4 shows the comparison results. As previously reported, the genotype panel of exome array alone showed the worst performance (Martin et al. 2014). The mean estimated r^2 was 0.332, 0.616, and 0.661 for genotype panels of exome array, GWAS chip, and combined approach, respectively. Combined genotype panel showed the best performance compared to other genotype panels ($P < 2.2 \times 10^{-16}$, about 7.3% increase in mean estimated r^2 compared to those of GWAS chip only). In Figure 2.5, most of imputed variants using combined approach showed better performance than that using genotype panel of GWAS chip alone. The increment in imputation accuracy was the largest when allele frequencies of imputed variants were below than 1%. The increment in estimated r^2 of combined genotype panel was about 10-11% for rare variants (MAF 0.5 – 1%) and extremely rare variants (MAF 0.3 – 0.5%) compared to the genotype panel with GWAS chip only. Mean estimated r^2 of GWAS and combined approach was 0.870 and 0.906 for MAF ≥ 5%, 0.653 and 0.706 for MAF 1 – 5%, 0.465 and 0.515 for MAF 0.5% - 1%, and 0.406 and 0.452 for MAF 0.3 – 0.5%.

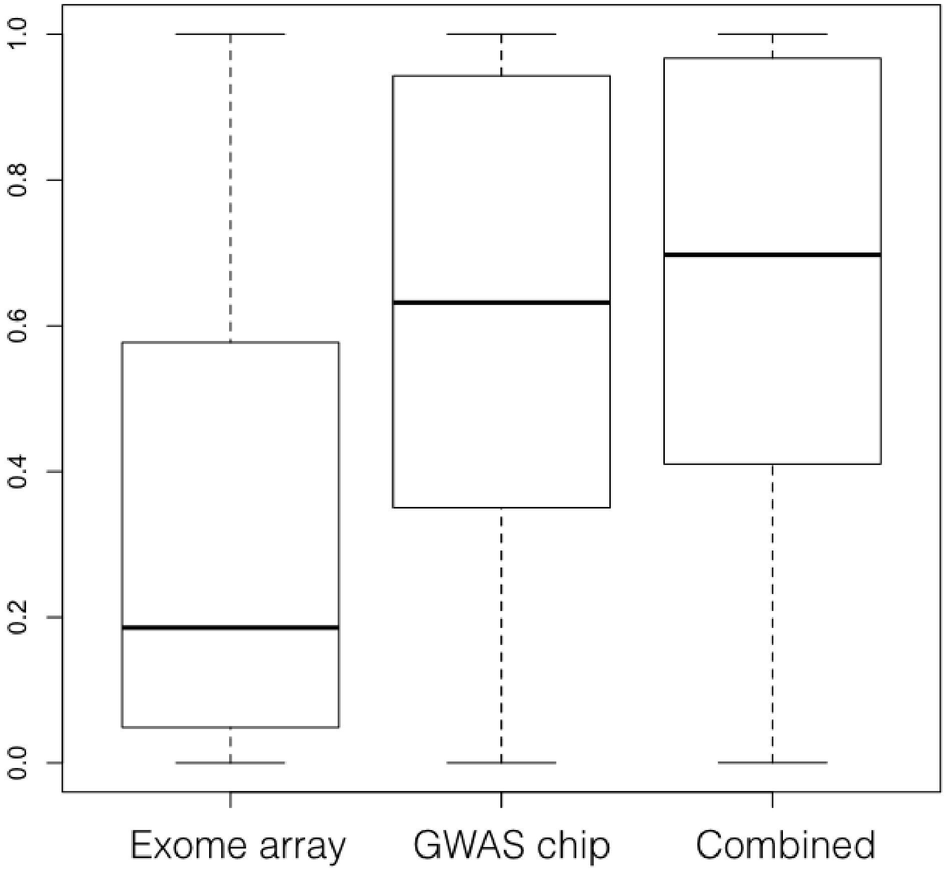


Figure 2.4 Boxplot of estimated r^2 of genotype panels

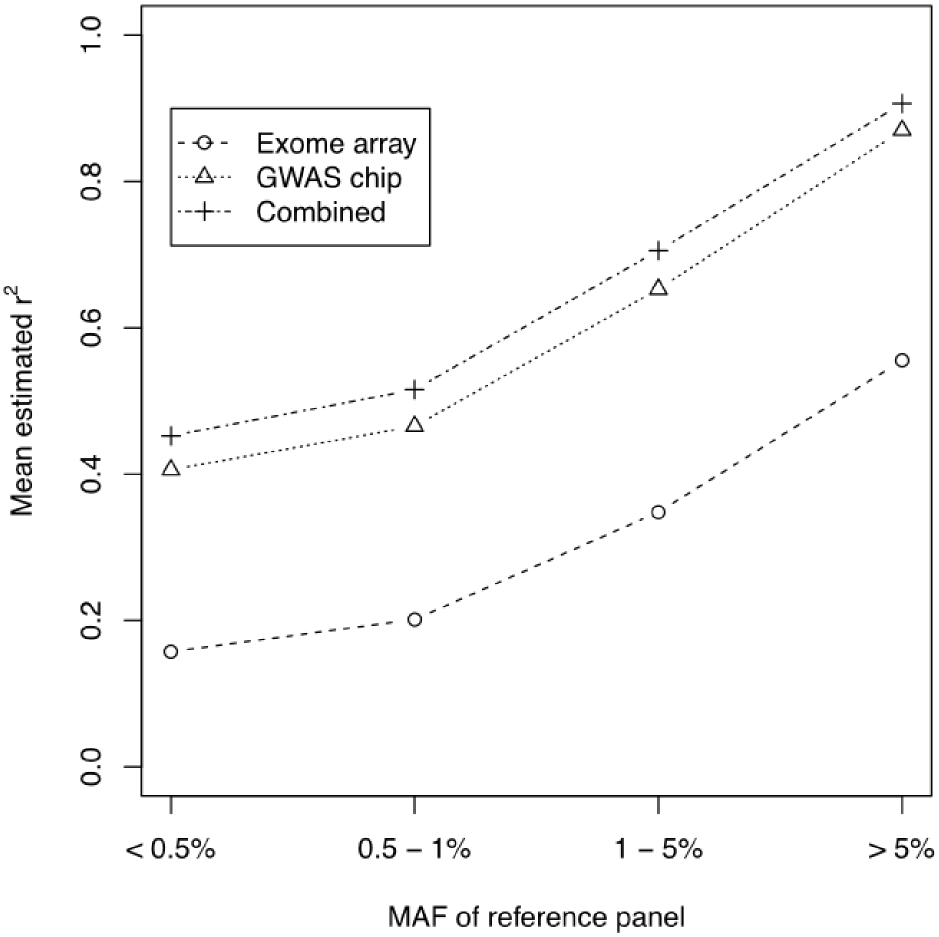


Figure 2.5 Mean estimated r^2 of genotype panels by MAF bins

2.3.3 Comparison of genomic coverage among genotype panels

Major advantage of imputation analysis is in obtaining dense set of imputed variants with relatively small number of markers of genotype panel. By using dense set of imputed markers, association mapping power can be increased via enhanced genomic coverage. This property has enabled us to perform *in silico* fine mapping in imputation based association studies. Recently, Nelson et al. reported imputation based genomic coverage of widely used genotyping arrays (Nelson et al. 2013). Imputation based genomic coverage is calculated as the number of imputed variants above imputation quality score threshold divided by total number of variants in the reference panel. In this study, we compared imputation based genomic coverage of genotype panels of GWAS chip only and combined approach. For genomic coverage, we selected 143,022 exonic variants including imputed and genotyped by exome array. Since we used exome sequencing data in constructing reference panel, 143,022 variants were regarded as virtual exome in this study. Number of variants were 56,326 (39.4%), 28,072 (19.6%), 22,931 (16.0%), and 35,693 (25.0%) for variant with MAF ≥ 5%, 1-5%, 0.5-1%, and < 0.5%, respectively. Table 2.2 summarized the results. We selected stringent cut-off as estimated r^2 of 0.8 and less stringent cut-off as estimated r^2 of 0.4. By using stringent cut-off, overall genomic coverage was 0.435 and 0.560 for GWAS chip only and combined approach. In overall, approximately 29% increase in genomic coverage was observed if combined approach used (r^2 threshold ≥ 0.8). For rare variants (MAF < 1%), however, genomic coverage of combined approach was about two times of those of genotype panel with GWAS chip only (r^2 threshold ≥ 0.8).

Table 2.2 Genomic coverage of genotype panels of GWAS chip only and combined approach

MAF bin	$r^2 \geq 0.8$		$r^2 \geq 0.4$	
	GWAS chip	**Combined**	**GWAS chip**	**Combined**
ALL	0.435	0.560	0.749	0.818
≥ 5%	0.794	0.901	0.953	0.983
1 – 5%	0.403	0.588	0.799	0.881
0.5 – 1%	0.146	0.290	0.585	0.686
0.3 – 0.5%	0.079	0.172	0.491	0.591

2.3.4 Sample size effect of reference panel and comparison analysis

Previously, numerous efforts have been reported to enhance imputation performance of rare variants (Joshi et al. 2013; Kreiner-Moller et al. 2014; Li et al. 2011; Duan et al. 2013; Deelen et al. 2014). Basically, there were three types of approaches. The first approach is to increase number of samples of the reference panel up to thousands of samples (Li et al. 2011). The second type of approach uses a study specific reference panel instead of public reference panel such as 1,000 genomes project reference panel (Duan et al. 2013; Deelen et al. 2012). Last strategy uses local reference panel consisting a subset of samples with an array containing many low frequency markers or local sequencing (Joshi et al. 2013; Kreiner-Moller et al. 2014). Local reference panel was used as complementary to public reference panel.

In this study, we studied sample size effect of reference panel on imputation performance of GWAS chip only and combined data. Additionally, we compared imputation performance of GWAS chip only, combined data, and previously reported two-step imputation approach that utilizes local reference panel (Kreiner-Moller et al. 2014). In this analysis, we used only chromosome 1 of the data. We used only imputed variants across all results. Number of imputed variants used for sample size effect analysis and comparison analysis were 10,624 and 10,912, respectively.

For studying sample size effect of reference panel, we performed imputation on GWAS chip and combined data with a subset of samples of original reference panel. Figure 2.6 shows mean estimated r^2 of GWAS chip only and combined data by MAF bins. Regardless of sample size of reference panel, combined data showed better imputation performance than GWAS chip only data. Combined data with 500 samples of reference panel showed enhanced imputation accuracy than GWAS chip only data with 500-848 samples of reference panel.

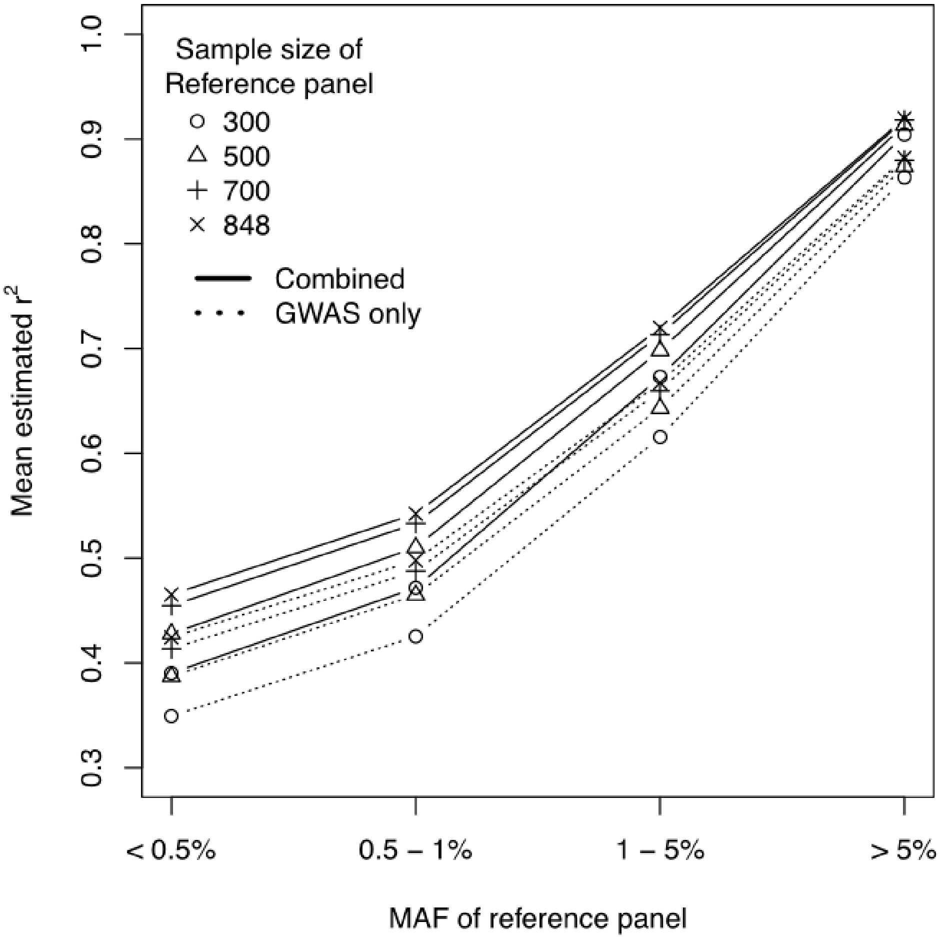


Figure 2.6 Mean estimated r^2 varied by sample size of reference panel

Next, we compared imputation performance of GWAS chip only, combined data, and a two-step approach. In this study, we modified the strategy of a two-step approach that a subset of samples of combined data was used as local reference panel. Table 2.3 and Table 2.4 summarized imputation results and genomic coverage of rare variants, respectively. In overall, combined data outperformed other approaches. Considering genotyping cost of samples, however, two-step imputation approach can be effective strategy since only additionally genotyped 1,000-2,000 samples can increase approximately 5% of mean estimated r^2 and achieve similar genomic coverage to those of combined data.

Table 2.3 Mean estimated r^2 of two-step imputation approach

Sample size of local reference panel	**0.3 – 0.5%**	**0.5 – 1%**	**1 – 5%**	**≥ 5%**
0 **(GWAS chip only)**	0.423	0.498	0.668	0.882
500	0.431	0.514	0.696	0.904
1,000	0.440	0.520	0.700	0.905
2,000	0.444	0.525	0.704	0.907
3,000	0.444	0.527	0.706	0.907
4,000	0.438	0.525	0.706	0.908
Combined	0.463	0.542	0.720	0.920

Table 2.4 Genomic coverage of two-step imputation approach ($r^2 \geq 0.8$)

Sample size of local reference panel	**0.3 – 0.5%**	**0.5 – 1%**
0 **(GWAS chip only)**	0.070	0.158
500	0.086	0.188
1,000	0.090	0.193
2,000	0.095	0.195
3,000	0.093	0.195
4,000	0.096	0.198
Combined	0.092	0.195

2.4 Discussion

In this study, we described the analysis strategy of combined approach that utilizes merged data of GWAS chip and exome array and following imputation analysis. We showed effectiveness of combined approach by analyzing imputation results using reference panel consisting of exome sequencing, exome array, and GWAS chip and genotype panel consisting of exome array and GWAS chip. As a result, the combined approach showed improved imputation accuracy and enhanced genomic coverage, especially for rare variants (MAF < 1%). Combined approach effectively increased imputation accuracy up to 11% and about two times of genomic coverage for rare variants.

Recently, various studies have been reported to increase imputation accuracy of rare variants (Joshi et al. 2013; Kreiner-Moller et al. 2014; Li et al. 2011; Duan et al. 2013; Deelen et al. 2012). Previous studies have mainly focused on utilization of reference panel by constructing the reference panel with sequenced samples (Duan et al. 2013; Deelen et al. 2012) or by increasing samples size of reference panel (Li et al. 2011) or by using complementary information retrieved from local sequencing (Joshi et al. 2013) or local ultra-high-density genotyping arrays (Kreiner-Moller et al. 2014). In a different aspect, our study suggests to use customized chips to increase imputation accuracy of rare variants. If customized chips are available for samples with previously genotyped using GWAS chips, the combined approach would be a possible cost-effective strategy for studying rare variants with increased accuracy and genomic coverage. Moreover, modified strategy adopting previously suggested approaches such as two-step approach can be used to efficiently design imputation based rare variant association study within a limited budget.

Rare variant contents of exome array used in this study have mainly designed based on data from sequenced samples with European ancestry (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). Since we studied samples with East Asian ancestry, exome array may not be the best complementary source of rare variants in this study. Well-designed customized chip based on sequencing data of a specific population would possibly show more increase in imputation accuracy and genomic coverage.

In the present study, we excluded non-imputable variants from the initial reference panel. The main reason for exclusion was to prevent misleading imputation results of extremely rare variants. As Li et al. reported, estimated r^2 of extremely rare variants was not likely to reflect their true value. Instead of excluding non-imputable variants from the initial reference panel, one would filter out those extremely rare variants after imputation analysis. However, there are two possible concerns in using reference panel with non-imputable variants. First, rare variants are computationally difficult to phase due to its few frequency in a haplotype context (Browning and Browning). In this study, we excluded extremely non-imputable rare variants with MAF below 0.3%. Those non-imputable rare variants were about 370K and 74% of exome sequencing data. Therefore, non-imputable variants would introduce possible phasing errors to the reference panel. In addition to possible errors, a large number of variants in the reference panel may require more computational time in imputation analysis.

As a reference panel, we only used study specific sequenced 848 samples. Imputation accuracy can be increased using large reference panels such as reference haplotypes from 1,000 genomes project data. However, Duan et al. reported previously that the reference panel consisting of study specific sequenced samples showed better imputation performance than using reference panel of 1,000 genomes project (Duan et al. 2013). Only modest gain of imputation accuracy (1.5 – 2.3%) was observed when combined reference panel of study specific reference panel and 1,000 genomes project data. Since rare variants are tend to be population specific, relatively small number of samples per a specific ancestry would be limitation of 1,000 genomes project data in imputing rare variants. Upcoming phase 3 of 1,000 genomes project data will provide approximately 2,500 multi-ethnic sequenced samples and may provide more samples with a specific population ancestry.

