Kim *et al. BMC Genomics* (2015) 16:1109
DOI 10.1186/s12864-015-2192-y

BMC Genomics

METHODOLOGY ARTICLE Open Access

A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data

Young Jin Kim[1,2], Juyoung Lee[2], Bong-Jo Kim[2], T2D-Genes Consortium and Taesung Park[1,3*]

Abstract

Background: Rare variants have gathered increasing attention as a possible alternative source of missing heritability. Since next generation sequencing technology is not yet cost-effective for large-scale genomic studies, a widely used alternative approach is imputation. However, the imputation approach may be limited by the low accuracy of the imputed rare variants. To improve imputation accuracy of rare variants, various approaches have been suggested, including increasing the sample size of the reference panel, using sequencing data from study-specific samples (i.e., specific populations), and using local reference panels by genotyping or sequencing a subset of study samples. While these approaches mainly utilize reference panels, imputation accuracy of rare variants can also be increased by using exome chips containing rare variants. The exome chip contains 250 K rare variants selected from the discovered variants of about 12,000 sequenced samples. If exome chip data are available for previously genotyped samples, the combined approach using a genotype panel of merged data, including exome chips and SNP chips, should increase the imputation accuracy of rare variants.

Results: In this study, we describe a combined imputation which uses both exome chip and SNP chip data simultaneously as a genotype panel. The effectiveness and performance of the combined approach was demonstrated using a reference panel of 848 samples constructed using exome sequencing data from the T2D-GENES consortium and 5,349 sample genotype panels consisting of an exome chip and SNP chip. As a result, the combined approach increased imputation quality up to 11 %, and genomic coverage for rare variants up to 117.7 % (MAF < 1 %), compared to imputation using the SNP chip alone. Also, we investigated the systematic effect of reference panels on imputation quality using five reference panels and three genotype panels. The best performing approach was the combination of the study specific reference panel and the genotype panel of combined data.

Conclusions: Our study demonstrates that combined datasets, including SNP chips and exome chips, enhances both the imputation quality and genomic coverage of rare variants.

Keywords: Combined approach, Exome chip, Imputation, Rare variant

* Correspondence: tspark@stats.snu.ac.kr
[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, South Korea
[3]Department of Statistics, Seoul National University, San 56-1, Shilim-dong, Kwanak-gu, Seoul 151-742, South Korea
Full list of author information is available at the end of the article

Background

Genome-wide association studies (GWAS) have revealed unprecedented numbers of disease-associated loci [1, 2]. However, previously reported loci explain only a small proportion of heritability [2–4]. Previous GWAS mainly focused on common variants that was readily accessible via initial genomic technologies [2]. Through recent advancements in high-throughput sequencing technology, more complete genome-wide assessment of variants has become possible [5]. Recent large scale sequencing studies reported that the population frequencies of a large proportion of discovered variants were rare (Minor Allele Frequency (MAF) < 1 %) [6–8]. Given their abundance, rare variants have been increasingly recognized as an alternative source of missing heritability [5, 7, 9]. However, large-scale, population-based genomic sequencing studies are not yet feasible, due to high cost and computation-intensive analysis [10, 11].

Alternatively, imputation has been widely used for studying rare variants. Imputation has estimated untyped rare variants using thousands of sequenced samples available as a reference panel such as the 1,000 genomes project data [12, 13]. Recent imputation-based association studies have revealed numerous uncommon or rare variants associated with coronary artery disease, blood cell traits, serum creatinine, chronic kidney disease, and adult body height [10, 12, 13]. However, imputing rare variants has been challenging, due to the low accuracy of imputed genotypes of rare variants [10, 14], and poorly imputed rare variants may mislead follow-up association studies.

Imputation requires a reference panel and genotype panels. The reference panel is the basis for imputation performance, and the genotype panel is made from observed data. From both the reference and genotype panels, the shared haplotype segments are estimated using variants present on both panels. Then, the untyped genotypes are imputed using these haplotypes [15]. The accuracy of imputation can be increased by improving the reference panel and the genotype panel.

