Published in final edited form as:

Genet Epidemiol. 2017 January ; 41(1): 41–50. doi:10.1002/gepi.22020.

PreCimp: Pre-collapsing Imputation approach increases imputation accuracy of rare variants in terms of collapsed variables

Young Jin Kim^{1,2}, Juyoung Lee², Bong-Jo Kim², T2D-Genes Consortium, and Taesung Park^{1,3,*}

¹Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, 151-742, Korea

²Division of Structural and Functional Genomics, Center for Genome Science, Korean National Institute of Health, Osong, Chungchungbuk-do, 363-951, Korea

³Department of Statistics, Seoul National University, Seoul, 151-742, Korea

Abstract

Imputation is widely used for obtaining information about rare variants. However, one issue concerning imputation is the low accuracy of imputed rare variants as the inaccurate imputed rare variants may distort the results of region-based association tests. Therefore, we developed a pre-collapsing imputation method (PreCimp) to improve the accuracy of imputation by using collapsed variables. Briefly, collapsed variables are generated using rare variants in the reference panel, and a new reference panel is constructed by inserting pre-collapsed variables into the original reference panel. Following imputation analysis provides the imputed genotypes of the 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. Compared with the results of post-collapsing method, PreCimp approach was shown to relatively increase imputation accuracy about 3.4 – 6.3% when dosage r^2 is between 0.6 and 0.8, 10.9 – 16.1% when dosage r^2 is between 0.4 and 0.6, and 21.4 ~ 129.4% when dosage r^2 is below 0.4.

Keywords

Next Generation Sequencing; Imputation; SNPs; genotyping; Population genetics

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 May 2015, 15,238 single nucleotide polymorphisms (SNPs) from 2,175

*Corresponding author: Phone: 82-2-880-8924, tspark@stats.snu.ac.kr. Address: Department of Statistics, Seoul National University San 56-1, Shilim-dong, Kwanak-gu, Seoul, 151-742, Korea.

published GWASs are included in the National Human Genome Research Institute GWAS catalogue, a curated resource of SNP-trait associations [Hindorff, et al. 2015; 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 [Lander 2011; Manolio, et al. 2009]. 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 [Cooke Bailey, et al. 2014; Cruchaga, et al. 2014; Lange, et al. 2014; Morrison, et al. 2013; 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 a large 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 [Auer, et al. 2012; Li, et al. 2011a]. 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. Generally, imputation uses estimated haplotype segments from reference panel using the overlapped variants between reference panel and genotype panel. Commonly used GWAS chips mainly contain common variants. When the genotype panel using GWAS chip is used for imputation analysis, poor performance of rare variant imputation would be caused by inaccurate estimation of rare haplotypes that are not well tagged by nearby common variants. 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. 2011a], or (2) using chips designed to tag rare variants and population-specific variants [Joshi, et al. 2013; Li, et al. 2011a]. 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.

Common tagging variants of a GWAS chip may not tag haplotypes carrying rare alleles well. However, common variants would possibly tag collapsed variables better. The collapsing method summarizes multiple rare variants within a region into a single variable [Bansal, et al. 2010]. The collapsed variable is generally binary, coded as 1 if an individual has one or more rare alleles and 0 otherwise. Minor allele (coded as 1) frequency of the collapsed variable is defined as the sum of frequencies of rare alleles [Li and Leal 2008] and is much higher than the frequency of each rare allele. Therefore, minor allele of collapsed variables may be more accurately estimated via imputation analysis, if haplotypes carrying rare alleles share common haplotype segments that are well tagged by common variants of a GWAS chip.

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 1). The proposed method uses variants from a phased reference panel to make collapsed variables and then inserts these pre-collapsed variables (PreCs) into the original reference panel to make a new reference panel. Typical imputation with the new reference panel can impute PreCs 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 I) [Cho, et al. 2009]. Then, imputation accuracy was assessed by comparing imputed PreCs and collapsed variables from exome array data of identical 5,349 samples.

Methods

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.

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 a study specific reference panel.

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 (Affymetrix website, <http://www.affymetrix.com>), SNPs 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 the genotype panel and the true dataset, respectively (Table I).