Although Next Generation Sequencing (NGS) is not efficient approach for a large scale genome study, NGS will become an essential tool in genomics as the cost is decreasing rapidly. Meanwhile, imputation based research strategy would be efficient approach to identify associations between diseases and variants including less common and rare variants.

Chapter 3. Pre-Collapsing Imputation approach

3.1 Introduction

Over the last decade, genome-wide association studies (GWASs) have been successful in unveiling the genetics of human diseases (Bush and Moore 2012). Certainly, GWAS have revealed unprecedented numbers of disease associated genetic variants (Hindorff et al. 2009). As of March 2014, 12,599 single nucleotide polymorphisms (SNPs) from 1,827 published GWASs are included in the National Human Genome Research Institute GWAS catalogue, a curated resource of SNP-trait associations (Hindorff et al. 2009; Welter et al. 2014). However, despite previous efforts to discover the genetic sources of diseases, variants identified by GWASs have been shown to explain only a small proportion of the phenotypic variance observed (Manolio et al. 2009; Lander 2011). Since previous GWASs were largely based on common variants, other possible sources of missing heritability would be rare variants (minor allele frequency [MAF] < 1-5%), structural variants, gene-gene interactions, and gene-environment interactions (Manolio et al. 2009).

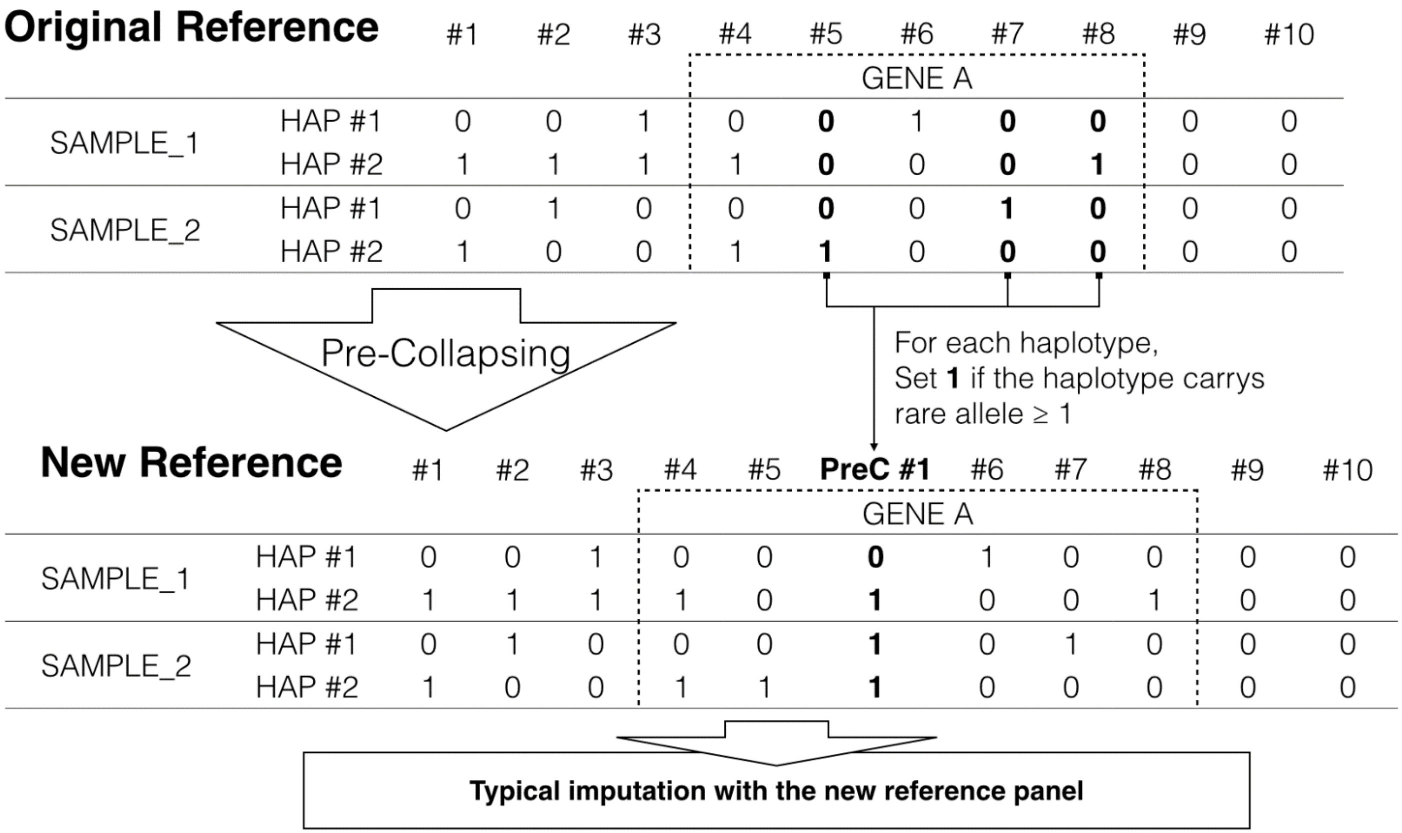
With the recent advances in massively parallel sequencing, rare variants are gaining increasing attention in GWASs (Zuk et al. 2014). Indeed, recent sequencing based association studies discovered previously unknown less common (MAF = 1-5%) and rare variants (MAF < 1%) associated with various phenotypes such as high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, schizophrenia, Alzheimer's disease, and nephropathy (Morrison et al. 2013; Cooke Bailey et al. 2014; Cruchaga et al. 2014; Lange et al. 2014; Purcell et al. 2014). Two approaches are commonly used in association studies utilizing rare variants (Lee et al. 2012; Zuk et al. 2014). One approach is the individual variant test that is typically used in

GWAS. Although it is the simplest to use, this strategy is underpowered because of the low allelic frequencies and abundance of rare variants (Bansal et al. 2010; Zuk et al. 2014). The second approach, which is more powerful, is the region-based association tests, which collapses sets of rare variants and then tests for an association between multiple variants and a phenotype (Bansal et al. 2010; Lee et al. 2012; Zuk et al. 2014).

Given the relatively high cost of the current high-throughput sequencing technology as well as the amount of computing power required, it is not yet feasible to use next-generation sequencing to analyze the number of samples required to identify associations between rare variants and phenotypes (Auer et al. 2012; Magi et al. 2012). Recently, imputation has been widely used as another approach to comprehensively and cost effectively search for rare variants in large-scale cohorts (Auer et al. 2012; Pasaniuc et al. 2012). Imputation estimates untyped markers that are not directly genotyped in the SNP chip (Marchini and Howie 2010). Typically, imputation analysis requires a reference panel with a dense set of markers. The thousands of sequenced samples obtained from the 1,000 Genomes Project are commonly used as an external reference (Howie et al. 2011; Huang et al. 2012; Sung et al. 2012). Study-specific reference panels (Auer et al. 2012; Pasaniuc et al. 2012) are also a powerful resource, especially for rare variants, since rare variants tend to be population specific (Bodmer and Bonilla 2008). For example, by imputation-based association analysis using the 1,000 Genomes Project, Magi et al. identified previously unknown variants associated with coronary artery disease from 17,000 Wellcome Trust Case Control Consortium study samples that had already been extensively analyzed (Magi et al. 2012). Another previous study also performed imputation based association analysis on blood cell traits by using study-specific reference panel containing whole exome sequenced samples (Auer et al. 2012). The other study reported that association analysis followed by imputation analysis using extremely low-coverage sequencing data increased power for GWAS (Pasaniuc et al. 2012).

Despite its cost effectiveness and efficiency, the use of imputation on rare variants has a substantial disadvantage because of the inaccuracy of imputed genotypes (Li et al. 2011; Auer et al. 2012). Auer et al. reported that only 7.3% of imputed rare variants (MAF = 0.1%-0.5%) were available after stringent imputation quality control (estimated r^2 threshold = 0.9) (Auer et al. 2012). The use of inaccurate imputed rare variants could distort the results of region-based association tests, which have become the standard method of analysis for rare variants. Moreover, estimated r^2, one of the quality metrics for imputation, is not a good estimator for extremely rare variants (MAF $\leq$ 0.5%) (Li et al. 2011). Two solutions for enhancing the accuracy of the imputation of rare variants have been proposed: (1) increasing the reference sample size by thousands of samples (Li et al. 2011), or (2) using chips designed to tag rare variants and population-specific variants (Li et al. 2011; Joshi et al. 2013). However, these solutions cannot be immediately applied to existing genotype data since additional experiments would be required. Therefore, a new method for increasing the accuracy of the imputation of rare variants is necessary.

In this study, we propose a pre-collapsing imputation (PreCimp) method to improve the imputation accuracy of rare variants in terms of collapsed variables (Figure 3.1 and Figure 3.2). The proposed method uses variants from a phased reference panel to make collapsed variables and then inserts these pre-collapsed variables (PCVs) into the original reference panel to make a new reference panel. Typical imputation with the new reference panel can impute PCVs into the genotypes from study samples at only a computational cost. To evaluate our method, we built a reference panel from 848 samples with data from exome sequencing, a GWAS chip, and an exome chip. PreCimp was then performed on 5,349 samples obtained from the Korea Association REsource (KARE) project (Table 3.1) (Cho et al. 2009).



Original Reference

		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
					GENE A						
SAMPLE_1	HAP #1	0	0	1	0	**0**	1	**0**	**0**	0	0
	HAP #2	1	1	1	1	**0**	0	**0**	**1**	0	0
SAMPLE_2	HAP #1	0	1	0	0	**0**	0	**1**	**0**	0	0
	HAP #2	1	0	0	1	**1**	0	**0**	**0**	0	0

For each haplotype,
Set **1** if the haplotype carrys
rare allele ≥ 1

New Reference

		#1	#2	#3	#4	#5	**PreC #1**	#6	#7	#8	#9	#10
					GENE A							
SAMPLE_1	HAP #1	0	0	1	0	0	**0**	1	0	0	0	0
	HAP #2	1	1	1	1	0	**1**	0	0	1	0	0
SAMPLE_2	HAP #1	0	1	0	0	0	**1**	0	1	0	0	0
	HAP #2	1	0	0	1	1	**1**	0	0	0	0	0

Typical imputation with the new reference panel

Figure 3.1 Pre-collapsing method of PreCimp

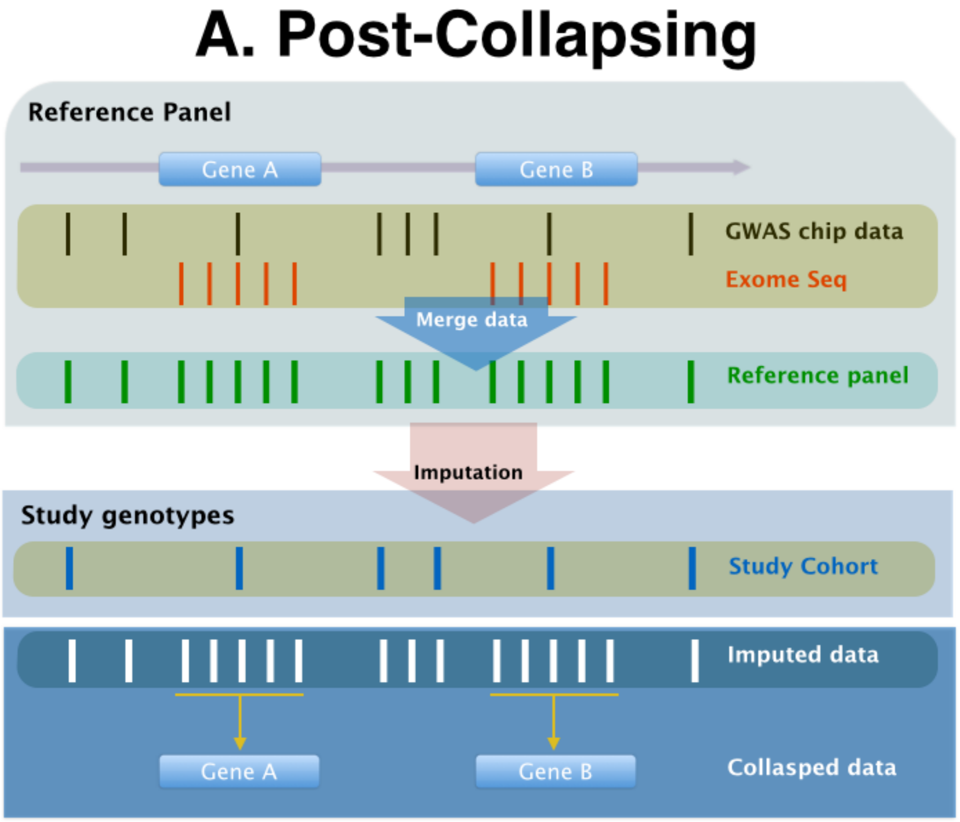


Figure 3.2 Schematic representations of the post-collapsing and pre-collapsing methods used in this study

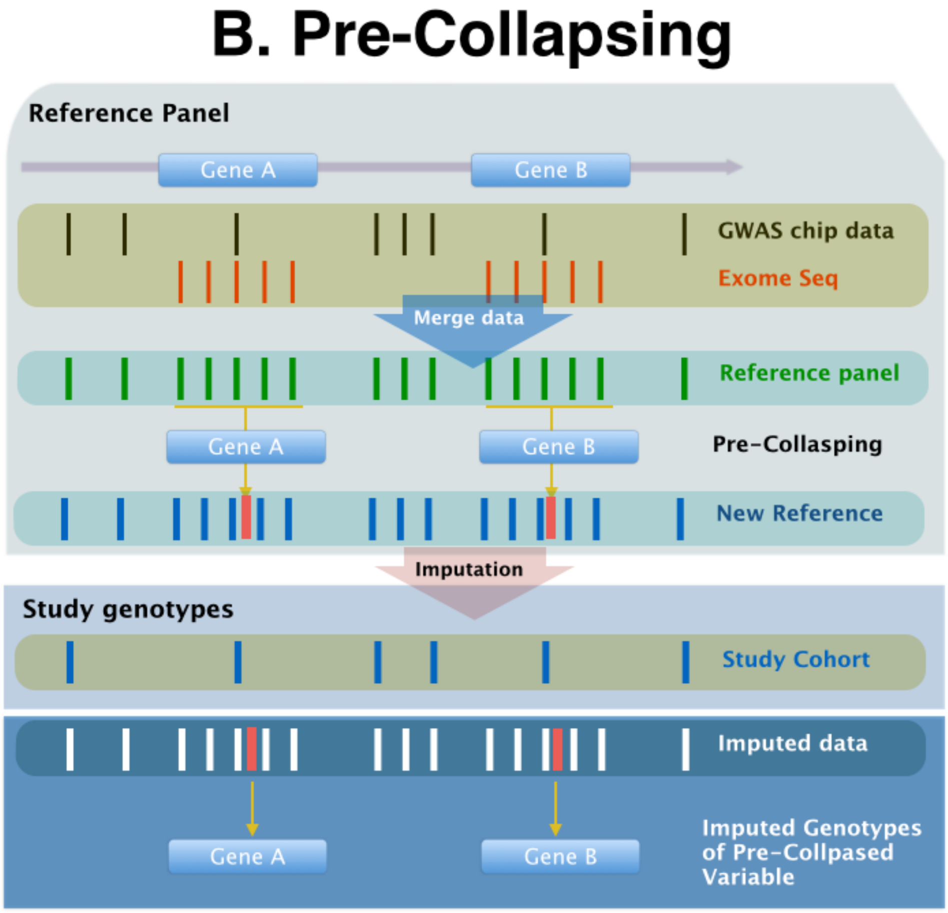


Figure 3.2 Schematic representations of the post-collapsing and pre-collapsing methods used in this study

Table 3.1 Datasets used in this study

Category **(# of samples)**	**Exome sequencing**	**GWAS chip** **AFFY 5.0**	**Exome chip**
# of variants	500,821	344,366	66,196
Reference panel (848)	O	O	O
Genotype panel (5,349)	X	O	X
True data (5,349)	X	X	O

3.2 Materials and Methods

3.2.1 Subjects

Study subjects from the KARE project were recruited from two prospective population-based cohorts as a part of the Korean Genome Epidemiologic Study project. A total of 10,038 participants aging from 40 to 69 years old were registered from both cohorts at the baseline study for two years starting from 2001. A detailed description of KARE has been given in a previous paper (Cho et al. 2009). The study using KARE samples was approved by two independent institutional review boards at Seoul National University and the National Institute of Health, Korea. Liver enzyme, aspartate aminotransferase (AST), was obtained in the morning before the first meal of the day. Participants were removed from subsequent analysis if taking any medication likely to influence on the liver enzyme trait (Kim et al. 2011)

3.2.2 Exome sequencing

Approximately 10,000 exomes (~18,000 genes) from five ethnic groups have been sequenced by the The Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Ethnic Samples Consortium at the Broad Sequencing Center using Agilent Human Exon v2 capture. Some of the KARE samples, including 538 samples from type 2 diabetes cases and 579 samples from controls, were included in this dataset. After quality control on DNA and sequenced samples, 1,087 samples were retained for further analysis. Alignment and variant calling process were performed based on the reference genome hg19. The Genome Analysis Toolkit v2 was used to call the variants (McKenna et al. 2010). In this study, we used 500,821 autosomal variants of 848 Korean samples to build population specific reference panel.

3.2.3 GWAS and exome chip genotyping

KARE study subjects were genotyped with two genotyping platforms: the Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix Inc., San Diego, CA, USA) and the Illumina HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA, USA) exome array. Genotyping using the Affymetrix SNP Array 5.0 and quality control procedures have been described in detail previously (Cho et al. 2009). Briefly, samples with a high missing rate (>4%), gender discrepancy, excessive heterozygosity, or cryptic first degree relatives were removed. Then, those SNPs with Hardy-Weinberg equilibrium p-values < 10^{-6}, genotype call rates < 95%, and MAF < 0.01 were also removed from the set. After the remaining SNPs were annotated using the Affymetrix annotation file (see Web Resources) without positional information were eliminated from further analysis. Finally, 8,842 samples with 344,366 autosomal SNPs remained, which were used for the imputation analysis. Of these previously genotyped samples, 6,197 samples were genotyped using exome array. All these samples passed the following exclusion criteria: call rate < 99%, excessive heterozygosity, and gender inconsistency. Then, variants with call rate < 0.95, Hardy-Weinberg equilibrium p-values < 10^{-6}, duplicated markers, and monomorphic markers were removed, so that 66,196 of the initial 242,901 variants were taken forward for further analysis. Among 6,197 samples, 848 samples used for constructing reference panel and remaining GWAS chip and exome chip of 5,349 samples were used as genotype panel and true dataset, respectively (Table 3.1).