Previously, numerous studies have reported enhanced imputation performance of rare variants [14, 16–19], using three types of basic approaches for improving reference panels. The first is to increase the number of samples of the reference panel [14]. The second approach uses a study-specific reference panel instead of a public reference panel (e.g., the 1,000 genomes project reference panel) [16, 17]. The last 'two step approach' uses a local reference panel consisting of a subset of samples with a chip containing many low frequency variants or local sequencing data [18, 19]. Such local reference panels are used to complement public reference panels. These three approaches mainly focus on improving the reference panels by constructing an independent, study-specific reference panel, or complementing an existing public reference panel.

Alternatively, the imputation accuracy of rare variants can also be increased by improving a genotype panel using a chip designed to contain rare variants or markers tagging nearby rare variants [14, 18]. For example, an exome chip is a customized chip containing about 250 K variants including rare functional coding variants selected from ~ 12,000 sequenced samples [20] that can be genotyped at less cost than commercial genome-wide single nucleotide polymorphism (SNP) chips containing rare variants. While exome chips alone were shown as insufficient for imputation, as compared to commercial SNP chips widely used for GWAS [21], exome chips additionally genotyped for previously SNP chip-genotyped samples would improve their utility as a good source of rare variants. Moreover, a genotype panel combining exome chips and SNP chip data can help construct population-specific haplotypes carrying rare variants, thus also increasing the imputation accuracy of rare variants.

In this study, we propose a new strategy to increase the accuracy of imputation of rare variants by improving a new genotype panel by combining exome chip with existing SNP chip data. We show that the new genotype panel of combined data of exome chip and SNP chip improves imputation quality of imputed rare variants. To demonstrate the effectiveness of our strategy for improving genotype panels, we compared imputation strategies based on three genotype panels: 1) exome chip only; 2) SNP chip only; and 3) combined SNP chip and exome chip. Performances were compared via imputation quality scores [22] and genomic coverages [23, 24].

We also performed a systematic investigation of the effect of the reference panel on the imputation quality of rare variants. Using 848 samples with whole exome sequencing data (WES), SNP chip data (GWAS), and exome chip data (EXOME), we built various reference panels: WES + GWAS + EXOME, WES + GWAS, WES + EXOME, WES, and the 1000 genomes phase 1 dataset (1KG). We then performed imputation on 5,349 samples with three genotype panels of exome chip, SNP chip, and combined data. Additionally, to assess the effect of the reference panel sample size on imputation performance, we varied the number of samples from 300 to 848, by increments of 200, to examine the performance of imputation strategies.

Results

Exclusion of poorly-imputable variants

In this study, we used the estimated r^2 ($\hat{r}^2$) values provided by the imputation software minimac [15, 25] as imputation quality scores. The $\hat{r}^2$ is an estimate of the true r^2 (dosage r^2), which is the squared correlation of the true genotypes and imputed genotypes [25]. It is given by comparing the variance of imputed genotype scores with the variance of expected genotype scores. Previously, Li et al. reported that $\hat{r}^2$ was poorly estimated for very rare variants (MAF ≤ 0.5 %) [14]. Almost all variants with poorly calibrated $\hat{r}^2$ were very rare variants. Also, the Pearson correlation coefficient between $\hat{r}^2$ and dosage r^2 was below 0.9 (0.783–0.825) for very rare variants, but it was more than 0.9 (0.951–0.983) for variants with MAF > 0.5 % [14]. If the $\hat{r}^2$ could not estimate the true value closely, then our comparative analysis of imputation performance would yield misleading results. In this context, we called very rare variants as “poorly-imputable variants” if the Pearson correlation coefficient $\hat{r}^2$ and dosage r^2 was below 0.9 and excluded them for further analysis.