Building reference panels

We then constructed a study-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. The numbers 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 minor allele count < 5 were excluded, 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]. 1,000 Genomes project phase 3 v5 reference panel data were downloaded from MaCH website. Prior to imputation analysis, we selected chromosome 1 data of 504 samples with East Asian ancestry. After removing extremely rare variants (minor allele count < 5), 811,572 variants were remained for further analysis.

Pre-collapsing and post-collapsing based imputation

The collapsing method is an approach that collapses rare variants within a region [Li and Leal 2008]. Figure 1 shows schematic representation of post-collapsing (PostC) and pre-collapsing imputation (PreCimp) methods in a side by side manner. For imputed rare variants, we defined PostC and PreCimp methods as follows.

PostC method is an approach that is typically used in region-based association studies using imputation data. First, imputation analysis was performed on the genotype panel of GWAS chip to impute rare variants. For further analysis, we used best-guess genotypes that is the imputed genotypes with maximum posterior probability [Marchini and Howie 2010]. Then, a collapsed variable X of imputed rare variants for the i^{th} individual is defined as

$$x_i = \begin{cases} 1 & \text{if the number of rare allele} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

The PreCimp method is an approach that collapses rare variants in a reference panel and generates a new reference panel by inserting these pre-collapsed variables (PreCs) 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

$$x_{ij} = \begin{cases} 1 & \text{if the number of rare allele} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, typical pre-phasing based imputation with the new reference panel was performed.

Since PreCs are artificially generated, these new markers need to be assigned to specific chromosomal positions in order to be incorporated into the reference panel. For the

comparison analysis, we used four 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), and the mean position of variants used for PreC (PreCimp-M). We only used mean position for PreCs to compare imputation performances between PostC and PreCimp, and for PreCimp methods when additional information was used.

In this study, pre-phasing-based imputation was performed for rare variants imputation [Howie, et al. 2012]. Imputation analysis was performed using minimac software [Howie, et al. 2012]

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 array 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.

Statistical analysis

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.

Results

PostC vs. PreCimp methods

We performed a comparison analysis of the imputation performances of PostC and PreCimp methods. For the purpose of comparison, we used only variants in an exome array and treated the exome array data as true data. For PostC, we used the study specific exome reference panel to impute the genotype panel of 5,349 samples genotyped with Affymetrix 5.0. Imputed genotypes with maximum posterior probability were used for further analysis. Then, collapsed variables were generated by collapsing imputed rare variants. For PreCimp, pre-collapsed variables were generated by collapsing rare variants of the study specific reference panel and new reference panel was constructed by inserting pre-collapsed variables into the original reference panel. Then, we used the new reference panel to impute the genotype panel of 5,349 samples genotyped with Affymetrix 5.0. To measure imputation accuracy, we used dosage r^2 by calculating squared Pearson correlation between collapsed variables of PreCimp and true data, collapsed variables of PostC and true data.

Two sets of collapsed variants were used, MAF1 (collapsing variants with MAF $\geq 1\%$) and MAF5 (collapsing variants with MAF $\geq 5\%$). 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 2. Figure 2A shows that imputation performance was enhanced by the PreCimp method. The proposed approach relatively 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 II). A Wilcoxon signed-rank test performed to test the statistical significance of difference in imputation performance showed that the PreCimp method significantly outperformed the PostC method ($p\text{-value} < 2.2 \times 10^{-16}$).

For rare variants, imputation performance depends on the frequency of haplotypes carrying rare alleles [Li, et al. 2011a]. Improved accuracy by PreCimp is produced by the increased frequency of haplotypes carrying minor allele of collapsed variable. For example, *ABCA10* of MAF5 set consists of six variants with allele frequencies ranging from 0.3% to 2.6%. For *ABCA10*, dosage r^2 of PostC and PreCimp was 0.24 and 0.36, respectively. Within a 100kb window of *ABCA10* region, common variants (MAF $\geq 5\%$), six rare variants, and the collapsed variable of six variants were selected to construct haplotypes using Haploview v4.2. The frequencies of haplotypes carrying rare alleles of six variants ranged from 0.2% to 1.4%. However, the frequency of the haplotype carrying minor allele of collapsed variable (sets of rare alleles) was 4.7%. Therefore, PreCimp would improve imputation performance by increasing the frequency of haplotype with minor allele of collapsed variable.