3.2.4 Building the population specific exome reference panel

We then constructed a population-specific exome reference panel by merging data obtained from 848 identical samples via exome sequencing, exome array, and GWAS chip. Initially, there were 344,366, 66,196, and 500,821 variants obtained from Affymetrix 5.0, exome chip, and exome sequencing data, respectively. Prior to merging these variants, we excluded variants in that overlapped among the three platforms. The inclusion priority was in the following order: exome sequencing data, exome chip data, and Affymetrix 5.0 data. The number of unique variants obtained from the Affymetrix 5.0, exome chip, and exome sequencing data were 337,058, 18,811, and 500,821, respectively. The merged panel initially contained 856,690 variants. After extremely rare variants with MAF < 0.3% were excluded (Li et al. 2011), the merged panel contained 487,381 variants and phased using the ShapeIT v2 program to build the phased reference panel for imputation analysis (Delaneau et al. 2012).

3.2.5 Pre-collapsing and post-collapsing based imputation

The collapsing method is an approach that collapses rare variants within a region (Li and Leal 2008). For imputed rare variants, we defined post-collapsing (PostC) and pre-collapsing imputation (PreCimp) methods as follows (see Figure 3.1 for pre-collapsing method of PreCimp, see Figure 3.2 for schematic representations of the PostC and PreCimp methods). PostC method is an approach that is typically used in region-based association studies. A collapsed variable X of imputed rare variants for the i^{th} individual is defined as

$$x_i = \begin{cases} 1\, if \;\; the\ number\ of \;\; rare\ alleles \geq 1 \\ 0 \qquad\qquad\quad otherwise \end{cases}$$

The PreCimp method is an approach that collapses rare variants in a reference panel and generates a new reference panel by inserting these PCVs into the original reference panel. For this method, variants for each haplotype in the reference panel are collapsed. A collapsed variable X for the *j*th haplotype of the *i*th individual in the reference panel is defined as

$$x_{ij} = \begin{cases} 1\, if\ the\ number\ of\ rare\ alleles \geq 1 \\ 0 \qquad\qquad otherwise \end{cases}$$

In this study, pre-phasing-based imputation was performed for rare variants imputation (Howie et al. 2012). Then the PostC method was applied to imputed rare variants after imputing single rare variants. Prior to PostC, genotypes with maximum posterior probabilities were assigned for imputed genotypes. The PreCimp method consists of two steps. First, a new reference panel containing PCVs was constructed using the PreCimp method. Since PCVs are artificially generated, these new markers need to be assigned to specific chromosomal positions in order to be incorporated into the reference panel. Here, if rare variants were only available in the reference panel, we used the mean positional value of rare variants as the positional value for the PCVs. If one or more rare variants were available in both the reference and genotype panels, we used five different positions: a position one base ahead of the position of the first rare variant (PreCimp-1), the position of the last rare variant (PreCimp-L), the position of the variant with the highest LD r^2 (PreCimp-R2), the mean position of variants used for PreC (PreCimp-M), and the weighted mean position of variants used for PreC (PreCimp-WM). For pre-collapsed variable with *m* rare variants (*k* = 1,2,…,*m*), weighted mean position was defined as

$$Weighted\ mean\ position = \sum_{k}^{m} MAF_k Position_k \,/\, \sum_{k}^{m} MAF_k$$

Next, typical pre-phasing based imputation with the new reference panel was performed. Imputation analysis was performed using minimac software (Howie et al. 2012).

3.2.6 Comparison of imputation performance

For gene-based collapsing, rare variants were selected for further analysis if they were available in the true dataset, the exome chip data. Rare variants of true data set were also collapsed using collapsing and collapsing based on haplotypes for PostC and PreCimp, respectively. To measure imputation accuracy, we used dosage r^2 that is squared Pearson correlation between imputed dosages and true genotypes.

3.2.7 Statistical analysis

Prior to association analysis, AST values were transformed with the reciprocal to follow the normal distribution. Region-based association tests were performed by linear regression adjusting age, gender, and recruitment area. Collapsed variables of imputed rare variants using post-collapsing method and dosage values of imputed pre-collapsed variables from pre-collapsing method were used as the independent variable for post-collapsing method and pre-collapsing based imputation method, respectively. To test the difference between dosage r^2 values between imputation results, the Wilcoxon signed-rank test was performed. Statistical analyses were performed using the R program.

3.3 Results

3.3.1 PostC vs. PreCimp methods

We performed a comparison analysis of the imputation performances of the PostC and PreCimp methods. Two sets of collapsed variants were used, MAF1 (collapsing variants with MAF = 0.3 - 1%) and MAF5 (collapsing variants with MAF = 0.3% - 5%). In total, 1,597 genes for MAF1 and 3,830 genes for MAF5 sets were available if a region was defined as a gene region with two or more rare variants. The results from the two sets are compared in Figure 3.3. Figure 3.3A shows that imputation performance was enhanced by the PreCimp method. The proposed approach increased imputation accuracy about 3.4 ~ 6.3% (dosage r^2 0.6 ~ 0.8), 10.9 ~ 16.1% (dosage r^2 0.4 ~ 0.6), and 21.4 ~ 129.4% (dosage r^2 below 0.4) compared with the results of post-collapsing method [Table 3.2]. A Wilcoxon signed-rank test was performed to test the statistical significance of difference in imputation performance and showed that the PreCimp method significantly outperformed the PostC method (p-value < 2.2×10^{-16}).

The difference in dosage r^2 using the PreCimp and PostC methods are shown in Figure 3.3B (MAF5 set). Although the PreCimp method showed increased imputation performance, some collapsed variables with poor performance were also observed. Since the PreCimp method utilizes rare variants in the reference panel based on haplotype information, two factors that could affect the performance would be gene length and the number of rare variants used for PreCimp. Figure 3.4 shows the scatter plot of the number of variants used for PreCimp and gene length in the MAF5 set. Red circles indicate poor performance when PreCimp was used, and the size of circle reflects the magnitude of difference in dosage r^2 between PreCimp and PostC. Genes < 200kb are shown in Figure 3.4A, and genes ≥ 200kb are shown in

Figure 3.4B.

Gene length was a major factor affecting the imputation performance of the PreCimp method. For large genes (about >200kb, about 3% of genes in MAF5 set), the PreCimp method may not be good for improving the imputation accuracy of collapsed variables. However, the performance of PreCimp can be improved by splitting large genes into several small-sized regions. For example, *ASTN2* in the MAF5 set is 803kb in size and has six variants. The values obtained by PostC and PreCimp for dosage r^2 were 0.65 and 0.24, respectively. However, splitting *ASTN2* into two sub-regions for PreCimp increased the value of the mean dosage r^2 for the two regions to 0.68. The increment in dosage r^2 were 0.03 and 0.44, as compared to the values obtained by PostC and PreCimp without splitting, respectively.

We next compared dosage r^2 values of PreCimp and PostC method using haplotype block information. Generally, imputation methods perform better in genomic regions with strong LD than regions with weak LD (Pei et al. 2008; Hao et al. 2009). To obtain haplotype block information, we used LD-based haplotype block recognition software MIG++ implemented in LDexplorer (Taliun et al. 2014). In this analysis, haplotype blocks were obtained using chromosome 1 of the reference panel. There were 42,454 variants in chromosome 1 and 5,970 blocks were detected using default option of MIG++. Median number of variants in haplotype blocks was four. Minimum and Maximum number of variants in haplotype blocks were 2 and 76, respectively. Regions used for collapsing were divided into two groups based on following criteria: regions in haplotype block if all variants used for collapsing were in a single haplotype block, and regions not in haplotype block otherwise. 107 regions were in haplotype blocks and 301 regions were located outside of haplotype blocks. As previously reported (Pei et al. 2008; Hao et al. 2009), PostC and PreCimp both performed better if regions were located in haplotype blocks than regions outside of the haplotype blocks (Figure 3.5). However, difference in dosage r^2 (PreCimp – PostC) was greater for regions in haplotype blocks than regions outside of haplotype blocks. Mean difference in dosage r^2 for regions in haplotype blocks and outside of blocks was 0.059 and 0.047, respectively.

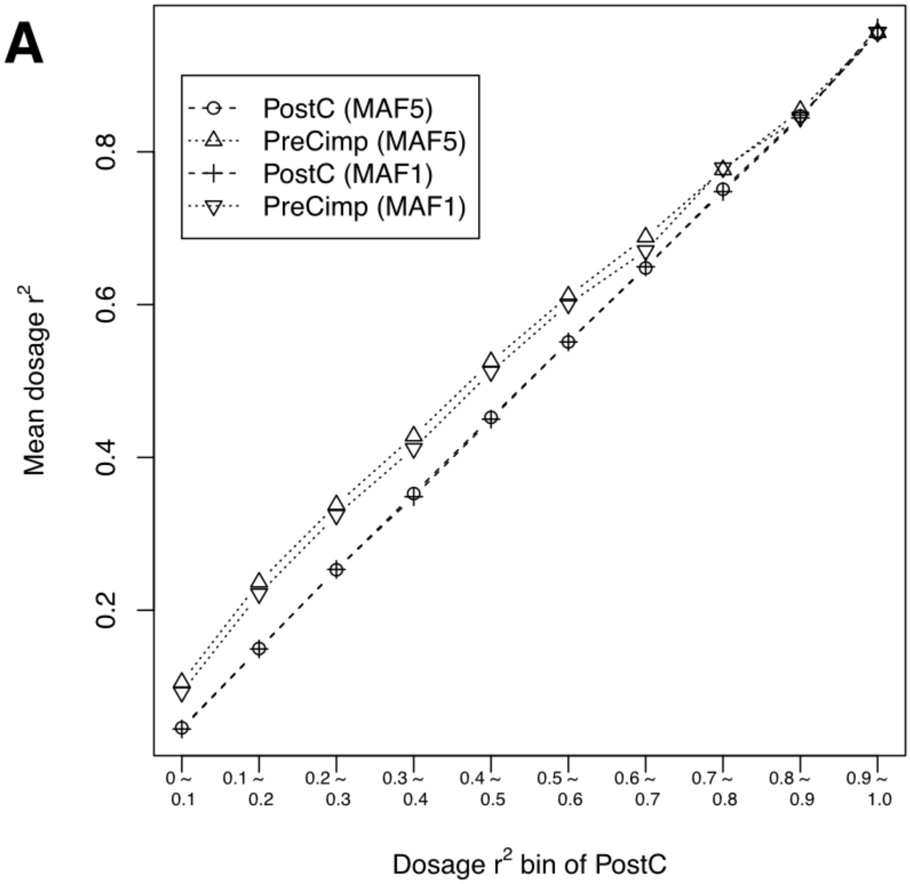


Figure 3.3A Comparison of imputation performance of post-collapsing, and pre-collapsing methods

(A) Comparison of mean dosage r^2 of methods by dosage r^2 bin of PostC method. B shows histogram of difference in dosage r^2 values for the pre- and post-collapsing imputation methods. The red dotted vertical line indicates no difference in dosage r^2.

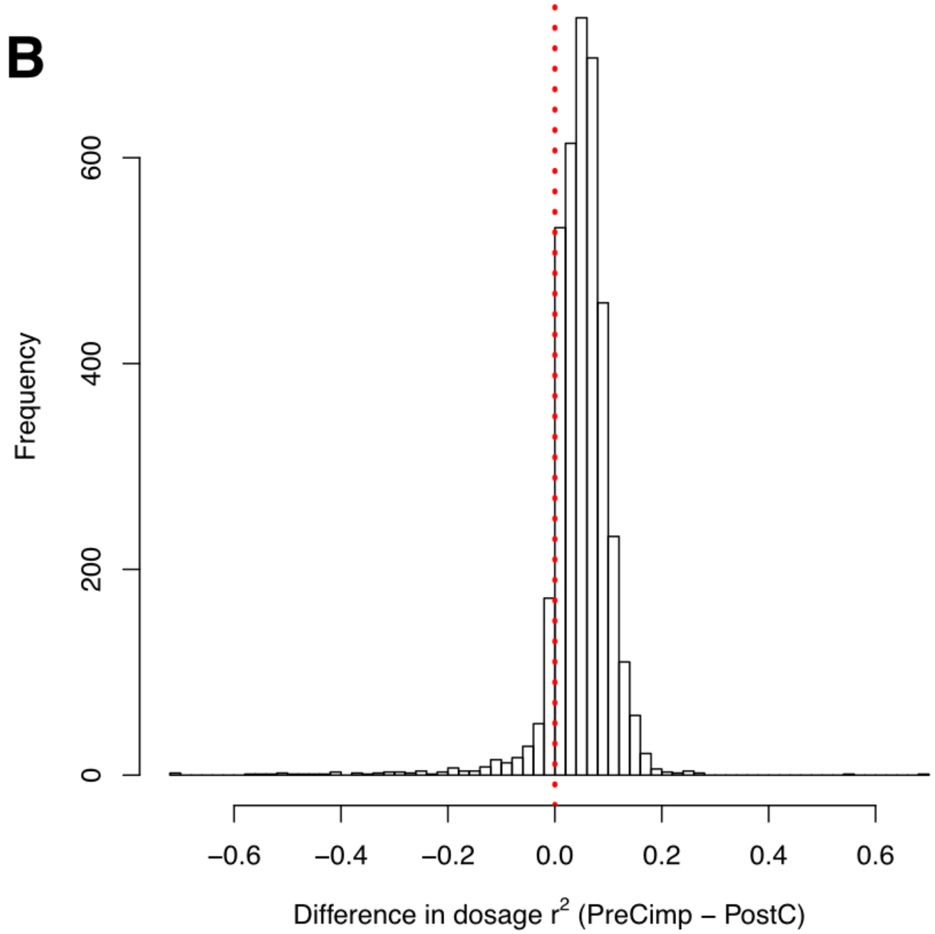


Figure 3.3B Comparison of imputation performance of post-collapsing, and pre-collapsing methods
(A) Comparison of mean dosage r^2 of methods by dosage r^2 bin of PostC method. B shows histogram of difference in dosage r^2 values for the pre- and post-collapsing imputation methods. The red dotted vertical line indicates no difference in dosage r^2.

Table 3.2 Enhanced imputation accuracy by the PreCimp method

Dosage r^2 bin of PostC	**Mean increased in dosage r^2**			**Increased in dosage r^2 (%) (PreCimp – PostC) / PostC**	
	All genes (3,830 genes) (# of genes)	< 200kb (3,717 Genes) (# of genes)	≥ 200kb (113 Genes) (# of genes)	< 200kb	All genes
0 ~ 0.1	0.060 (236)	0.060 (236)	- (0)	129.4%	129.4%
0.1 ~ 0.2	0.087 (230)	0.088 (228)	-0.072 (2)	58.8%	57.9%
0.2 ~ 0.3	0.085 (282)	0.086 (275)	0.039 (7)	34.1%	36.2%
0.3 ~ 0.4	0.075 (357)	0.078 (351)	-0.055 (6)	22.0%	21.4%
0.4 ~ 0.5	0.073 (435)	0.076 (423)	-0.036 (12)	16.8%	16.1%
0.5 ~ 0.6	0.060 (485)	0.064 (464)	-0.028 (21)	11.6%	10.9%
0.6 ~ 0.7	0.040 (506)	0.048 (487)	-0.149 (19)	7.4%	6.3%
0.7 ~ 0.8	0.025 (469)	0.035 (450)	-0.206 (19)	4.7%	3.4%
0.8 ~ 0.9	0.008 (422)	0.018 (401)	-0.196 (21)	2.1%	0.8%
0.9 ~ 1.0	0.001 (408)	0.003 (402)	-0.147 (6)	0.3%	0.1%

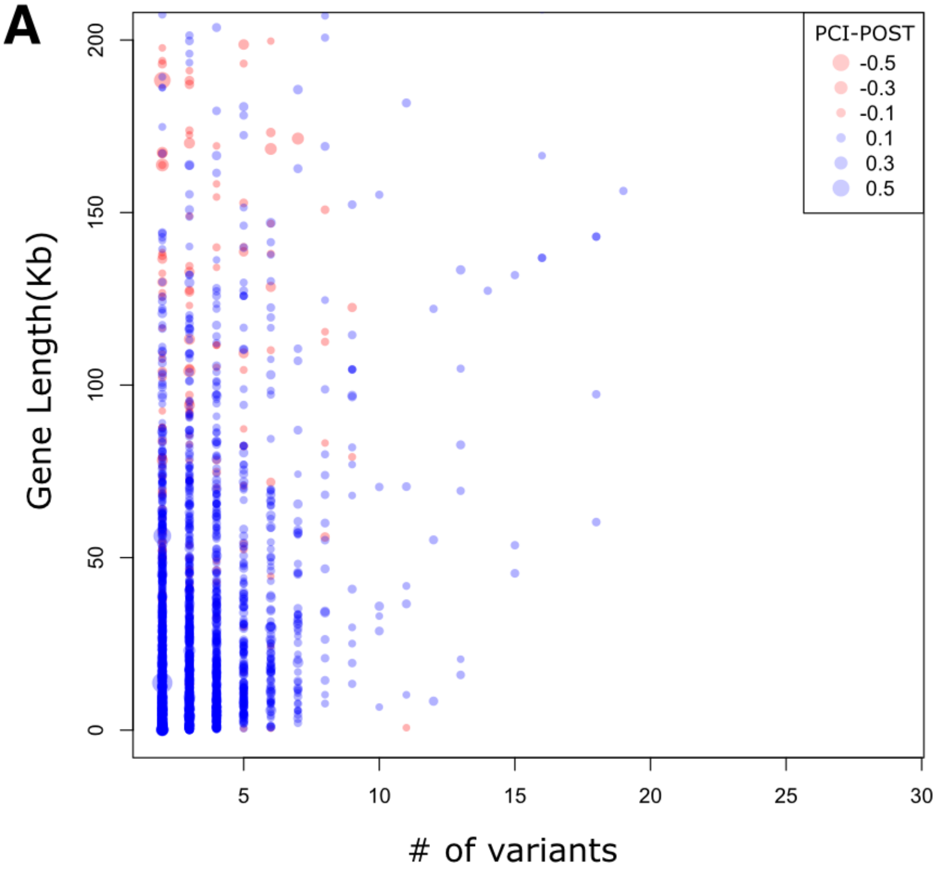


Figure 3.4A Difference in dosage r^2 values by gene size and length
(A) Scatter plot of the number of variants used for pre-collapsing vs. gene length for genes in the MAF5 set with size < 200kb (B) Scatter plot of the number of variants used for pre-collapsing vs. gene length for genes in the MAF5 set with size ≥ 200kb. Circle size represents the magnitude of difference in dosage r^2. Blue color indicates that the pre-collapsing method performs better than the post-collapsing method. Red color indicates that the pre-collapsing performs worse than the post-collapsing method.