We compared $\hat{r}^2$ and dosage r^2 in four different MAF bins. We divided the variants into four MAF bins based on MAF: common (MAF ≥ 5 %), less common (MAF 1–5 %), rare (MAF 0.5–1 %), and very rare (MAF < 0.5 %) [14]. To measure the strength of the linear relationship between $\hat{r}^2$ and dosage r^2, Pearson correlation coefficients were calculated for each MAF bin. We first performed imputation on the genotype panel containing the SNP chip only by using our WES + GWAS + EXOME reference panel. Among imputed variants, 45,802 variants from 5,349 samples were compared to the corresponding variants obtained from an exome chip constructed using identical samples. Figure 1 shows the imputation results of variants by MAF bins. The Pearson correlation coefficients were 0.98, 0.97, 0.94, and 0.77 for MAF bins ≥5 %, 1–5 %, 0.5–1 %, and < 0.5 %, respectively. As Li et al. reported, $\hat{r}^2$ did not reflect the true value, dosage r^2, for very rare variants (MAF < 0.5 %, Fig. 1d). However, $\hat{r}^2$ became closer to the dosage r^2, as the MAF increased (Fig. 1a–c).

In this study, we used whole exome sequencing data (~18,000 genes) for studying imputation performance. However, Li et al. studied imputation performance using sequencing data of only 202 genes [14]. Since the results may be different depending on the data, we thoroughly analyzed very rare variants to study the lower bound of allele frequency showing good estimation of dosage r^2. Very rare variants were split into discrete MAF bins of width 0.001. The Pearson correlation coefficients were 0.91, 0.90, 0.87, 0.78, and 0.55 for MAF bins 0.4–0.5 %, 0.3–0.4 %, 0.2–0.3 %, 0.1–0.2 %, and 0–0.1 %, respectively. The Pearson correlation coefficients were 0.98, 0.97, 0.94, 0.77 for MAF bins ≥ 5 %, 1–5 %, 0.5-1 %, and < 0.5 %, respectively. Among very rare variants, MAF bins with MAF < 0.3 % showed that the Pearson correlation coefficients dropped below 0.9 (Fig. 1e–f). If these variants are included in the analysis, poorly estimated $\hat{r}^2$ may cause