The difference in dosage r^2 using the PreCimp and PostC methods are shown in Figure 2B (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 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 $< 200\text{kb}$ are shown in Figure 3A, and genes $\geq 200\text{kb}$ are shown in Figure 3B.

Gene length was a major factor affecting the imputation performance of the PreCimp method. For large genes (about $>200\text{kb}$, 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 subregions 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.

Application to 1,000 Genomes project data

We also applied PreCimp to 1,000 Genomes project data that is widely used as public reference panel. 1,000 Genomes project phase 3 v5 reference panel data was downloaded from MaCH website (<http://csg.sph.umich.edu/abecasis/MACH/>). To simulate dataset for the analysis of the imputation performance, we selected chromosome 1 data and applied PreCimp using variants overlapped with an exome array that are treated as true data. Since rare variants tend to be population specific [Bodmer and Bonilla 2008], multi-ethnic samples of 1,000 Genomes project would not share haplotype background for rare variants.

Therefore, we only selected 504 samples with East Asian ancestry. Variants were collapsed for each gene if MAF $\leq 5\%$. There were 279 genes consisting of 660 variants. For comparison purpose, we also performed PostC analysis using imputation data based on the same reference panel. To measure imputation accuracy, we used dosage r^2 by calculating squared Pearson correlation between collapsed variables of PreCimp and true data, and collapsed variables of PostC and true data. As expected, PreCimp increased imputation accuracy about 3.8 – 7.6% (dosage r^2 0.6 – 0.8), 11.7 – 14.9% (dosage r^2 0.4 – 0.6), and 29.1 – 425.2% (dosage r^2 below 0.4) compared with the results of PostC method. Wilcoxon signed-rank test showed that the PreCimp method significantly outperformed the PostC method (p-value $< 2.2 \times 10^{-16}$).

The results showed that PreCimp can be applied to study specific reference panel and also the public reference panel of 1,000 Genomes project. Moreover, the proposed method improved imputation performance of rare variants compared to PostC method regardless of reference panels used for imputation analysis.

Although PreCimp improved imputation performance of rare variants using 1,000 Genomes project data, 1,000 Genomes data would not be as good as study specific sequencing data as a source of rare variants. For example, chromosome 1 of study specific reference panel had 422 genes when collapsing variants with MAF $\leq 5\%$ while only 279 genes were available for 1,000 Genomes project phase 3 reference panel. Among them, only 175 genes had the same number of variants used for pre-collapsing in both study specific reference panel and 1,000 Genomes project data. If 1,000 Genomes project data was used as reference panel, mean dosage r^2 of the 175 genes were 0.561 and 0.495 for PreCimp and PostC, respectively. Study specific reference panel showed mean dosage r^2 as 0.690 and 0.646 for PreCimp, PostC, respectively. The study specific reference panel had more rare variant information and showed better imputation performance.

Difference in rare variant information and imputation performance between reference panels would be caused by two factors. The first factor is a sample size. Note that there were only 504 samples with East Asian ancestry in 1,000 Genomes Project phase 3 dataset. The sample size 504 is much smaller than that of 848 of the study specific reference panel used in this study. It is previously reported that the more sample size of the reference panel provides, the more rare variants [Li and Leal 2009] and the better imputation performance [Li, et al. 2011a]. The second factor is the reference panel. The study specific reference panel has more shared haplotype segments with the genotype panel of the study samples [Duan, et al. 2013]. Previously, the imputation approach using the study specific reference panel was also shown to provide better imputation quality than that using 1,000 Genomes data as a reference panel [Duan, et al. 2013]. Therefore, PreCimp would perform best if the study specific reference panel is available.