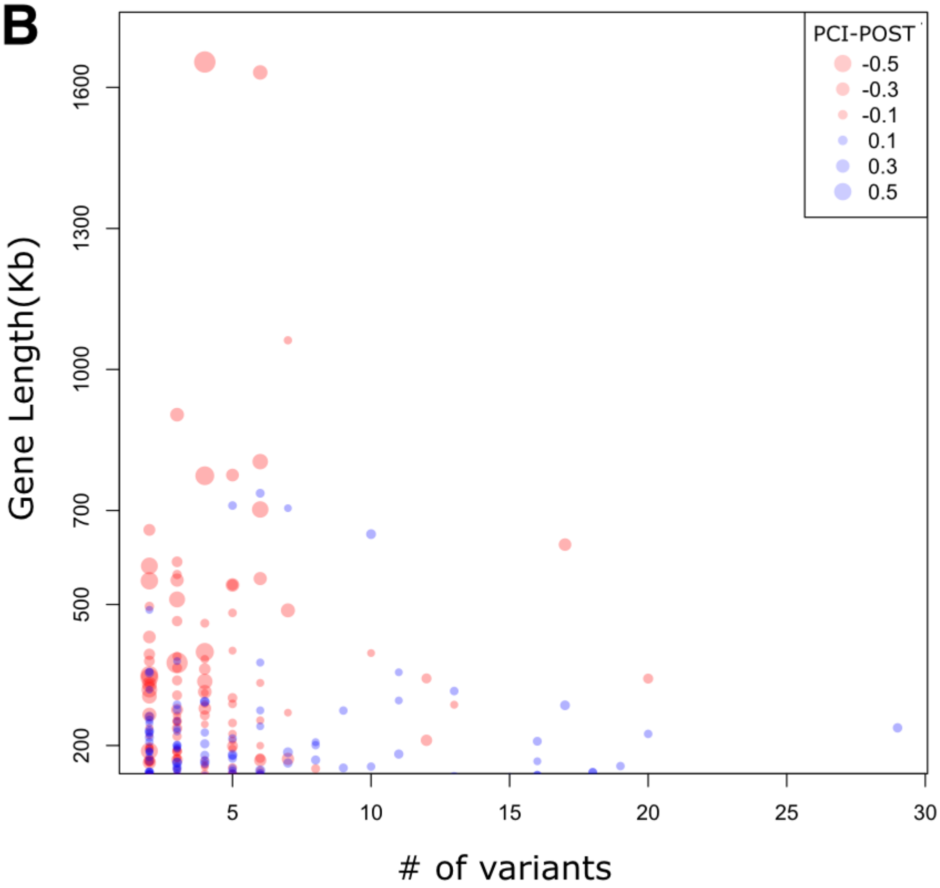


Figure 3.4B Difference in dosage r^2 values by gene size and length
(A) Scatter plot of the number of variants used for pre-collapsing vs. gene length for genes in the MAF5 set with size < 200kb (B) Scatter plot of the number of variants used for pre-collapsing vs. gene length for genes in the MAF5 set with size ≥ 200kb. Circle size represents the magnitude of difference in dosage r^2. Blue color indicates that the pre-collapsing method performs better than the post-collapsing method. Red color indicates that the pre-collapsing performs worse than the post-collapsing method.

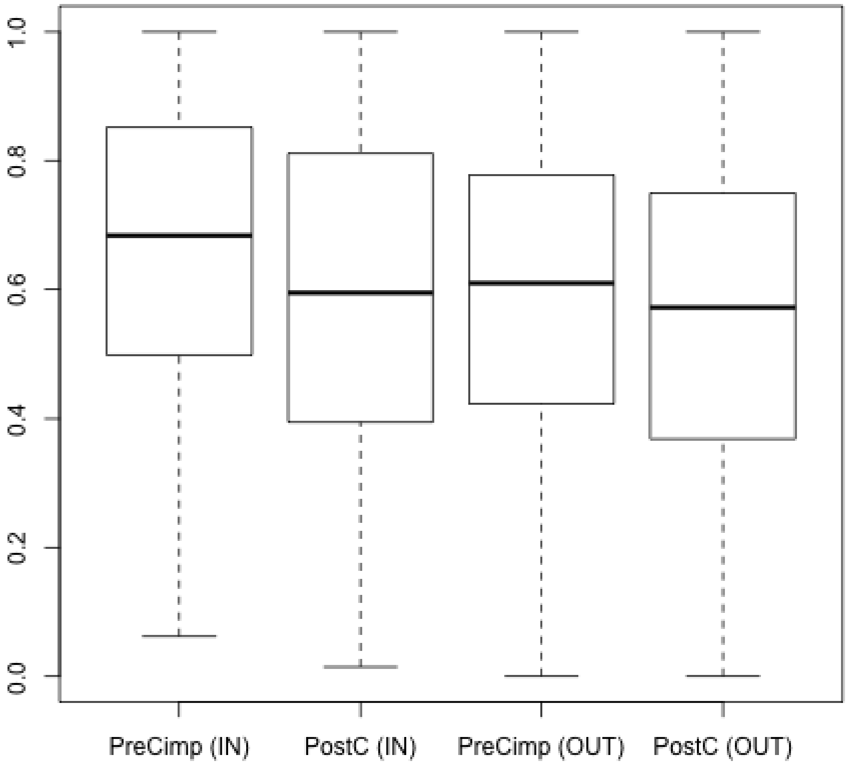
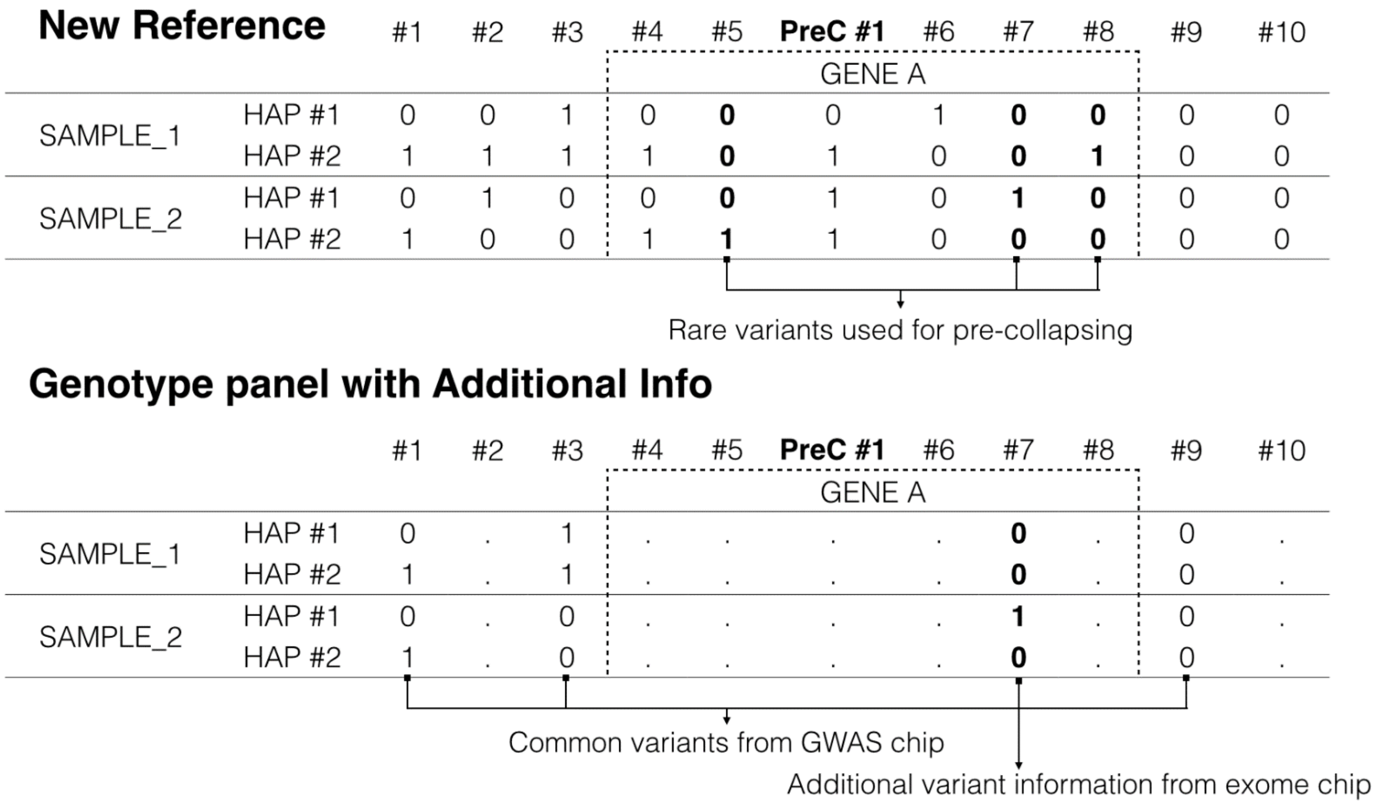


Figure 3.5 Boxplot of dosage r^2 values of PreCimp and PostC. First two boxplot was shown for regions in haplotype blocks. Last two boxplot was shown for regions outside haplotype blocks.

3.3.2 PreCimp with additional information

The PreCimp method greatly enhances imputation accuracy if additional information is used. Since rare variants used for PreCimp are more likely to correlate with PCVs, PreCimp would perform better if one or more rare variants used for PreCimp were available in both the reference and genotype panels. For example, low-cost customized chips containing rare variants, such as exome chip and metabo chip, can be powerful sources of rare variants with additional information (Figure 3.6). Therefore, we analyzed the effect of additional information on the imputation performance of the PreCimp method by adding a variant used for PreCimp into the genotype panel. To maximize the performance, a rare variant with the highest LD r^2 with PCV was selected. Figure 3.7A shows the mean dosage r^2 values obtained by PreCimp without additional information, and PostC (PostC-ADD) and PreCimp (PreCimp-ADD) when additional information was used for imputation (MAF5 set).

The results show that the imputation performance of PreCimp and PostC was greatly improved when an additional variant was added. Furthermore, PreCimp also outperformed PostC. Overall, the mean difference in dosage r^2 values was 0.338 when PreCimp was used either with or without additional information. While dosage r^2 was greatly improved overall, large genes showed relatively small increases in dosage r^2 (Figure 3.7B). For example, in the MAF5 set, there are 2,976 genes with $\leq$ 3 variant (77.7%) and 854 genes with > 3 variants (22.3%). The mean differences of dosage r^2 were 0.233 and 0.368 for genes with $\leq$ 3 variants and those genes with > 3 variants, respectively. For genes > 200kb with > 3 variants (60 genes, 1.6%), the increment of dosage r^2 was dropped to 0.142.



New Reference

		#1	#2	#3	#4	#5	**PreC #1**	#6	#7	#8	#9	#10
							GENE A					
SAMPLE_1	HAP #1	0	0	1	0	**0**	0	1	**0**	**0**	0	0
	HAP #2	1	1	1	1	**0**	1	0	**0**	**1**	0	0
SAMPLE_2	HAP #1	0	1	0	0	**0**	1	0	**1**	**0**	0	0
	HAP #2	1	0	0	1	**1**	1	0	**0**	**0**	0	0

Rare variants used for pre-collapsing

Genotype panel with Additional Info

		#1	#2	#3	#4	#5	**PreC #1**	#6	#7	#8	#9	#10
							GENE A					
SAMPLE_1	HAP #1	0	.	1	.	.	.	.	**0**	.	0	.
	HAP #2	1	.	1	.	.	.	.	**0**	.	0	.
SAMPLE_2	HAP #1	0	.	0	.	.	.	.	**1**	.	0	.
	HAP #2	1	.	0	.	.	.	.	**0**	.	0	.

Common variants from GWAS chip

Additional variant information from exome chip

Figure 3.6 Example of PreCimp approach with additional information

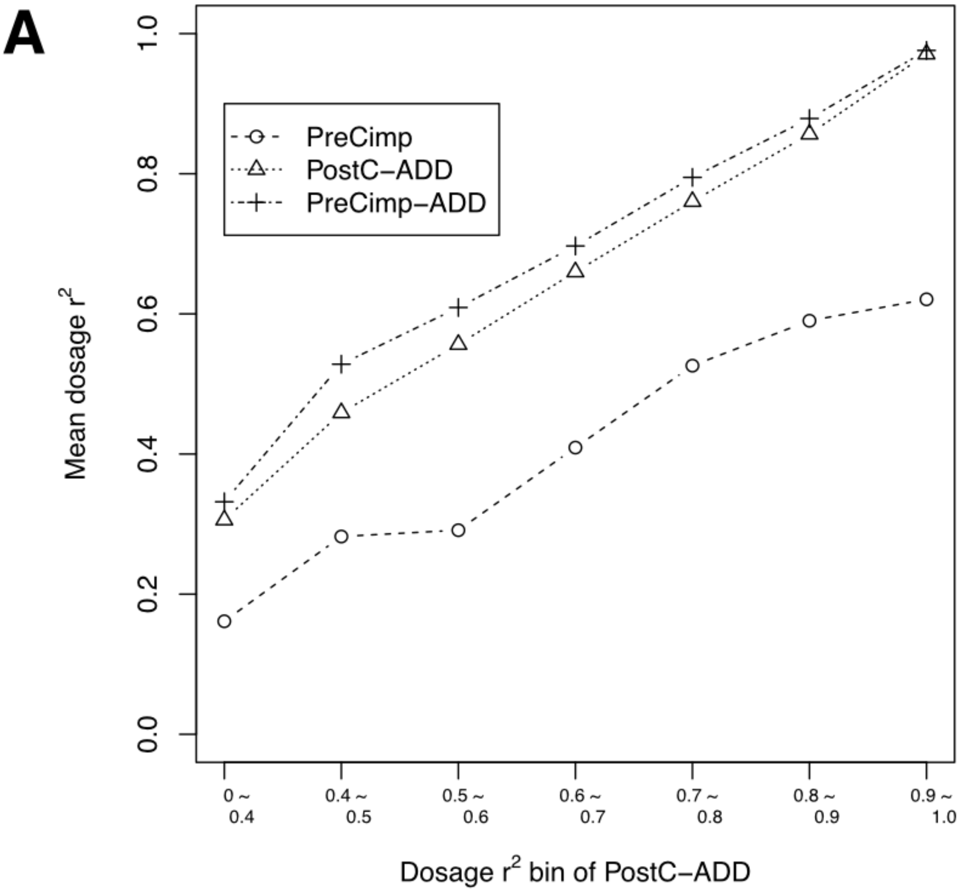


Figure 3.7A Effect of additional information on imputation performance
(A) Comparison of mean dosage r^2 values obtained by the PreCimp without additional information, PostC with additional information (PostC-ADD), and PreCimp with additional information (PreCimp-ADD) are plotted by dosage r^2 bin of PostC method with additional information. (B) The linkage disequilibrium r^2 between pre-collapsed variables and added variants is shown on the x-axis, and the difference between dosage r^2 values obtained using pre-collapsing with additional information and those obtained using the pre-collapsing method without additional information is shown on the y-axis. Circle size represents number of variants used for collapsing.

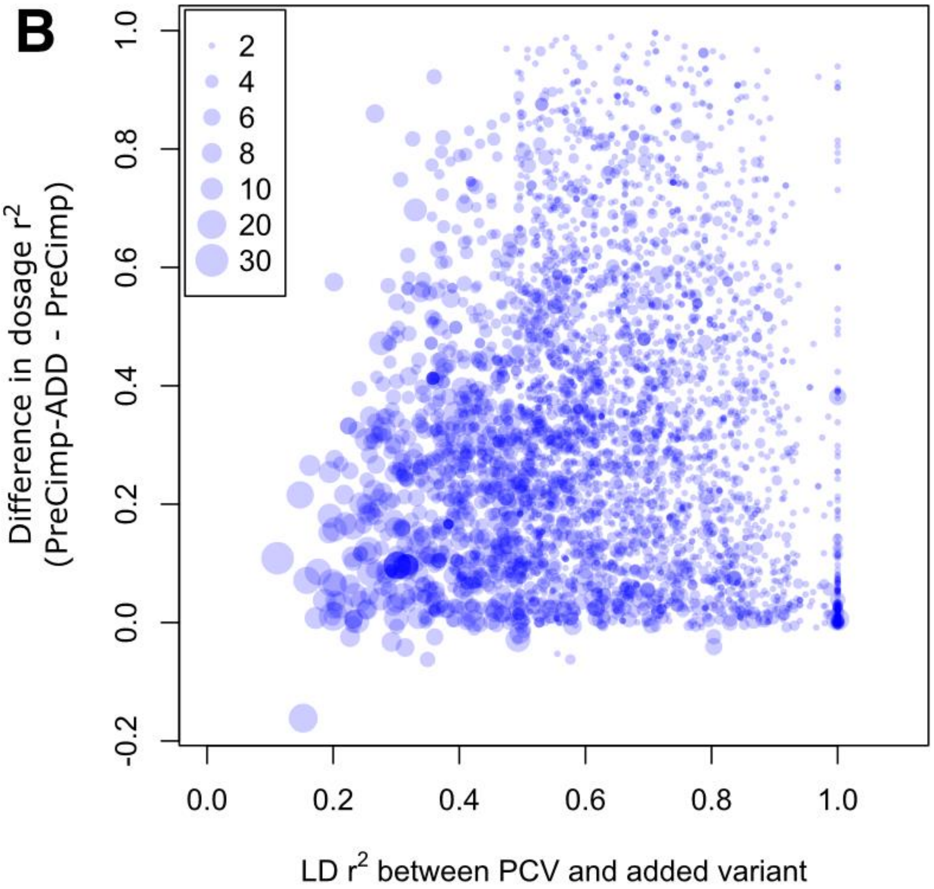


Figure 3.7B Effect of additional information on imputation performance
(A) Comparison of mean dosage r^2 values obtained by the PreCimp without additional information, PostC with additional information (PostC-ADD), and PreCimp with additional information (PreCimp-ADD) are plotted by dosage r^2 bin of PostC method with additional information. (B) The linkage disequilibrium r^2 between pre-collapsed variables and added variants is shown on the x-axis, and the difference between dosage r^2 values obtained using pre-collapsing with additional information and those obtained using the pre-collapsing method without additional information is shown on the y-axis. Circle size represents number of variants used for collapsing.