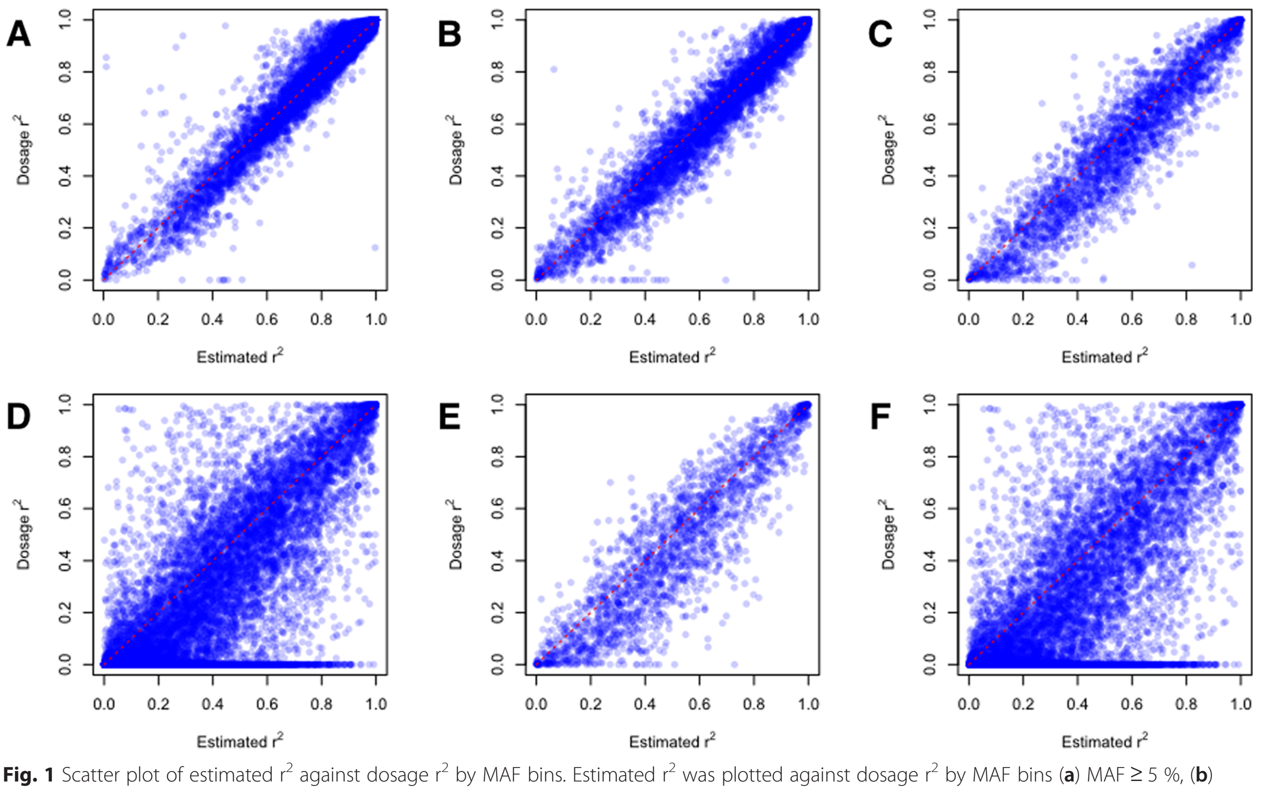

Fig. 1 Scatter plot of estimated r² against dosage r² by MAF bins. Estimated r² was plotted against dosage r² by MAF bins (**a**) MAF ≥ 5 %, (**b**) MAF = 1–5 %, (**c**) MAF = 0.5–1 %, (**d**) MAF < 0.5 %, (**e**) MAF = 0.3–0.5 %, and (**f**) MAF < 0.3 %. The red dotted line represents the diagonal

less consistent results to those using dosage r^2. Therefore, the variants with MAF < 0.3 % (369,309 of 856,690 variants) were regarded as poorly-imputable and excluded from further study.

Comparison of imputation quality among genotype panels

Using the WES + GWAS + EXOME reference panel excluding poorly imputable variants, we performed imputation on three genotype panels, including exome chip only, SNP chip only, and combined data of exome chip and SNP chip. For the purpose of comparison, we selected 108,682 imputed variants overlapping the three genotype panels. The $\hat{r}^2$ value was used to measure imputation quality. The numbers of variants were 35,443 (32.6 %), 21,191 (19.5 %), 19,527 (18.0 %), and 32,547 (29.9 %) for variants with MAF ≥ 5 %, 1–5 %, 0.5–1 %, and < 0.5 %, respectively. Figure 2 shows the comparison results. As previously reported, the genotype panel of the exome chip alone showed the worst performance [21]. The mean $\hat{r}^2$ values were 0.332, 0.616, and 0.661 for the genotype panels of exome chip, SNP chip, and combined approach, respectively. Thus, the combined genotype panel showed the best performance compared to the other genotype panels (Wilcoxon signed rank sum test *P*-values < 2.2x10^{-16}, a 7.3 % relative increase in the mean of the $\hat{r}^2$ compared to those of SNP chip only). In Fig. 2, most imputed variants using the combined approach performed better than the genotype panel of SNP chip alone. The mean values of the $\hat{r}^2$ of SNP and combined approach were 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 <0.5 %, respectively. The relative increment in $\hat{r}^2$ of the combined genotype panel was 4.1 % for common variants (MAF ≥ 5 %), 8.1 % for less common variants (MAF 1–5 %), 10.7 % for rare variants (MAF 0.5–1 %), and 11.4 % for very rare variants (MAF < 0.5 %), compared to the genotype panel with SNP chip only. Thus, the increment in imputation quality was largest when the minor allele frequencies of the imputed variants were below 1 %.