PreCimp with additional information

If the genotype panel includes rare variants, imputation accuracy of rare variants can be increased [Joshi, et al. 2013; Li, et al. 2011a]. In this study, PreCimp uses PreCs for imputing collapsed variables. Since PreCs were generated by collapsing rare variants in the reference panel, rare variants used for PreCs are more likely to correlate with PreCs than

nearby common variants. Therefore, PreCimp would perform better if one or more rare variants used for PreCimp were available in both the reference and genotype panels. In this context, we next 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 simulate the genotype panel containing rare variants, rare variants with the highest LD r^2 with PreCs were added to the existing genotype panel. Figure 4 shows the mean dosage r^2 values of MAF5 set obtained by PreCimp without additional information, and PostC (PostC+) and PreCimp (PreCimp+) when additional information was used for imputation. The results show that the imputation performance of PreCimp and PostC was greatly improved when an additional variant was added to the genotype panel. Mean dosage r^2 was 0.602, 0.931, and 0.934 for PreCimp, PostC+, and PreCimp+, respectively. As expected, PreCimp+ outperformed PostC+. PreCimp+ approach relatively increased imputation accuracy about 4.7% (dosage r^2 0.6 – 0.8), 11.2% (dosage r^2 0.4 – 0.6), and 8.5% (dosage r^2 below 0.4) compared with the results of PostC+. A Wilcoxon signed-rank test showed that the PreCimp+ significantly outperformed the PostC+ method (p-value < 2.2×10^{-16}).

Effect of PreC position on imputation performance

PreC is an artificial value and has no specific genomic position. Thus, the position of PreCs should be assigned arbitrarily. Since the imputation method predicts untyped markers based on haplotype patterns consisting of sets of correlated variants, the position of PreCs could affect imputation performance. For the comparison analysis, we used four 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), and the mean position of variants used for PreC (PreCimp-M).

For PreCimp, all PreCimp methods showed similar imputation performances. There was no statistically significant difference among the positions of PreCimp method. This result indicates that the position of PreCs does not affect the estimates of haplotype segments using nearby common variants. For PreCimp+, PreCimp-R2 showed the best performance. Figure 5 shows mean dosage r^2 values obtained using the different PreCimp methods (MAF5 set). PreCimp-R2 showed an improved performance over the four PreCimp methods and significantly outperformed the other PreCimp methods (Wilcoxon signed rank test p-values < 6.17×10^{-8} , for MAF1 and MAF5). PreCimp-R2 puts PreCs near the highly correlated rare variants, which may increase the probability that PreCs and the highly correlated rare variants are included in the same estimated haplotype segments. Thus, the more accurate imputation of PreCimp-R2 would be caused by better estimation of haplotypes containing PreCs.

Discussion

In this study, we proposed the 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 accuracy of collapsed variables. For example, the imputation accuracy of genes with low dosage r^2 (<

0.6) of PostC method was improved by approximately 10.9 – 129.4% (Table II). Moreover, the performance was greatly improved when 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 PreC position on imputation performance. Our results show that, if additional rare variants are added to the existing genotype panel, imputation performance is increased by placing PreCs next to the added variants with the highest LD.

In addition to enhancing imputation accuracy, PreCimp enables us to use more collapsed variables after filtering out low quality imputation variants. For example, the imputed variants of low quality with estimated $r^2 < 0.3$ are often removed after imputation analysis [Li, et al. 2011b]. The estimated r^2 is an imputation quality metric provided by MACH and minimac software [Li, et al. 2011b]. After filtering out imputed variants of low quality, PreCimp retained 90.5% of imputed collapsed variables (3,487 of 3,830 genes, MAF5 set) while 71.5% (2,757 of 3,830 genes, MAF5 set) of collapsed variables were remained for PostC. PreCimp not only enhances imputation accuracy but also retains 19% more collapsed variables than PostC.

Improved imputation accuracy by PreCimp would lead to an increase in statistical power of an association study. Reduced imputation error provides an increase in statistical power [Huang, et al. 2009]. Huang et al. reported that 5% ~ 13% more samples are required for the 1% increase of imputation error [Huang, et al. 2009]. Also Liu et al. showed that 0.1 increase in dosage r^2 improves more than 5% in statistical power (Figure 2 of Liu et al. 2013, assuming MAF=0.1, odds-ratio=1.28, 1,000 cases and 1,000 controls) [Liu, et al. 2013]. In this study, there were 465 genes (12% of MAF5 set, 3,830 genes) with more than 0.1 increase in dosage r^2 . Based on the results by Liu et al., one can expect to gain more than 5% of statistical power for about 12% of gene sets by using PreCimp. However, statistical power varies depending on minor allele frequency, genetic effect size, and sample size [Liu, et al. 2013]. Therefore, a further study is warranted for a thorough investigation of statistical power gain by PreCimp in various conditions.