3.3.3 Effect of PCV position on imputation performance

PCV is an artificial value and has no specific genomic position. Thus, the position of PCVs should be assigned arbitrarily. Since the imputation method predicts untyped markers based on haplotype patterns consisting of sets of correlated variants, the position of PCVs could affect imputation performance, especially if additional variants are used for PreCimp. For the comparison analysis, we used five different positions: a position one base ahead of the position of the first rare variant (PreCimp-1), the position of the last rare variant (PreCimp-L), the position of the variant with the highest LD r^2 (PreCimp-R2), the mean position of variants used for PreC (PreCimp-M), and the weighted mean position of variants used for PreC (PreCimp-WM). In this analysis, we only used chromosome 1. Figure 3.8 shows mean dosage r^2 values obtained using the different PreCimp methods (MAF5 set). PreCimp-R2 showed an improved performance over other PreCimp methods.

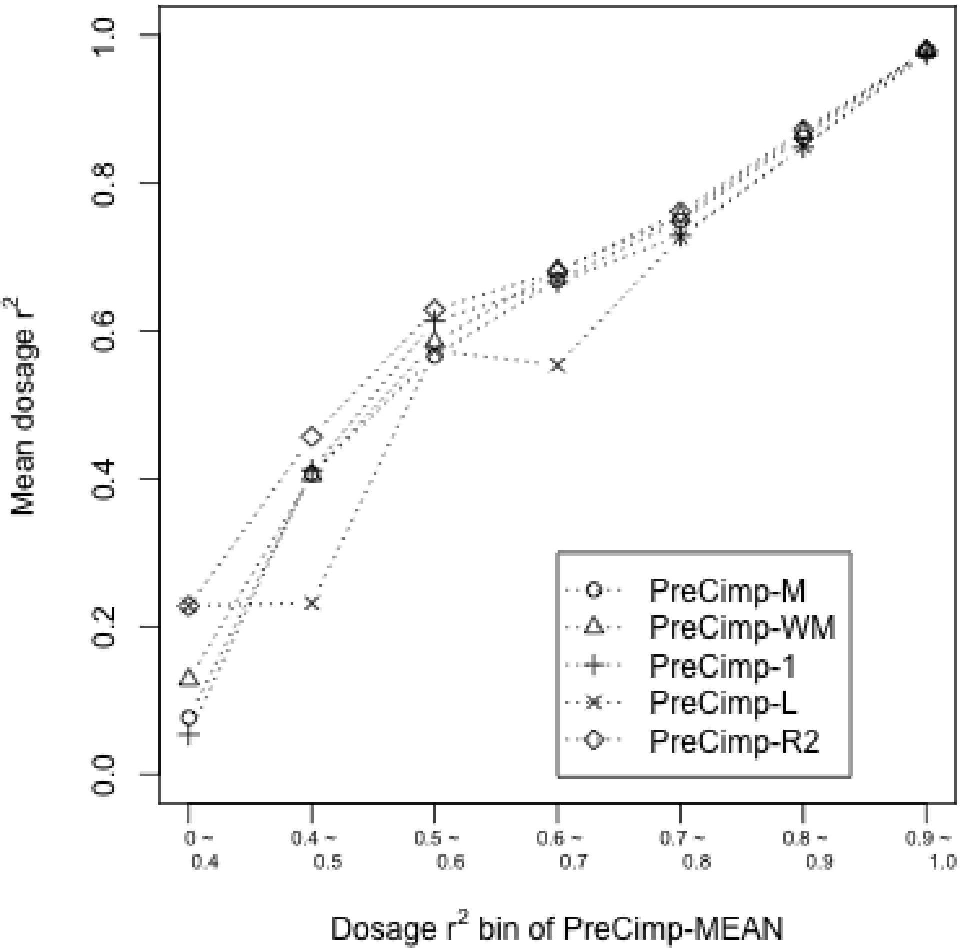


Figure 3.8 The effect of pre-collapsed variable position on imputation performance Comparison of mean dosage r^2 values obtained by the pre-collapsed imputation (PreCimp) method using various pre-collapsed variable positions including mean position of rare variants (PreCimp-M), weighted mean position of rare variants (PreCimp-WM), a position one base ahead of the position of the first rare variant (PreCimp-1), the position of the last rare variant (PreCimp-L), and the position of the variant with the highest LD r^2 (PreCimp-R2)

3.3.4 Example of PreCimp and PostC in association study

We next compared association analysis results of PreCimp and PostC method with those of true dataset. A liver enzyme trait, aspartate aminotransferase (AST), was tested with multiple linear regression analysis on collapsed variables after adjusting recruitment area, gender, and age. For comparison analysis, we used only variants available in the true dataset. Figure 3.9 shows scatter plots using –log10(p-value) of true data, PreCimp, and PostC method. In Figure 3.9B, PostC method showed two false positives that were not significant in the results of true data (p-value > 0.05). Two false results were summarized in Table 3.3. Collapsing of best guessed genotypes from low imputation quality variants caused misleading results. For example, dosage r^2 values of two variants used for PostC was 0.157 and 0.173. However, dosage r^2 of collapsed variable by PostC was 0.022 while dosage r2 of imputed pre-collapsed variable by PreCimp was 0.148.

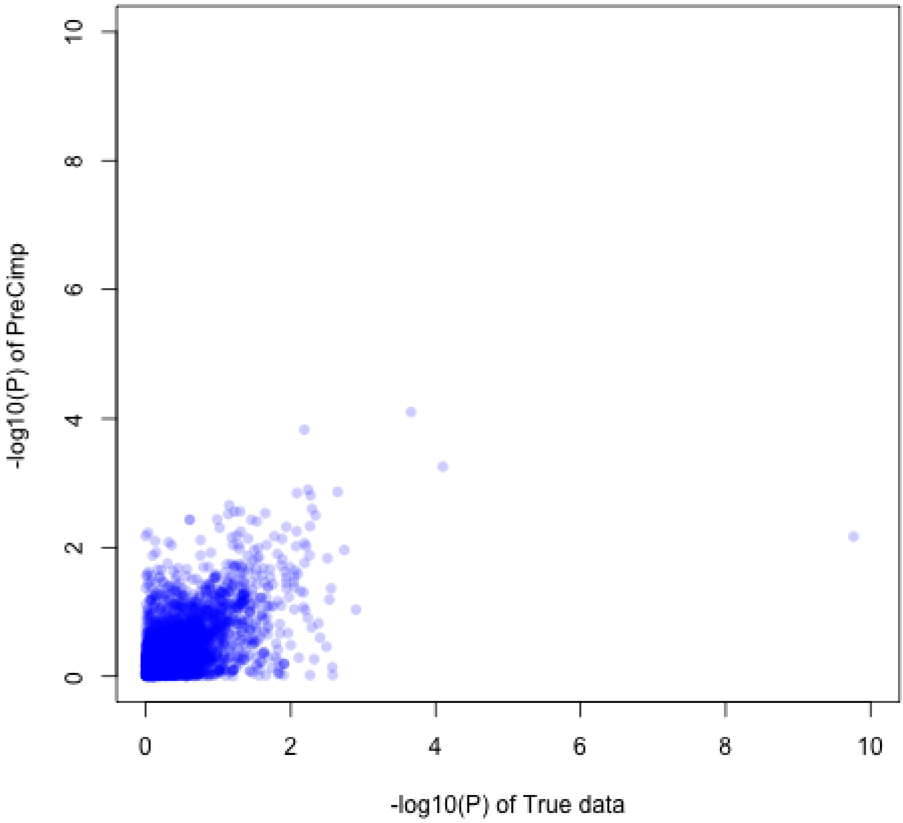


Figure 3.9 Scatter plot of –log10(p-value) (A) True dataset vs. PreCimp

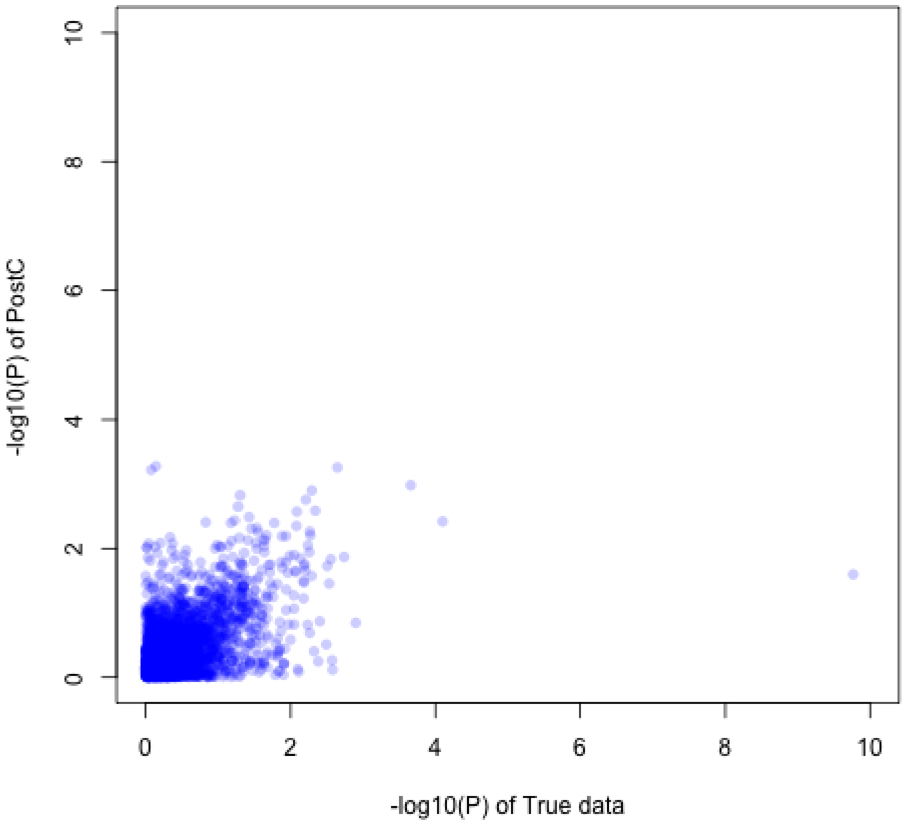


Figure 3.9 Scatter plot of –log10(p-value) (B) True dataset vs. PostC.

Table 3.3 False results by PostC method

Gene	Dosage r^2			P-value		
	Single variant	PostC	PreCimp	PostC	PreCimp	True data
CNOT3	0.157 0.173	0.022	0.148	5.3×10^{-4}	6.8×10^{-2}	7.2×10^{-1}
GAS6-AS1	0.165 0.271	0.062	0.160	6.0×10^{-4}	2.7×10^{-2}	8.3×10^{-1}

3.4 Discussion

In this study, we proposed a PreCimp method to improve the accuracy of imputation of rare variants by using collapsed variables. Using exome sequencing and chip data, we demonstrated that the proposed PreCimp method enhances the imputation performance of collapsed variables. For example, the imputation accuracy of genes with low dosage r^2 (< 0.6) was improved by approximately 10.9 – 129.4% (Table 3.2). Moreover, the performance was greatly enhanced if the variants used for PreCimp were also used in the imputation analysis. If available, customized chips such as exome chip and metabo chip can provide additional rare variants to the genotype panel so that the imputation accuracy of collapsed variables would be greatly increased. In addition, we investigated the effect of PCV position on imputation performance. Our results show that, if additional variants are available, imputation performance is increased by placing PCVs next to the added variants with the highest LD.

The major advantages of the proposed approach are feasibility and flexibility in implementation. The PreCimp method simply builds a new reference panel and then performs standard imputation analysis with the new reference, which can impute collapsed variables more accurately. Since PreCimp uses the information of phased reference haplotypes, construction of new reference panel using PreCimp is computationally feasible and doesn't require a compute intensive process such as haplotype estimation of reference panel. In addition, a coding scheme utilizing the PreCimp method would make it possible to identify disease-associated rare variants on the basis of haplotype. During PreCimp, rare variants are collapsed by each haplotype, and PCVs can be coded as 0, 1,or 2 depending on the number of haplotypes with rare variants.

Despite these advantages, however, the proposed PreCimp method has three main limitations. First, PreCimp showed poor performance with large genes (>200kb, Table 3.1). Generally, the distance between two variants is negatively correlated with LD, and there is weak correlation between rare variants due to their low allelic frequency. Therefore, collapsing multiple rare variants within large-sized region would result in a low correlation with common markers in the reference panel. It might be that the performance of PreCimp can be improved by splitting large genes into several small sub-regions. Since genes larger than 200kb are likely to show poor performance, we recommend that split large-sized genes into chunks smaller than 200kb. Second, we used imputation via a pre-phasing method based on haplotype information using a bi-allelic coding scheme. Thus, the imputed collapsed variable can only be used as a variable indicating the presence or absence of rare variants. If another imputation strategy is used, a coding scheme based on counting can be used in the PreCimp method. Lastly, the imputed collapsed variables can only be used for burden type association tests. Non-burden type tests such as the weighting method and the sequence kernel association test (Wu et al. 2011) are difficult to use for the imputed collapsed variables. Thus, the proposed method will have to be extended in order to consider various aspects of rare variants in association analyses.

Larger reference panel such as 1,000 genomes project data would enhance imputation accuracy, especially for rare variants. However, rare variants are likely to be population specific (Bodmer and Bonilla 2008). Considering ancestry, 1,000 genomes would not be a good source of rare variants compared to study specific sequencing data. For example, there are only 286 samples with East Asian ancestry in 1,000 genomes project phase 1 dataset. 286 samples are much lower than 848 samples used in this study. Duan et al. reported that imputation performance using study specific reference panel showed better imputation quality than using the reference panel of 1,000 genomes data (Duan et al. 2013). Also concatenation of study specific reference panel and 1,000 genomes reference panel showed only modest gains over study specific reference panel in imputation quality (1.5 ~ 2.3%). Therefore, PreCimp would perform best if population specific reference panel is available. However, PreCimp also can be applied to public reference panel such as 1,000 genomes project data. 1,000 genomes project data provides ancestry information of samples. One can select samples with similar ancestry of study population and apply PreCimp on subset of the reference panel. It is expected that there will be more than 500 samples with a specific ancestry in reference panel of 1,000 genomes project phase 3.

In conclusion, next-generation sequencing technology is becoming an essential research tool in genomics. Although next-generation sequencing is not yet applicable to large-scale population based genome studies, the cost for sequencing is rapidly decreasing. In the meantime, genotype imputation of rare variants is a cost-efficient way to comprehensively search for rare variants. Thus, our PreCimp method is valuable for increasing imputation performance of collapsed variables because it has the ability to enhance the imputation performance of rare variants.

Chapter 4. Imputation based association analysis on liver enzyme traits

4.1 Introduction

Elevated level of γ-glutamyl transferase (GGT), alanine aminotransferase (ALT), and aspartate aminotransferase (AST), plasma liver enzymes, are well known indicator of increased risk of liver diseases (van Beek et al. 2013). Liver enzymes have been reported to be an index of liver injury (Pratt and Kaplan 2000) and a marker of fatty liver (Schindhelm et al. 2006; Targher et al. 2009; Vernon et al. 2011) and oxidative stress (Lee et al. 2008). Therefore, finding factors influencing liver enzyme levels is very important to understand individual difference and also underlying mechanism of liver related diseases.

Heritability of liver enzymes was 32-69%, 22-64%, and 21-61% for GGT, ALT, and AST, respectively (Whitfield and Martin 1985; Bathum et al. 2001; Whitfield et al. 2002; Pilia et al. 2006; Lin et al. 2009; Makkonen et al. 2009; Nilsson et al. 2009; Rahmioglu et al. 2009; Loomba et al. 2010; Sung et al. 2010). As genetic factors have substantial influence on the variation of liver enzymes, numerous GWASs have been conducted to identify associated variants (Yuan et al. 2008; Chambers et al. 2011; Kim et al. 2011). However, reported loci failed to fully explain phenotypic variance. Since previous GWASs mainly focused on common variations (MAF >

5%), identification of less common (MAF 1-5%) and rare variants (MAF < 1%) is warranted.

In this context, we performed exome-wide association analysis by whole-exome imputation on 8,749 samples of combined data comprising of GWAS chip and exome array. Whole-exome imputation and genotyped data using exome array enabled us to examine functional variants among previously known regions and less common or rare variants associated with liver enzyme levels.

4.2 Materials and Methods

4.2.1 Subjects

Korea Association REsource (KARE) project is initiated in 2007. Two prospective cohorts as a part of Korean Genome Epidemiologic Study (KoGES) were participated in this project. There were 10,038 participants aging from 40 to 69 years old. In these prospective cohorts, participants were examined clinical records, anthropometric, and biochemical traits for every two year. A detailed description of KARE has been reported previously(Cho et al. 2009).