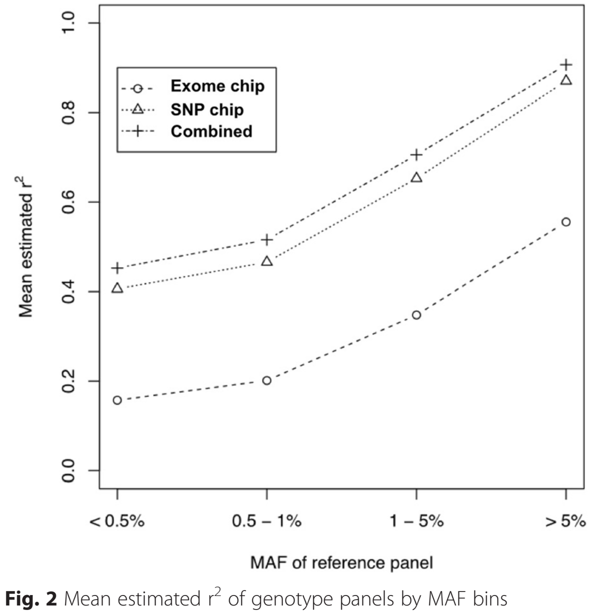

Fig. 2 Mean estimated r^2 of genotype panels by MAF bins

Comparison of genomic coverage among genotype panels

We next compared the genotype panels in terms of their genomic coverage, i.e., the proportion of variants captured by a genotyping chip [24]. The larger the genomic coverage, the better the association mapping performance. One major advantage of imputation lies in obtaining a large dense set of imputed variants from a relatively small number of observed variants on the genotype panel. These imputed variants enhanced genomic coverage, resulting in increased association power [15], enabling us to perform *in silico* fine mapping in imputation-based association studies. Likewise, Nelson et al. recently reported imputation-based genomic coverage of widely used genotyping chips [24]. Imputation-based genomic coverage is calculated as the number of imputed variants with quality scores above the threshold value (info score ≥ 0.8) divided by the total number of variants in the reference panel [24].

In this study, we compared imputation-based genomic coverage of three genotype panels. We imputed genotype panels using the WES + GWAS + EXOME reference panel. For genomic coverage, we selected 143,022 exonic variants, including those imputed and genotyped by exome chip. Since we used exome sequencing data to construct the reference panel, 143,022 variants were regarded as a "virtual" exome. The numbers of variants were 56,326 (39.4 %), 28,072 (19.6 %), 22,931 (16.0 %), and 35,693 (25.0 %) with MAFs ≥ 5 %, 1–5 %, 0.5–1 %, and < 0.5 %, respectively. Table 1 summarizes the genomic coverages. We also selected two cut-off values for $\hat{r}^2$: 0.8 as a stringent cut-off, and 0.4 as a less stringent cut-off. This 0.8 stringent cut-off provided a genomic coverage of 0.435 for the SNP chip only and 0.560 for the combined approach, respectively, while the less stringent cut-off ($\hat{r}^2 \geq$ 0.4) provided genomic coverages of 0.749 and 0.818 for SNP chip only and the combined approach, respectively. Overall, the combined approach yielded approximately 9.2 % ($\hat{r}^2 \geq$ 0.4) and 29 % ($\hat{r}^2 \geq$ 0.8) relative increases in genomic coverage of all variants. However, for rare variants (MAF < 1 %) applying stringent cut-offs ($\hat{r}^2 \geq$ 0.8), the genomic coverage of the combined approach increased by 98.6 % and 117.7 %, compared to that of the SNP chip only, for

Table 1 Genomic coverage of genotype panels of SNP chip only and combined approach

MAF bin	Estimated $r^2 \geq 0.8$			Estimated $r^2 \geq 0.4$		
	Exome chip	SNP chip	Combined	Exome chip	SNP chip	Combined
ALL	0.367	0.435	0.560	0.492	0.749	0.818
≥5 %	0.600	0.794	0.901	0.756	0.953	0.983
1–5 %	0.374	0.403	0.588	0.510	0.799	0.881
0.5–1 %	0.192	0.146	0.290	0.290	0.585	0.686
<0.5 %	0.107	0.079	0.172	0.192	0.491	0.591

variants with MAFs of 0.5–1 % and < 0.5 %, respectively. Thus, the combined approach greatly improved genomic coverages of rare variants.

Systematic effect of various reference panels on imputation quality

Systematic analysis of reference panels could establish an effective strategy for improving imputation quality of rare variants. We first investigated systematically the effects of various reference panels on imputation quality using five different reference panels, including 1KG, WES, WES + EXOME, WES + GWAS, and WES + GWAS+ EXOME. We then compared imputation qualities by using 15 possible combinations of three genotype panels and five reference panels. $\hat{r}^2$ values were used to measure imputation quality. For comparisons, we selected 66,920 variants imputed by all 15 combinations. Figure 3 shows the imputation qualities by genotype panels used: Fig. 3a for exome chip, 3b for SNP chip, and 3c for the combined. Imputation quality was influenced by reference panels, and the same reference panel showed different imputation quality when used for imputing different genotype panels. The WES + EXOME reference panel was the best performing genotype panel of exome chips, and the WES + GWAS + EXOME reference panel was best for the genotype panels of SNP chip and the combined data. The 1KG reference panel was the worst performing genotype panel of exome chip, and the WES reference panel for the genotype panels of SNP chip and the combined data.