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 does not require an intensive computing process such as haplotype estimation of reference panel. In this study, we showed that PreCimp can be applied to imputation by using a specific reference panel and also public reference panels. In addition, a coding scheme utilizing the PreC method would make it possible to identify disease-associated rare variants on the basis of haplotype. During PreC, rare variants are collapsed by each haplotype, and PreCs can be coded as 0, 1, or 2 depending on the number of haplotypes with rare variants. PreCimp can also be used for pathway based association tests using imputed data in combination with PostC method. Since PreCimp mainly utilizes LD and haplotype information, PreCimp cannot be directly applied to estimating a pathway-level pre-collapsed

variable consisting of numerous genes with different LD structures. Instead, PreCimp can provide an imputed pre-collapsed variable for each gene and then PostC method can be used for collapsing the imputed pre-collapsed variables, representing genes in a pathway, into a single pathway-level variable.

Despite these advantages, however, the proposed PreCimp method has some limitations. First, PreCimp showed poor performance with large genes (>200kb, Table I). Generally, the distance between two variants is negatively correlated with LD. In addition, 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. However, the performance of PreCimp can be improved by splitting large genes into several small sub-regions. Since genes sized 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 be incorporated with the proposed method. Thus, the proposed method will have to be extended in order to consider various aspects of rare variants in association analyses.

Recently, various studies have proposed rare variant association tests using imputed genotypes [Li, et al. 2010; Zawistowski, et al. 2010]. Our proposed method mainly focused on increasing the accuracy of imputed collapsed variables. By combining PreCimp method and previously introduced association methods for imputed genotypes, association analysis for rare variants would be more powerful for identifying associated rare variants.

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.

Acknowledgments

The authors thank Goncalo Abecasis for helpful comments. This work was supported by the Bio & Medical Technology Development Program of the National Research Foundation of Korea (NRF) grant (2013M3A9C4078158), by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI15C2165), and an intramural grant from the Korea National Institute of Health (2014-NI73001-00), the Republic of Korea. This study was provided with data from the Korean Genome Analysis Project (4845-301), the Korean Genome and Epidemiology Study (4851-302), and Korea Biobank Project (4851-307, KBP-2013-000) that were supported by the Korea Center for Disease Control and Prevention, Republic of Korea. Sequencing data from the T2D-GENES Consortium was supported by NIH/NIDDK U01's DK085501, DK085524, DK085526, DK085545 and DK085584.