The HEXA cohort is one of the KoGES population-based cohorts which were initiated in 2001 aiming to identify risk factors of life-style related complex diseases such as type 2 diabetes, hypertension, and dyslipidemia. Approximately 3,700 of 1,200,000 subjects aged 40-69 from the HEXA cohort were randomly selected as a shared control group for the Korean cancer and coronary artery disease (CAD) GWA studies. Genotyping was conducted with the Affymetrix Genome-Wide Human SNP array 6.0 in 2008.

4.2.2 GWAS and exome chip genotyping

Initially, KARE and HEXA samples were genotyped using three different platforms. The Affymetrix Genome-Wide Human SNP Array 5.0 and SNP Array 6.0 (Affymetrix Inc., San Diego, CA, USA) was used for genotyping samples of KARE and HEXA, respectively. And the Illumina HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA, USA) exome array was used for genotyping a subset of KARE and HEXA samples that were previously genotyped using GWAS chips. A quality control procedure of GWAS chips of both cohorts are described in detail previously (Cho et al. 2009; Kim et al. 2011). Briefly, samples with a high missing rate (>4%), gender discrepancy, excessive heterozygosity, or cryptic first degree relatives were removed. Then, those SNPs with Hardy-Weinberg equilibrium p-values < 10^{-6}, genotype call rates < 95%, and MAF < 0.01 were also removed from the set. After the remaining SNPs were annotated using the Affymetrix annotation file (see Web Resources) without positional information were eliminated from further analysis. Finally, 8,842 samples with about 344K autosomal SNPs and 3,703 samples with 650K autosomal SNPs were remained for KARE and HEXA, respectively. Amog these previously genotyped samples, 6,197 KARE and 3,400 HEXA samples were genotyped using exome array. All these samples passed the following exclusion criteria: call rate < 99%, excessive heterozygosity, and gender inconsistency. Then, variants with call rate < 0.95, Hardy-Weinberg equilibrium p-values < 10^{-6}, duplicated markers, and monomorphic markers were removed, so that 66,196 of the initial 242,901 variants were taken forward for further analysis. Among 6,197 KARE samples, 848 samples used for constructing reference panel and remaining GWAS chip and exome chip of KARE and HEXA samples were used as genotype panel

4.2.3 Building the population specific exome reference panel

Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Ethnic Samples (T2D-GENES) consortium was initiated to identify functional variants associated with type 2 diabetes and its related risk factors. From five ethnic groups, about 10,000 exomes were sequenced at the Broad Sequencing Center using Agilent Human Exon v2 capture (capturing ~18,000 genes). Among ten thousands of samples, 538 type 2 diabetes and 579 control samples from KARE project were included. 1,087 samples were remained for further analysis after quality control on DNA and sequenced samples. For reference genome, hg19 was used for alignment and variant calling process. During the variant calling process, the Genome Analysis Toolkit v2 was used (McKenna et al. 2010). For 1,087 samples, we used 500,821 autosomal variants of 848 Korean samples to construct whole-exome reference panel.

We then constructed a population-specific whole-exome reference panel by merging data of exome sequencing, exome array, and GWAS chip of 848 identical KARE samples. The detailed description is reported in a separate paper (Kim et al. submitted). After merging process, initial reference panel contained 856,690 variants. After excluding non-imputable variants (extremely rare variants with MAF < 0.3%) (Kim et al. submitted), final whole-exome reference panel included 487,381 variants and phased using the ShapeIT v2 program to build the phased reference panel for imputation analysis (Delaneau et al. 2012).

4.2.4 Statistical analysis

For imputation analysis, we used typical pre-phasing based imputation analysis on combined genotype panels consisting of GWAS chip and exome chip (Howie et al. 2012). We used minimac software, a low memory and computationally efficient implementation of the MaCH algorithm (Li et al. 2010). The association of imputed and genotypes SNPs with liver enzymes was tested by linear regression adjusting age, gender, and recruitment area (in case of KARE) using EPACTS (http://genome.sph.umich.edu/wiki/EPACTS). Prior to analysis, all imputed genotypes were assigned as best-guessed genotypes based on posterior probabilities. The meta-analysis was performed using a weighted average method assuming fixed effects with inverse variance using metal software (Willer et al. 2010). Statistical analyses and visualization of the results were performed using the R program.

4.3 Results

We performed whole-exome imputation on combined data consisting of GWAS chip and exome array from KARE and HEXA samples. Since we constructed whole-exome reference panel using Affymetrix SNP 5.0, only a subset of GWAS data of HEXA cohort (Affymetrix SNP 6.0) matched with reference panel was used for imputation analysis. As a result, a total of 8,529 samples were imputed and 487,381 imputed variants were generated. For association analysis, KARE and HEXA samples were analyzed separately. After association analysis, meta-analysis was conducted merging KARE and HEXA association results.

Figure 4.1 is manhattan plot of AST association results of KARE samples. Quantile-quantile plot of AST association results of KARE is shown in Figure 4.2. As displayed in Figure 4.1 and 4.2, spurious signals with very strong statistical significance were observed from imputed variants with low imputation quality score (rsq < 0.4). Therefore, we excluded imputed variants with low imputation quality score (rsq < 0.4) and 461,295 variants were remained.

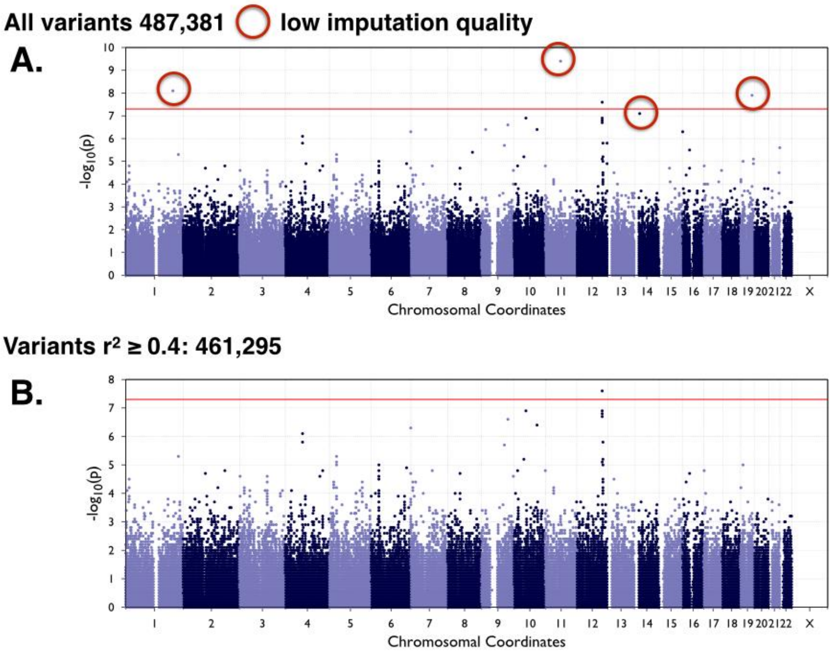


Figure 4.1 manhattan plot of AST association results of KARE samples. (A) initial association results (B) association results after excluding low quality imputed variants

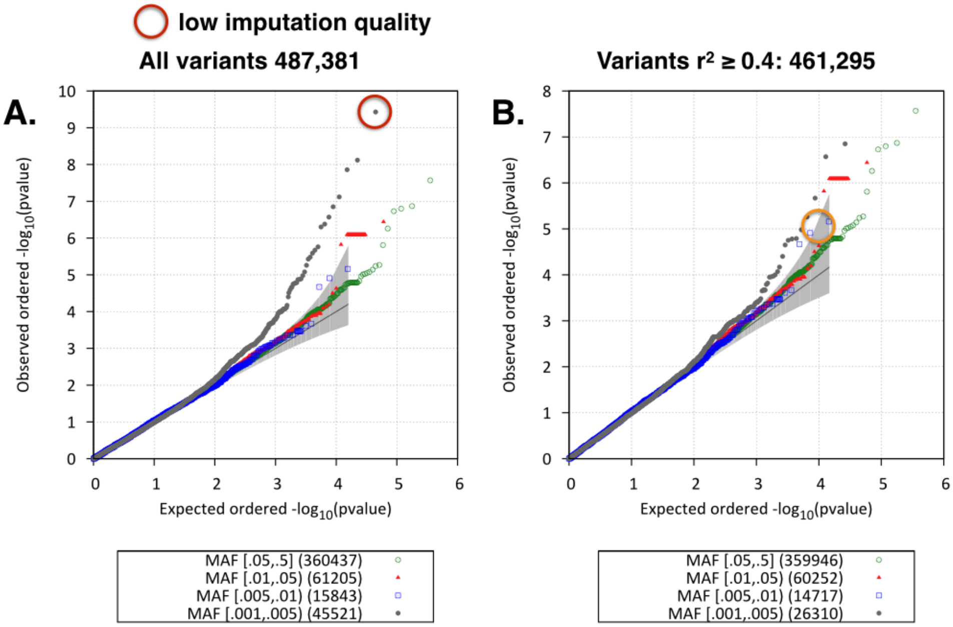


Figure 4.2 quantile-quantile plot of AST association results of KARE samples. (A) initial association results (B) association results after excluding low quality imputed variants

After quality control on imputed variants, there was no false positives by low quality imputed variants (Figure 4.3 and Figure 4.4). We performed meta-analysis with quality controlled association results of KARE and HEXA samples. As a result, we discovered 20 loci with p-value < 5x10^{-6}. Although most of loci were previously reported, we discovered 7 novel loci among them after excluding 2 possible false positives with statistically significant p-value ($P < 0.05$) from heterogeneity test. However, no novel loci reached at the genome-wide significance level (P = 5x10^{-8}). Top signals from ALT, AST, and GGT are shown in Table 4.1, 4.2, and 4.3, respectively.

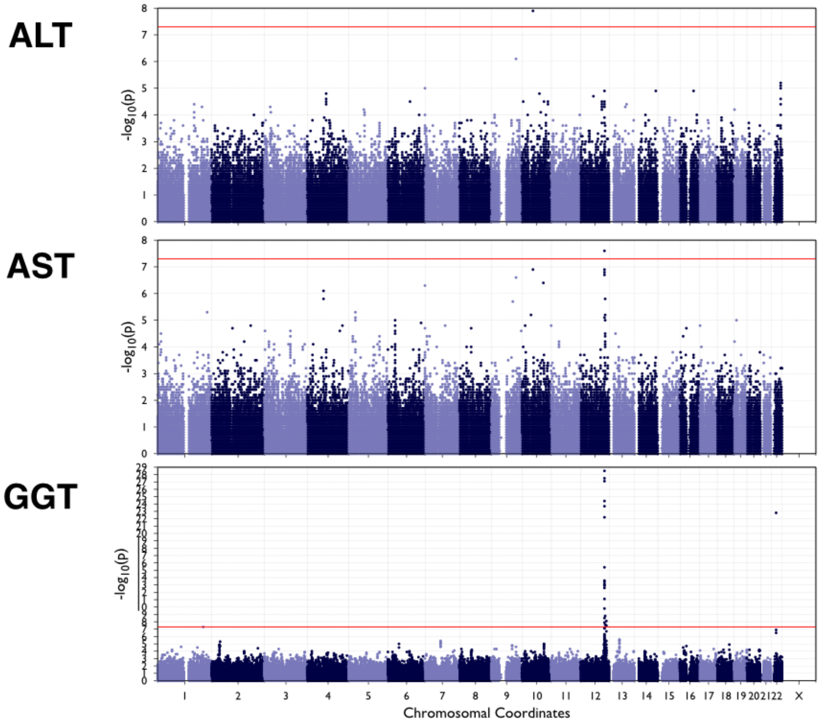


Figure 4.3 manhattan plot of ALT, AST, and GGT association results of KARE samples.

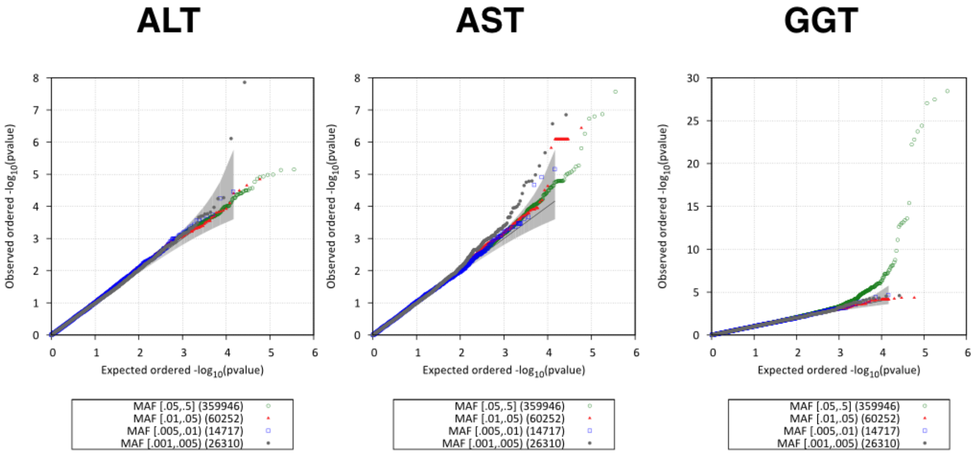


Figure 4.4 quantile-quantile plot of ALT, AST, and GGT association results of KARE samples. (A) initial association results (B) association results after excluding low quality imputed variants

Table 4.1 Top signals from ALT association results

CHR	Function	MAF	P-value KARE	P-value CITY	P-value META	P-value Het	Known
2	Intron: LHCGR|STON1-GTF2A1L	0.224	2.43E-03	2.69E-04	3.66E-06	3.02E-01	X
8	Intergenic	0.285	4.82E-04	1.54E-03	2.64E-06	6.64E-01	O
10	Missense: A1CF [Lys -> Gln]	0.001	1.37E-08	2.51E-01	1.31E-07	1.61E-02	X
12	Intron ALDH2	0.159	1.35E-05	3.27E-02	1.79E-06	3.96E-01	O
22	Synonymous: PNPLA3	0.419	7.29E-06	7.85E-05	2.37E-09	7.05E-01	O

Table 4.2 Top signals from AST association results

CHR	Function	MAF	P-value KARE	P-value CITY	P-value META	P-value Het	Known
1	Intron:OBSCN	0.001	5.42E-06	4.23E-02	1.53E-06	1.86E-01	X
10	Missense:RET Arg -> His	0.009	6.87E-06	6.14E-02	2.74E-06	1.82E-01	X
10	Missense: A1CF Lys->Gln	0.001	1.40E-07	4.11E-01	2.19E-06	1.39E-02	X
10	Intergenic	0.2262	2.74E-04	3.25E-03	2.84E-06	9.21E-01	X
10	Missense: GOT1 Gln->Glu	0.014	3.59E-07	8.98E-05	1.28E-10	9.20E-01	O
12	Missense: ALDH2 Glu->Lys	0.159	2.69E-08	6.73E-04	7.99E-11	5.60E-01	O
13	Intron: COL4A1	0.002	5.77E-02	1.22E-06	2.98E-06	1.19E-02	X
22	Synonymous: PNPLA2	0.419	2.39E-03	7.97E-07	5.19E-08	4.43E-02	O

Table 4.3 Top signals from GGT association results

CHR	Function	MAF	P-value KARE	P-value CITY	P-value META	P-value Het	Known
7	Intron:MLXIPL	0.10	4.02E-06	1.34E-01	2.89E-06	1.97E-01	O
8	Intron:AC135352.1|KIAA1456	0.21	1.23E-04	8.64E-03	3.26E-06	8.71E-01	X
12	missense:ALDH2 Glu -> Lys	0.16	3.40E-29	1.95E-04	6.12E-31	6.04E-03	O
12	Intron:HNF1A	0.48	7.30E-09	1.22E-02	4.83E-10	2.89E-01	O
13	Intergenic	0.46	2.62E-06	9.60E-01	1.07E-03	7.58E-03	X
22	Intron:GGT1	0.25	1.29E-07	1.55E-05	1.16E-11	4.03E-01	O

Intronic and missense variants were newly associated with ALT trait. Nearby genomic region of variants at intron of *LHCGR-STON1-GTF2A1L* have been previously associated with obesity, endometrial cancer, and bipolar disorder. Missense variant at *A1CF* changes amino acid Lys to Gln. *A1CF* (APOBEC1 complementation factor) is a protein-coding gene. *A1CF* was previously associated with anisometropia, and hydrocele. *A1CF* was previously reported to modulate liver regeneration via post-transcriptional regulation (Blanc et al. 2010). Regional association plot of missense variant at *A1CF* gene is shown in Figure 4.5.

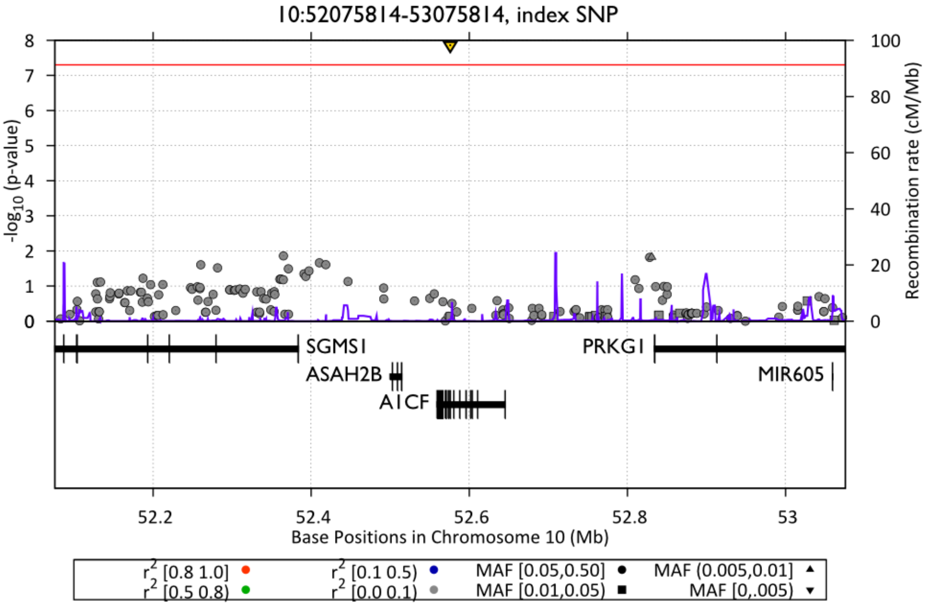


Figure 4.5 regional association plot of missense variant at *A1CF* gene

For AST trait, 4 newly associations were discovered. 4 loci were located at intron of *OBSCN*, exon of *RET*, intergenic region, and intron of *COL4A1*. OBSCN (obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF) is a protein-coding gene. Diseases associated with OBSCN include leiomyosarcoma, and gastrointestinal stromal tumor. *RET* (ret proto-oncogene) is a protein-coding gene. RET has been reported to be associated with diseases such as thyroid cancer, childhood, and sipple syndrome. Also p.G533C mutation of RET was reported to confer predisposition to multiple endocrine neoplasia Type 2A (Oliveira et al. 2011). Regional association plot of variants at *OBSCN* and *RET* is displayed in Figure 4.6 and 4.7, respectively.