We next investigated why imputation quality differed across the reference panels, and found two factors: 1) differences in the number of overlapping variants between the reference and genotype panels; and 2) shared haplotype patterns between the reference and genotype panels.

First, overlapping variants between the reference and genotype panels play important roles in imputation. For predicting untyped genotypes, imputation uses estimated haplotype segments from the reference panel for the overlapping variants between reference and genotype panels. The lower the number of overlapping variants between the reference and genotype panels, the worse the imputation quality, due to inaccurate estimation of haplotype segments. The numbers of overlapping variants are summarized in Table 2. For study-specific reference panels that included WES data, the worst performing reference panel had the smallest number of overlapping variants, while the best performing reference panel had the most overlapping variants. This is the reason why the best and worst performing reference panels were different according to the genotype panels used. For example, WES + EXOME reference panel showed the best performance for the genotype panel of exome chip (Fig. 3a). The number of overlapped variants was 38,243,

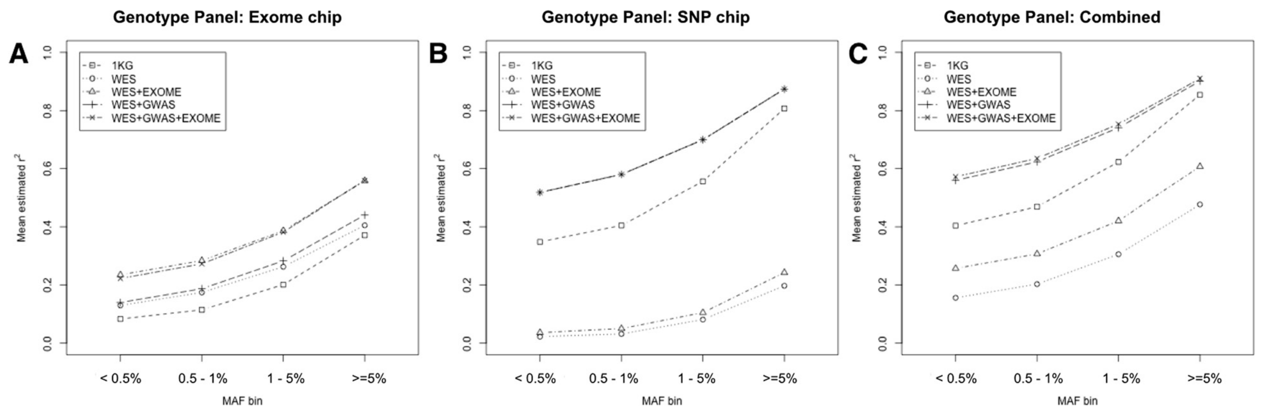

Fig. 3 Mean estimated r^2 of various combinations of reference panels and genotype panels. Reference panels are the 1000 genomes phase 1 dataset (1KG) and various combinations of whole exome sequencing data (WES), SNP chip data (GWAS), and exome chip data (EXOME)

Table 2 The number of overlapped variants between reference panels and genotype panels

Reference panels	Exome chip	SNP chip	Combined
WES	21,120	4,472	24,514
WES + EXOME	38,243	7,323	41,637
WES + GWAS	23,972	344,359	364,402
WES + GWAS + EXOME	38,243	344,359	378,695
1KG	49,286	344,359	389,715

which was the largest number among the number of overlapping variants between the four study-specific reference panels and genotype panel of exome chip (Table 2). On the other hand, WES + EXOME was the 2rd worst performing for other genotype panel, having the 2nd smallest number of overlapping variants (Figs. 3a and 3b, Table 2).