References

- Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, Nalls MA, Zhang G, Franceschini N, Fox K, Lange EM, et al. 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.* 2012; 91(5):794–808. [PubMed: 23103231]
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010; 11(11):773–85. [PubMed: 20940738]
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008; 40(6):695–701. [PubMed: 18509313]
- Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.* 2012; 8(12):e1002822. [PubMed: 23300413]
- Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, Yoon D, Lee MH, Kim DJ, Park M, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet.* 2009; 41(5):527–34. [PubMed: 19396169]
- Cooke Bailey JN, Palmer ND, Ng MC, Bonomo JA, Hicks PJ, Hester JM, Langefeld CD, Freedman BI, Bowden DW. Analysis of coding variants identified from exome sequencing resources for association with diabetic and non-diabetic nephropathy in African Americans. *Hum Genet.* 2014
- Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature.* 2014; 505(7484):550–4. [PubMed: 24336208]
- Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2012; 9(2):179–81.
- Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, Jun G, Bizon C, Jiao S, Buyske S, Franceschini N, et al. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics.* 2013; 29(21):2744–9. [PubMed: 23956302]
- Hindorf L, MacArthur J, Morales J, Junkins H, Hall P, Klemm A, Manolio T. [Accessed [May 2015]] A Catalog of Published Genome-Wide Association Studies. 2015. Available at: www.genome.gov/gwasudies
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 2009; 106(23):9362–7. [PubMed: 19474294]
- Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet.* 2012; 44(8):955–9. [PubMed: 22820512]
- Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda).* 2011; 1(6):457–70. [PubMed: 22384356]
- Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet.* 2012; 20(7):801–5. [PubMed: 22293688]
- Huang L, Wang C, Rosenberg NA. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am J Hum Genet.* 2009; 85(5):692–8. [PubMed: 19853241]
- Joshi PK, Prendergast J, Fraser RM, Huffman JE, Vitart V, Hayward C, McQuillan R, Glodzik D, Polasek O, Hastie ND, et al. Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies. *PLoS One.* 2013; 8(7):e68604. [PubMed: 23874685]
- Lander ES. Initial impact of the sequencing of the human genome. *Nature.* 2011; 470(7333):187–97. [PubMed: 21307931]
- Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang ZZ, Bizon C, Lange EM, Smith JD, Turner EH, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet.* 2014; 94(2):233–45. [PubMed: 24507775]
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics.* 2012; 13(4):762–75. [PubMed: 22699862]

- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83(3):311–21. [PubMed: 18691683]
- Li B, Leal SM. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.* 2009; 5(5):e1000481. [PubMed: 19436704]
- Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One.* 2011a; 6(9):e24945. [PubMed: 21949800]
- Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Am J Hum Genet.* 2010; 87(5):728–35. [PubMed: 21055717]
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2011b; 34(8):816–34.
- Liu K, Luedtke A, Tintle N. Optimal methods for using posterior probabilities in association testing. *Hum Hered.* 2013; 75(1):2–11. [PubMed: 23548776]
- Magi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP. Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases. *Genet Epidemiol.* 2012
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–53. [PubMed: 19812666]
- Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010; 11(7):499–511. [PubMed: 20517342]
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20(9):1297–303. [PubMed: 20644199]
- Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Muzny D, Yu F, Rice K, Zhu C, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet.* 2013; 45(8):899–901. [PubMed: 23770607]
- Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet.* 2012; 44(6):631–5. [PubMed: 22610117]
- Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kahler A, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature.* 2014; 506(7487):185–90. [PubMed: 24463508]
- Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC. Performance of genotype imputations using data from the 1000 Genomes Project. *Hum Hered.* 2012; 73(1):18–25. [PubMed: 22212296]
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014; 42(Database issue):D1001–6. [PubMed: 24316577]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89(1):82–93. [PubMed: 21737059]
- Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet.* 2010; 87(5): 604–17. [PubMed: 21070896]
- Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. Searching for missing heritability: Designing rare variant association studies. *Proc Natl Acad Sci U S A.* 2014; 111(4):E455–64. [PubMed: 24443550]

Figure 1. Pre-collapsing imputation approach

#1 – #10 represent variants in the reference panel. PreC #1 indicates a pre-collapsed variable by collapsing rare variants (#5, #7, and #8) in the reference panel. In the genotype panel, rare variants and pre-collapsed variables are untyped markers (bold faced and italicized) estimated via imputation analysis.

Figure 2. 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) 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. 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 4. Effect of additional information on imputation performance

Comparison of mean dosage r^2 values obtained by the PreCimp without additional information, PostC with additional information (PostC+), and PreCimp with additional information (PreCimp+) are plotted by dosage r^2 bin of PostC method with additional information.

Figure 5. 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), 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)

Table I

Datasets used in this study

Panel (# of samples)	Exome Seq	GWAS chip (AFFY 5.0)	Exome array
# 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

“O” indicates the data is used for constructing a panel and “X” otherwise.

Table II

Enhanced imputation accuracy by the PreCimp method

Dosage r^2 bin of PostC	Mean increase in dosage r^2		Relative increase in dosage r^2 (%) [*]	
	ALL (3,830) (# of genes)	< 200kb (3,717) (# of genes)	200kb(113) (# of genes)	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%

^{*} Relative increase in dosage r^2 was calculated as (PreCimp–PostC)/PostC.