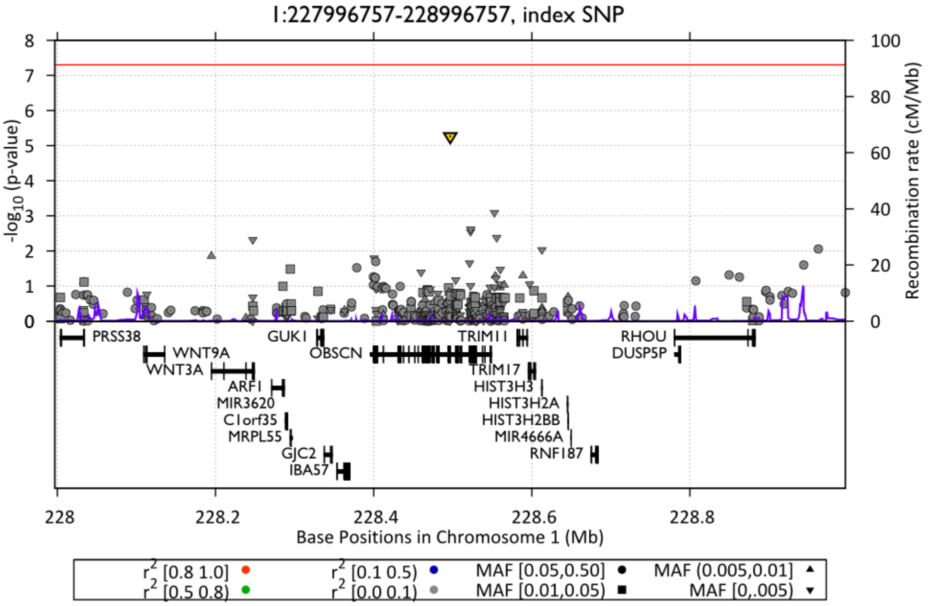


Figure 4.6 regional association plot of missense variant at *OBSCN* gene

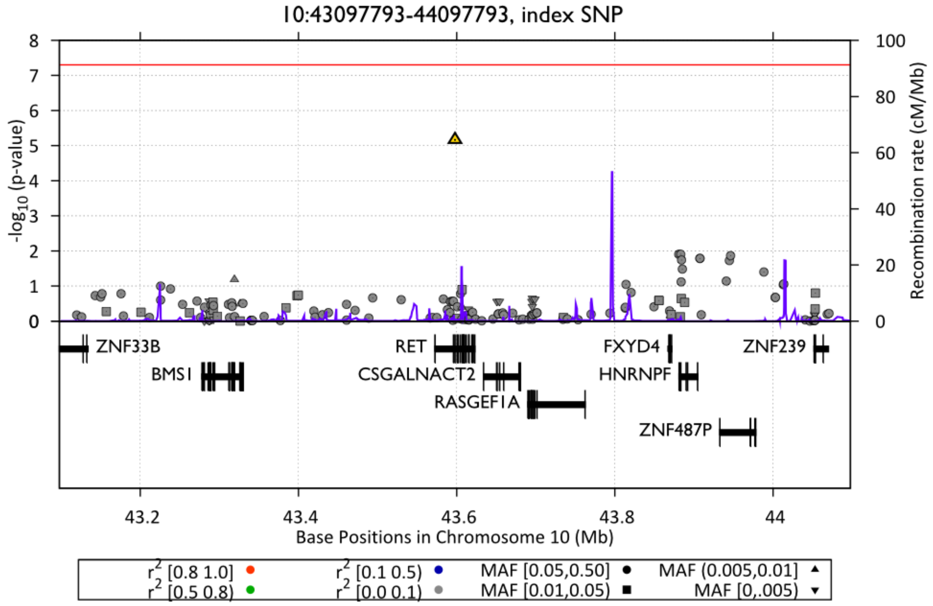


Figure 4.7 regional association plot of missense variant at *RET* gene

4.4 Discussion

In this study, we performed whole-exome imputation on 8,529 samples and subsequent imputation based association study. Meta-analysis of two cohort samples revealed 7 novel associations including two missense functional variants. Interestingly, one missense variant was discovered at *A1CF*. The gene was previously reported to modulate liver regeneration via post-transcriptional regulation (Blanc et al. 2010). Our study would shed light on understanding underlying genetics of liver enzyme related functional variants and its effect on phenotypes.

Although the newly discovered variants in the present study are valuable, those novel variants didn't take forward to replication study in an independent cohort. We reduced the possible chance of discovering false positives by performing meta-analysis on two cohorts and discarding spurious signals with statistically significant from heterogeneity test. However, it would be necessary to perform replication study to further confirm the associations in an independent cohort.

One limitation of our study is the construction of whole-exome reference panel and following imputation. Since extremely rare variants (MAF < 0.3%) were excluded from the original reference panel, we only performed association analysis with limited number of rare variants. Our study will be more powerful by increasing the number of samples in the reference panel or using additional information such as phase 3 reference panel from 1,000 genomes project.

In summary, our study reported 7 novel associations responsible for liver enzymes. Although those associations were not confirmed through replication study, by finding two functional variants, it would be valuable to understand the genetics of liver enzymes.

Chapter 5. Summary and Conclusion

Rare variants have gathered increasing attention as a possible alternative source of missing heritability. Since next generation sequencing technology is not yet efficient in a large scale genomic study, two approaches, imputation and customized chips such as exome array and Metabochip, have been widely used in large scale genome studies. Two approaches have successfully identified numerous less common or rare variants associated with various phenotypes. However, this imputation approach has a limitation due to low accuracy of imputed rare variants, and customized chips are designed only for the specific targets. Various previous studies have reported analysis strategies for improving imputation accuracy of rare variants. Since, previous studies have mainly focused on utilization of reference panel, different aspects of imputation strategy and methodological approach are warranted to more efficiently improve imputation accuracy of rare variants.

For a new strategy, we proposed the combined approach that adopts advantages of imputation and customized chip was described. In this approach, we constructed exome reference panel using 848 identical samples with whole exome sequencing data, GWAS chip, and exome array data. Using this population specific whole-exome reference panel, we performed imputation analysis on 5,349 samples of combined data including GWAS chip and exome array. We compared imputation results of exome array, GWAS chip only, and combined data. As a result, the combined approach increase about 11% in imputation accuracy and enhanced about two times of genomic coverage for rare variants (MAF < 1%) compared to imputation results of genotype panel with GWAS chip alone. Regardless of samples size of reference panel, combined approach showed better imputation performance.

Also combined approach outperformed previously reported two-step imputation approach.

Besides the analysis strategy for enhancing imputation accuracy of rare variants, we develop a method to improve imputation performance, which is Pre-collapsing based imputation approach (PreCimp) is described in chapter 3. PreCimp method consists of two steps. In the first step, collapsed variables are generated using rare variants in the reference panel and new reference panel is constructed by inserting pre-collapsed variables (PCVs) into the reference panel. Next, typical imputation analysis with the new reference provides the imputed genotypes of collapsed variables. We demonstrated the performance of PreCimp on 5,349 genotyped samples using a Korean population specific reference panel including 848 samples of exome sequencing, Affymetrix 5.0, and exome chip. PreCimp outperformed a traditional post-collapsing method that collapses imputed variants after single rare variant imputation analysis. Although PreCimp poorly performed for genes larger than 200kb (about 3% of all genes), its performance would be improved by splitting large-sized genes into small sub-regions. PreCimp approach was shown to increase imputation accuracy about 3.4 ~ 6.3% (dosage r^2 0.6 ~ 0.8), 10.9 ~ 16.1% (dosage r^2 0.4 ~ 0.6), and 21.4 ~ 129.4% (dosage r^2 below 0.4) compared with the results of post-collapsing method.

With the proposed methods, we performed imputation based association analysis on liver enzymes. 8,529 samples were imputed using whole-exome reference panel. Following association analysis and meta-analysis on two cohort including KARE and HEXA samples revealed 20 loci at the p-value 5×10^{-6}. Among them, most loci were previously reported and 7 novel loci were discovered in this study. However, none of 7 new associations didn't reach the genome-wide significance level (5×10^{-8}). Novel loci included two missense variants and one of them located at *A1CF* that is known to be a modulator of liver regeneration. Despite the valuable of the findings, further replication study is warranted to confirm the genetic effect of discovered variants in an independent cohort.

In summary, we propose a combined approach for analysis strategy and develop PreCimp method to improve imputation accuracy of rare. Combined approach enhanced imputation accuracy about 11% and two times of genomic coverage for rare variants compared to previously used genotype panel consists of GWAS chip only. Pre-collapsing based imputation approach enhanced imputation accuracy of rare variants in forms of collapsed variables. PreCimp increased imputation accuracy about 10.9 ~ 129.4% for imputed variants with imputation quality score below 0.6. In the following imputation based association analysis, we performed imputation analysis using whole-exome sequencing data on genotyped samples comprising 8,529 samples. Subsequent association analysis discovered 7 novel loci including two missense variants. Our investigation of analysis strategy and methodological approach for enhancing imputation accuracy of rare variants, and following imputation based association study would be efficient analysis approaches and valuable resource for understanding rare variants and its association to various phenotypes.

References

Auer, P. L., J. M. Johnsen, A. D. Johnson, B. A. Logsdon, L. A. Lange, M. A. Nalls, G. Zhang, N. Franceschini, K. Fox, E. M. Lange, S. S. Rich, C. J. O'Donnell, R. D. Jackson, R. B. Wallace, Z. Chen, T. A. Graubert, J. G. Wilson, H. Tang, G. Lettre, A. P. Reiner, et al. (2012). "Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project." Am J Hum Genet **91**(5): 794-808.

Auer, P. L., A. Teumer, U. Schick, A. O'Shaughnessy, K. S. Lo, N. Chami, C. Carlson, S. de Denus, M. P. Dube, J. Haessler, R. D. Jackson, C. Kooperberg, L. P. Perreault, M. Nauck, U. Peters, J. D. Rioux, F. Schmidt, V. Turcot, U. Volker, H. Volzke, et al. "Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits." Nat Genet **46**(6): 629-34.

Bansal, V., O. Libiger, A. Torkamani and N. J. Schork (2010). "Statistical analysis strategies for association studies involving rare variants." Nat Rev Genet **11**(11): 773-85.

Bathum, L., H. C. Petersen, J. U. Rosholm, P. Hyltoft Petersen, J. Vaupel and K. Christensen (2001). "Evidence for a substantial genetic influence on biochemical liver function tests: results from a population-based Danish twin study." Clin Chem **47**(1): 81-7.

Blanc, V., K. J. Sessa, S. Kennedy, J. Luo and N. O. Davidson (2010). "Apobec-1 complementation factor modulates liver regeneration by post-transcriptional regulation of interleukin-6 mRNA stability." J Biol Chem **285**(25): 19184-92.

Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." Nat Genet **40**(6): 695-701.

Browning, S. R. and B. L. Browning "Haplotype phasing: existing methods and new developments." Nat Rev Genet **12**(10): 703-14.

Bush, W. S. and J. H. Moore (2012). "Chapter 11: Genome-wide association studies." PLoS Comput Biol **8**(12): e1002822.

Chambers, J. C., W. Zhang, J. Sehmi, X. Li, M. N. Wass, P. Van der Harst, H. Holm, S. Sanna, M. Kavousi, S. E. Baumeister, L. J. Coin, G. Deng, C. Gieger, N. L. Heard-Costa, J. J. Hottenga, B. Kuhnel, V. Kumar, V. Lagou, L. Liang, J. Luan, et al. (2011). "Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma." Nat Genet **43**(11): 1131-8.

Cho, Y. S., M. J. Go, Y. J. Kim, J. Y. Heo, J. H. Oh, H. J. Ban, D. Yoon, M. H. Lee, D. J. Kim, M. Park, S. H. Cha, J. W. Kim, B. G. Han, H. Min, Y. Ahn, M. S. Park, H. R. Han, H. Y. Jang, E. Y. Cho, J. E. Lee, et al. (2009). "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits." Nat Genet **41**(5): 527-34.

Cooke Bailey, J. N., N. D. Palmer, M. C. Ng, J. A. Bonomo, P. J. Hicks, J. M. Hester, C. D. Langefeld, B. I. Freedman and D. W. Bowden (2014). "Analysis of coding variants identified from exome sequencing resources for association with diabetic and non-diabetic nephropathy in African Americans." Hum Genet.

Cruchaga, C., C. M. Karch, S. C. Jin, B. A. Benitez, Y. Cai, R. Guerreiro, O. Harari, J. Norton, J. Budde, S. Bertelsen, A. T. Jeng, B. Cooper, T. Skorupa, D. Carrell, D. Levitch, S. Hsu, J. Choi, M. Ryten, J. Hardy, D. Trabzuni, et al.

(2014). "Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease." Nature **505**(7484): 550-4.

Deelen, P., A. Menelaou, E. M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, L. C. Francioli, J. J. Hottenga, L. C. Karssen, K. Estrada, E. Kreiner-Moller, F. Rivadeneira, J. van Setten, J. Gutierrez-Achury, H. J. Westra, L. Franke, D. van Enckevort, M. Dijkstra, H. Byelas, C. M. van Duijn, et al. (2014). "Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'." Eur J Hum Genet.

Delaneau, O., J. Marchini and J. F. Zagury (2012). "A linear complexity phasing method for thousands of genomes." Nat Methods **9**(2): 179-81.

Du, M., P. L. Auer, S. Jiao, J. Haessler, D. Altshuler, E. Boerwinkle, C. S. Carlson, C. L. Carty, Y. D. Chen, K. Curtis, N. Franceschini, L. Hsu, R. Jackson, L. A. Lange, G. Lettre, K. L. Monda, D. A. Nickerson, A. P. Reiner, S. S. Rich, S. A. Rosse, et al. "Whole-exome imputation of sequence variants identified two novel alleles associated with adult body height in African Americans." Hum Mol Genet.

Duan, Q., E. Y. Liu, P. L. Auer, G. Zhang, E. M. Lange, G. Jun, C. Bizon, S. Jiao, S. Buyske, N. Franceschini, C. S. Carlson, L. Hsu, A. P. Reiner, U. Peters, J. Haessler, K. Curtis, C. L. Wassel, J. G. Robinson, L. W. Martin, C. A. Haiman, et al. (2013). "Imputation of coding variants in African Americans: better performance using data from the exome sequencing project." Bioinformatics **29**(21): 2744-9.

Edwards, A. O., R. Ritter, 3rd, K. J. Abel, A. Manning, C. Panhuysen and L. A. Farrer (2005). "Complement factor H polymorphism and age-related macular degeneration." Science **308**(5720): 421-4.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz and C. I. Amos (2008). "Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms." Am J Hum Genet **82**(1): 100-12.

Haines, J. L., M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. A. Postel and M. A. Pericak-Vance (2005). "Complement factor H variant increases the risk of age-related macular degeneration." Science **308**(5720): 419-21.

Hao, K., E. Chudin, J. McElwee and E. E. Schadt (2009). "Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies." BMC Genet **10**: 27.

Hindorff, L., J. MacArthur, J. Morales, H. Junkins, P. Hall, A. Klemm and T. Manolio "A Catalog of Published Genome-Wide Association Studies. Available at:www.genome.gov/gwasudies. Accessed [6th Mar 2014]."

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-7.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Holmen, O. L., H. Zhang, Y. Fan, D. H. Hovelson, E. M. Schmidt, W. Zhou, Y. Guo, J. Zhang, A. Langhammer, M. L. Lochen, S. K. Ganesh, L. Vatten, F. Skorpen, H. Dalen, S. Pennathur, J. Chen, C. Platou, E. B. Mathiesen, T. Wilsgaard, I. Njolstad, et al. "Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk." Nat Genet **46**(4): 345-51.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini and G. R. Abecasis "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." Nat Genet **44**(8): 955-9.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini and G. R. Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." Nat Genet **44**(8): 955-9.

Howie, B., J. Marchini and M. Stephens (2011). "Genotype imputation with thousands of genomes." G3 (Bethesda) **1**(6): 457-70.

Huang, J., D. Ellinghaus, A. Franke, B. Howie and Y. Li (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data." Eur J Hum Genet **20**(7): 801-5.

Huyghe, J. R., A. U. Jackson, M. P. Fogarty, M. L. Buchkovich, A. Stancakova, H. M. Stringham, X. Sim, L. Yang, C. Fuchsberger, H. Cederberg, P. S. Chines, T. M. Teslovich, J. M. Romm, H. Ling, I. McMullen, R. Ingersoll, E. W. Pugh, K. F. Doheny, B. M. Neale, M. J. Daly, et al. "Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion." Nat Genet **45**(2): 197-201.

Joshi, P. K., J. Prendergast, R. M. Fraser, J. E. Huffman, V. Vitart, C. Hayward, R. McQuillan, D. Glodzik, O. Polasek, N. D. Hastie, I. Rudan, H. Campbell, A. F. Wright, C. S. Haley, J. F. Wilson and P. Navarro "Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies." PLoS One **8**(7): e68604.

Joshi, P. K., J. Prendergast, R. M. Fraser, J. E. Huffman, V. Vitart, C. Hayward, R. McQuillan, D. Glodzik, O. Polasek, N. D. Hastie, I. Rudan, H. Campbell, A. F. Wright, C. S. Haley, J. F. Wilson and P. Navarro (2013). "Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies." PLoS One **8**(7): e68604.