Second, the use of shared haplotype patterns between the reference and genotype panels also improved the accuracy of imputation. Although the 1KG reference panel had the largest number of overlapped variants, its imputation quality was worse than using study-specific reference panels such as WES + EXOME for exome chip (Fig. 3a) and WES + GWAS + EXOME for the genotype panels of SNP chip and the combined data (Fig. 3b and 3c). As previously reported [17], the better performance of study-specific reference panels over 1KG was due to more shared haplotype segments between study-specific reference panels and the genotype panels used [17, 26]. All samples of study-specific reference panels and genotype panels were from a Korean population, whereas the 1KG reference panel consists of samples from various populations such as Europeans, Africans, and Asians. Thus, the 1KG reference panel and the genotype panel of Korean population do not have many shared haplotypes, resulting in poor imputation.

Comparative analysis of the 15 different combinations provided us effective imputation strategies for imputing rare variants when a specific genotype panel is given with a specific reference panel. For example, compared to the case of the SNP chip genotype panel with the 1KG reference panel, three strategies could increase the imputation quality of rare variants: 1) use the best performing reference panel, e.g., WES + GWAS + EXOME, instead of 1KG; 2) construct the genotype panels of combined data by additional genotyping of an exome chip for those samples already genotyped using the SNP chip; and 3) use the genotype panel of combined data from (2, above) with WES + GWAS + EXOME as a reference panel. The first strategy increased imputation quality from 0.445 to 0.622 for rare variants and from 0.377 to 0.553 for very rare variants. When the second strategy was applied, the imputation quality was 0.503 and 0.426 for rare and very rare variants, while the best performing strategy was the third, increasing imputation quality to 0.664 and 0.595 for rare and very rare variants, respectively.

Sample size effect of reference panel on imputation quality

It was previously shown that the larger the sample size of the reference panel, the better the imputation quality [14]. Here, we systematically investigated the effect of sample size on imputation quality. Since the WES + GWAS + EXOME reference panel performed the best, we studied the sample size effect of WES + GWAS + EXOME reference panel on imputation quality. We performed imputation on the genotype panels of SNP chip and combined data with reference sample sizes of 300, 500, 700, and 848. This analysis was restricted to variants on chromosome 1, to focus only on overlapping imputed variants. The total number of imputed variants used for this analysis was 10,624. Figure 4 shows the mean $\hat{r}^2$ of SNP chip only and combined data by different sample sizes for the reference panel. As expected, imputation quality increased as the sample size of reference panel increased.

For the imputed rare variants (MAF 0.5–1 %) using the combined data genotype panel, relative increases in mean $\hat{r}^2$ values were 8.2 % and 4.5 % for increasing the reference panel sample size from 300 to 500, and 500 to 700, respectively (Table 3), and 0.93 % for sample size increases from 700 to 848. Increments in mean $\hat{r}^2$ values were the lowest between 700 and 848 samples used for the reference panel. Therefore, increasing the sample size to more than 700 would not be cost-effective for improving imputation quality of rare variants. Similarly, imputation qualities of very rare variants (MAF <0.5 %) showed similar patterns to those of rare variants. When we performed the same analysis for the SNP chip genotype panel,

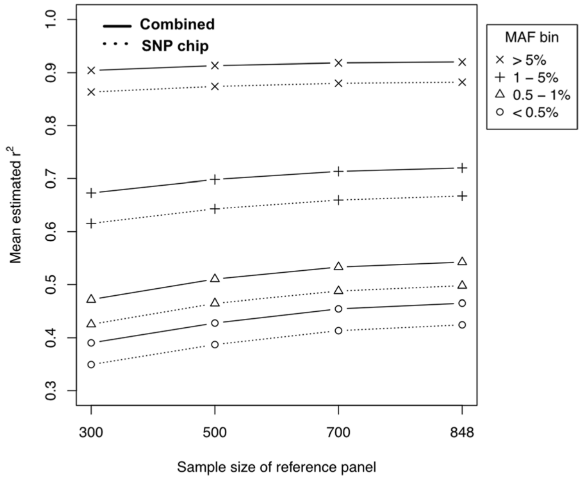

Fig. 4 Mean estimated r^2 varied by sample size of reference panel

Table 3 Relative increase in mean estimated r^2 by reference sample size (MAF 0.5–1 %)

Genotype panel	300 to 500	500 to 700	700 to 848
SNP chip only	9.27 %	4.90 %	2.09 %
Combined (SNP + exome chip)	8.18 %	4.46 %	1.74 %

similar patterns were observed for both rare and very rare variants.