Kim, Y. J., M. J. Go, C. Hu, C. B. Hong, Y. K. Kim, J. Y. Lee, J. Y. Hwang, J. H. Oh, D. J. Kim, N. H. Kim, S. Kim, E. J. Hong, J. H. Kim, H. Min, Y. Kim, R. Zhang, W. Jia, Y. Okada, A. Takahashi, M. Kubo, et al. (2011). "Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits." Nat Genet **43**(10): 990-5.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable and J. Hoh (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-9.

Kreiner-Moller, E., C. Medina-Gomez, A. G. Uitterlinden, F. Rivadeneira and K. Estrada "Improving accuracy of rare variant imputation with a two-step imputation approach." Eur J Hum Genet.

Lander, E. S. (2011). "Initial impact of the sequencing of the human genome." Nature **470**(7333): 187-97.

Lange, L. A., Y. Hu, H. Zhang, C. Xue, E. M. Schmidt, Z. Z. Tang, C. Bizon, E. M. Lange, J. D. Smith, E. H. Turner, G. Jun, H. M. Kang, G. Peloso, P. Auer, K. P. Li, J. Flannick, J. Zhang, C. Fuchsberger, K. Gaulton, C. Lindgren, et al. (2014). "Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol." Am J Hum Genet **94**(2): 233-45.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segre, E. K. Speliotes, E. Wheeler, N. Soranzo, J. H. Park, J. Yang, D. Gudbjartsson, et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height." Nature **467**(7317): 832-8.

Lee, S., M. C. Wu and X. Lin (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-75.

Lee, T. H., W. R. Kim, J. T. Benson, T. M. Therneau and L. J. Melton, 3rd (2008). "Serum aminotransferase activity and mortality risk in a United States community." Hepatology **47**(3): 880-7.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-21.

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson and M. G. Ehm "Performance of genotype imputation for rare variants identified in exons and flanking regions of genes." PLoS One **6**(9): e24945.

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson and M. G. Ehm (2011). "Performance of genotype imputation for rare variants identified in exons and flanking regions of genes." PLoS One **6**(9): e24945.

Li, Y., C. J. Willer, J. Ding, P. Scheet and G. R. Abecasis "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genet Epidemiol **34**(8): 816-34.

Lin, J. P., C. J. O'Donnell, C. S. Fox and L. A. Cupples (2009). "Heritability of serum gamma-glutamyltransferase level: genetic analysis from the Framingham Offspring Study." Liver Int **29**(5): 776-7.

Loomba, R., F. Rao, L. Zhang, S. Khandrika, M. G. Ziegler, D. A. Brenner and D. T. O'Connor (2010). "Genetic covariance between gamma-glutamyl transpeptidase and fatty liver risk factors: role of beta2-adrenergic receptor genetic variation in twins." Gastroenterology **139**(3): 836-45, 845 e1.

Magi, R., J. L. Asimit, A. G. Day-Williams, E. Zeggini and A. P. Morris "Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases." Genet Epidemiol.

Magi, R., J. L. Asimit, A. G. Day-Williams, E. Zeggini and A. P. Morris (2012). "Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases." Genet Epidemiol.

Makkonen, J., K. H. Pietilainen, A. Rissanen, J. Kaprio and H. Yki-Jarvinen (2009). "Genetic factors contribute to variation in serum alanine aminotransferase activity independent of obesity and alcohol: a study in monozygotic and dizygotic twins." J Hepatol **50**(5): 1035-42.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-53.

Marchini, J. and B. Howie "Genotype imputation for genome-wide association studies." Nat Rev Genet **11**(7): 499-511.

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

Martin, A. R., G. Tse, C. D. Bustamante and E. E. Kenny "Imputation-based assessment of next generation rare exome variant arrays." Pac Symp Biocomput: 241-52.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010).

"The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-303.

Morrison, A. C., A. Voorman, A. D. Johnson, X. Liu, J. Yu, A. Li, D. Muzny, F. Yu, K. Rice, C. Zhu, J. Bis, G. Heiss, C. J. O'Donnell, B. M. Psaty, L. A. Cupples, R. Gibbs and E. Boerwinkle (2013). "Whole-genome sequence-based analysis of high-density lipoprotein cholesterol." Nat Genet **45**(8): 899-901.

Nelson, S. C., K. F. Doheny, E. W. Pugh, J. M. Romm, H. Ling, C. A. Laurie, S. R. Browning, B. S. Weir and C. C. Laurie "Imputation-based genomic coverage assessments of current human genotyping arrays." G3 (Bethesda) **3**(10): 1795-807.

Nilsson, S. E., S. Read, S. Berg and B. Johansson (2009). "Heritabilities for fifteen routine biochemical values: findings in 215 Swedish twin pairs 82 years of age or older." Scand J Clin Lab Invest **69**(5): 562-9.

Oliveira, M. N., J. P. Hemerly, A. U. Bastos, R. Tamanaha, F. R. Latini, C. P. Camacho, A. Impellizzeri, R. M. Maciel and J. M. Cerutti (2011). "The RET p.G533C mutation confers predisposition to multiple endocrine neoplasia type 2A in a Brazilian kindred and is able to induce a malignant phenotype in vitro and in vivo." Thyroid **21**(9): 975-85.

Pasaniuc, B., N. Rohland, P. J. McLaren, K. Garimella, N. Zaitlen, H. Li, N. Gupta, B. M. Neale, M. J. Daly, P. Sklar, P. F. Sullivan, S. Bergen, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, S. M. Purcell, D. W. Haas, L. Liang, S. Sunyaev, et al. (2012). "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." Nat Genet **44**(6): 631-5.

Pei, Y. F., J. Li, L. Zhang, C. J. Papasian and H. W. Deng (2008). "Analyses and comparison of accuracy of different genotype imputation methods." PLoS One **3**(10): e3551.

Peloso, G. M., P. L. Auer, J. C. Bis, A. Voorman, A. C. Morrison, N. O. Stitziel, J. A. Brody, S. A. Khetarpal, J. R. Crosby, M. Fornage, A. Isaacs, J. Jakobsdottir, M. F. Feitosa, G. Davies, J. E. Huffman, A. Manichaikul, B. Davis, K. Lohman, A. Y. Joon, A. V. Smith, et al. "Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks." Am J Hum Genet **94**(2): 223-32.

Pilia, G., W. M. Chen, A. Scuteri, M. Orru, G. Albai, M. Dei, S. Lai, G. Usala, M. Lai, P. Loi, C. Mameli, L. Vacca, M. Deiana, N. Olla, M. Masala, A. Cao, S. S. Najjar, A. Terracciano, T. Nedorezov, A. Sharov, et al. (2006). "Heritability of cardiovascular and personality traits in 6,148 Sardinians." PLoS Genet **2**(8): e132.

Pratt, D. S. and M. M. Kaplan (2000). "Evaluation of abnormal liver-enzyme results in asymptomatic patients." N Engl J Med **342**(17): 1266-71.

Purcell, S. M., J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kahler, L. Duncan, E. Stahl, G. Genovese, E. Fernandez, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. Magnusson, E. Banks, K. Shakir, et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia." Nature **506**(7487): 185-90.

Rahmioglu, N., T. Andrew, L. Cherkas, G. Surdulescu, R. Swaminathan, T. Spector and K. R. Ahmadi (2009). "Epidemiology and genetic epidemiology of the liver function test proteins." PLoS One **4**(2): e4435.

Schindhelm, R. K., M. Diamant, J. M. Dekker, M. E. Tushuizen, T. Teerlink and R. J. Heine (2006). "Alanine aminotransferase as a marker of non-alcoholic fatty liver disease in relation to type 2 diabetes mellitus and cardiovascular disease." Diabetes Metab Res Rev **22**(6): 437-43.

Scott, R. A., V. Lagou, R. P. Welch, E. Wheeler, M. E. Montasser, J. Luan, R. Magi, R. J. Strawbridge, E. Rehnberg, S. Gustafsson, S. Kanoni, L. J. Rasmussen-Torvik, L. Yengo, C. Lecoeur, D. Shungin, S. Sanna, C. Sidore, P. C. Johnson, J. W. Jukema, T. Johnson, et al. "Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways." Nat Genet **44**(9): 991-1005.

Sung, J., K. Lee and Y. M. Song (2010). "Heritabilities of Alcohol Use Disorders Identification Test (AUDIT) scores and alcohol biomarkers in Koreans: the KoGES (Korean Genome Epi Study) and Healthy Twin Study." Drug Alcohol Depend **113**(2-3): 104-9.

Sung, Y. J., L. Wang, T. Rankinen, C. Bouchard and D. C. Rao (2012). "Performance of genotype imputations using data from the 1000 Genomes Project." Hum Hered **73**(1): 18-25.

Sveinbjornsson, G., E. Mikaelsdottir, R. Palsson, O. S. Indridason, H. Holm, A. Jonasdottir, A. Helgason, S. Sigurdsson, A. Sigurdsson, G. I. Eyjolfsson, O. Sigurdardottir, O. T. Magnusson, A. Kong, G. Masson, P. Sulem, I. Olafsson, U. Thorsteinsdottir, D. F. Gudbjartsson and K. Stefansson "Rare mutations associating with serum creatinine and chronic kidney disease." Hum Mol Genet.

Taliun, D., J. Gamper and C. Pattaro (2014). "Efficient haplotype block recognition of very long and dense genetic sequences." BMC Bioinformatics **15**: 10.

Targher, G., M. Chonchol, L. Miele, G. Zoppini, I. Pichiri and M. Muggeo (2009). "Nonalcoholic fatty liver disease as a contributor to hypercoagulation and thrombophilia in the metabolic syndrome." Semin Thromb Hemost **35**(3): 277-87.

Thompson, J. R., J. Attia and C. Minelli (2011). "The meta-analysis of genome-wide association studies." Brief Bioinform **12**(3): 259-69.

van Beek, J. H., M. H. de Moor, E. J. de Geus, G. H. Lubke, J. M. Vink, G. Willemsen and D. I. Boomsma (2013). "The genetic architecture of liver enzyme levels: GGT, ALT and AST." Behav Genet **43**(4): 329-39.

Vernon, G., A. Baranova and Z. M. Younossi (2011). "Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults." Aliment Pharmacol Ther **34**(3): 274-85.

Visscher, P. M., S. E. Medland, M. A. Ferreira, K. I. Morley, G. Zhu, B. K. Cornes, G. W. Montgomery and N. G. Martin (2006). "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." PLoS Genet **2**(3): e41.

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff and H. Parkinson (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic Acids Res **42**(Database issue): D1001-6.

Whitfield, J. B. and N. G. Martin (1985). "Individual differences in plasma ALT, AST and GGT: contributions of genetic and environmental factors, including alcohol consumption." Enzyme **33**(2): 61-9.

Whitfield, J. B., G. Zhu, J. E. Nestler, A. C. Heath and N. G. Martin (2002). "Genetic covariation between serum gamma-glutamyltransferase activity and cardiovascular risk factors." Clin Chem **48**(9): 1426-31.

Willer, C. J., Y. Li and G. R. Abecasis (2010). "METAL: fast and efficient meta-analysis of genomewide association scans." Bioinformatics **26**(17): 2190-1.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke and X. Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.

Yuan, X., D. Waterworth, J. R. Perry, N. Lim, K. Song, J. C. Chambers, W. Zhang, P. Vollenweider, H. Stirnadel, T. Johnson, S. Bergmann, N. D. Beckmann, Y. Li, L. Ferrucci, D. Melzer, D. Hernandez, A. Singleton, J. Scott, P. Elliott, G. Waeber, et al. (2008). "Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes." Am J Hum Genet **83**(4): 520-8.

Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev and E. S. Lander "Searching for missing heritability: designing rare variant association studies." Proc Natl Acad Sci U S A **111**(4): E455-64.

Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev and E. S. Lander (2014). "Searching for missing heritability: Designing rare variant association studies." Proc Natl Acad Sci U S A **111**(4): E455-64.

초 록

희귀변이는 잃어버린 유전성을 설명할 수 있을 것으로 기대되는 주요 유전 요인 중 하나로써 많은 관심을 받고 있다. 최근 급성장한 차세대염기서열분석 기법으로 인해 희귀변이의 발굴과 분석이 가능하게 되었다. 이처럼 차세대염기서열 분석이 유전체 연구에 강력한 연구기법으로 활용되고 있으나, 실험을 위한 비용과 분석에 필요한 높은 계산력의 문제로 대규모 인구집단 기반 유전체 연구에 활용하기에는 아직 어려운 실정이다. 그 대안으로, 최근에는 사용자화칩(메타보칩, 엑솜칩 등)과 임퓨테이션 기법이 대규모 인구집단 기반 유전체 연구에 널리 활용되고 있다. 임퓨테이션 기법은 컴퓨터 계산을 통한 예측 분석으로 희귀변이 정보를 얻을 수 있다. 그러나, 임퓨테이션을 통해 얻은 희귀변이의 정확도가 낮다는 문제점이 있다. 또한, 낮은 정확도의 임퓨테이션 결과를 이용하여 지역 기반 연관성 분석을 수행하게 된다면 위양성 결과가 발생할 가능성이 있다. 사용자화칩의 경우 희귀변이가 기본적으로 포함되도록 설계되어 있으나 대부분 특수한 목적으로 설계되었다는 점에서 한계를 가지고 있다. 따라서, 희귀변이 정보를 얻기 위한 새로운 분석 전략과 방법에 대한 요구가 증대되고 있다.

첫번째로, 본 연구에서는 통합 정보를 이용하는 방법에 대한 분석 전략을 수립하였다. 이 방법은 전장유전체칩과 엑솜칩을 통합하고 임퓨테이션 분석을 진행하는 것이다. 이를 위해 848명의 동일한 샘플에 대해 생산된 엑솜염기서열정보, 전장유전체칩, 엑솜칩 정보를 이용하여 참조패널을 구축하였다. 실제 임퓨테이션 분석의 대상이 되는 유전형

패널의 경우 5,349명의 동일한 샘플에 대해 엑솜칩으로만 구성된 패널, 전장유전체칩으로만 구성된 패널, 통합된 패널을 이용하여 임퓨테이션 시 나타나는 정확도 변화를 분석하였다. 그 결과로 희귀변이 연구에서 통합정보 패널을 이용하는 경우에 전장유전체칩으로만 구성된 패널을 사용하는 경우보다 약 11%의 정확도 향상과 두 배의 유전체연구범위가 향상되는 것을 관찰하였다. 참조패널의 샘플 수에 관계없이 통합패널이 항상 더 좋은 결과를 보여주었다. 또한, 통합패널을 이용한 방법은 기존에 소개된 두 단계 임퓨테이션 방법보다 더 높은 정확도를 보여주었다.

본 연구에서는 전략적 분석 방법 이외에도 분석 방법을 개발하여 높은 정확도의 희귀변이 정보를 얻고자 하였다. 선병합 방법을 통한 임퓨테이션을 이용하여 병합된 변수에 대한 정확도를 향상 시키고자 하였다. 선병합 임퓨테이션 방법은 두 단계로 구성되어있다. 첫째로 참조패널의 정보를 이용하여 병합된 정보를 생산하고 기존 참조패널에 추가함으로써 새로운 참조패널을 생성한다. 다음으로 새로 생성된 참조패널 정보를 이용하여 일반적인 임퓨테이션 분석을 수행하여 병합된 정보에 대한 예측 값을 얻을 수 있다. 본 연구에서는 848명의 동일인을 대상으로 생산된 엑솜염기서열정보, 전장유전체칩 정보, 엑솜칩 정보를 활용하여 엑솜참조패널을 구축하였다. 구축된 참조패널은 선병합 방법을 이용하여 새로운 패널을 구성하였고, 이 패널을 이용하여 5,349명의 전장유전체칩을 임퓨테이션 분석하였다. 분석된 결과는 동일한 5,349명을 대상으로 생산된 엑솜칩 정보와 비교하여 정확도를 측정하였다. 그 결과로 선병합 방법은 기존에 사용되었던 임퓨테이션 후 병합하는

방법보다 더 좋은 결과를 보여주었다. 약 3%에 해당하는 크기가 200kb 이상의 유전자에서는 좋지 않은 결과를 보여주었으나, 작은 단위로 나눠서 선병합 방법을 적용하는 경우 다른 결과와 마찬가지로 정확도 향상을 관찰 할 수 있었다. 선병합 방법은 임퓨테이션 후 병합하는 방법에 비해 약 3.4 ~ 6.3% (dosage r^2 0.6 ~ 0.8), 10.9 ~ 16.1% (dosage r^2 0.4 ~ 0.6), 21.4 ~ 129.4% (dosage r^2 0.4 이하)의 정확도 향상을 보여주었다.

마지막으로 본 연구에서는 상기 개발된 분석 전략과 방법을 이용하여 임퓨테이션 기반 연관성 분석을 수행하여 간 효소에 연관된 유전요인을 발굴하고자 하였다. 먼저 엑솜염기서열 정보를 포함하여 엑솜 참조패널을 구성하였으며, 이를 총 8,529명에서 생산된 전장유전체칩과 엑솜칩을 통합한 정보의 임퓨테이션 분석에 활용하였다. 임퓨테이션 후 연관성 분석을 수행하여 간 효소에 연관된 20개의 유전자좌를 발굴하였다 (유의확률 < $5x10^{-6}$). 발굴된 20개의 유전자좌 중, 7개는 본 연구에서 새롭게 발굴된 것이며 2개의 변이는 단백질 형성에 영향을 주는 것으로 알려져 있었다.

본 연구에서는 희귀변이 연구를 위해 희귀변이 임퓨테이션 시 정확도 향상을 위한 분석 전략과 방법론을 개발하였으며, 이를 연관성 분석에 활용하여 간 효소에 연관된 새로운 7개의 유전변이를 발굴하였다. 본 연구에서 개발된 효율적으로 높은 정확도의 희귀변이 정보를 얻을 수 있는 방법과 간 효소에 연관된 유전변이 신규 발굴 정보는 희귀변이 연구와 그것이 표현형에 미치는 연구에 널리 활용될 수 있는 것으로 기대된다.

주요어: 단일염기다형성, 유전변이, 희귀변이, 전장유전체연관성분석, 연관성 분석
학 번: 2009-30105