This study of sample size also provided us cost-effective sample size determination strategies for imputing rare variants. Instead of constructing a reference panel with a large number of samples, it would be more cost-effective to combine the reference panel with a smaller number of samples and the genotype panel of combined data. Table 4 (MAF 0.5–1 %) summarizes the relative increases in imputation quality. For example, the genotype panel of combined data with 500 samples of reference panel showed better imputation quality (mean $\hat{r}^2$ = 0.510) than the genotype panel of SNP chip only data with 848 samples of reference panel (mean $\hat{r}^2$ = 0.498) (Fig. 4).

Discussion

In this study, we proposed a new strategy for increasing imputation quality of rare variants, i.e., a combined approach that uses the genotype panel of combined data including SNP chip and exome chip for imputation. Using a WES + GWAS + EXOME reference panel, we showed that the genotype panel of combined data yielded better imputation quality than other genotype panels. For rare variants (MAF < 1 %), the combined approach relatively increased imputation quality up to 11 % and enhanced genomic coverage up to 117.7 %, as compared to the genotype panel of SNP chip only.

In addition, we systematically investigated the effect of various reference panels on imputation quality. We believe the current study is the first to systematically analyze imputation qualities by combining various reference panels and genotype panels. The best imputation quality for rare variants was obtained using the study-specific reference panel WES + GWAS + EXOME and the genotype panel of combined data.

Table 4 Relative increase in mean estimated r^2 by using combination of reference panel and combined approach (MAFs 0.5–1 %)

	R300-C	R500-C	R700-C	R848-C
R300-G	10.88 %	19.95 %	25.30 %	27.49 %
R500-G	1.47 %	9.77 %	14.67 %	16.67 %
R700-G	−3.27 %	4.64 %	9.31 %	11.21 %
R848-G	−5.25 %	2.50 %	7.07 %	8.94 %

The names of panels were abbreviated as follows: R (reference panel), G (the genotype panel of SNP chip only), and C (the genotype panel of combined data). R300, R500, R700, and R848 indicates sample sizes of 300, 500, 700, and 848 for reference panel, respectively. R(sample size) -G represents combination of R(sample size) reference panel and G genotype panel. R(sample size)-C represents the combination of R(sample size) reference panel and C genotype panel

Our study also provides a guideline for researchers to establish more cost-effective imputation strategies for increasing the imputation quality of rare variants. As shown in the results, combining a reference panel with a reasonable sample size and the combined data genotype panel is a cost-effective approach to increasing imputation quality of rare variants. For example, we reported that the genotype panel of combined data with 500 samples of reference panel outperformed the genotype panel of SNP chip with 848 samples of reference panel. The cost per sample of exome sequencing ($750) is about 11 times more expensive than those of exome chip ($70) [27]. If less than 3,700 samples were genotyped, generating exome chip data would be much more cost-effective than producing additional 348 samples of exome sequencing data. As an alternative to producing exome sequencing data or genotyping exome chip data, the merged panel approach which combine the concatenation of a public reference panel (e.g. 1,000 genomes project data) and a study specific reference panel, can be considered to increase the sample size of reference panel. Recent studies reported that the merged panel enhanced imputation performance [17, 28]. Also, 1,000 genomes project phase 3 data became available, providing 2,504 samples of whole genome sequencing data [29]. Since rare variants tend to be population specific [30], the merged panel approach would be effective if study samples were closely related with populations of 1KG data. Populations of 1KG phase 3 are African (661 samples), American (347 samples), East Asian (504 samples), European (503 samples), and South Asian (489 samples). Customized genotyping chip can also be an alternative approach if none of SNP chip and exome chip data is available. Considering the frequency of rare variant in a specific population, one can design a study specific genotyping chip containing rare variants. For example, UK biobank designed a chip containing 821 K SNPs [31]. Among them, about 111 K variants were rare coding variants (see url: http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/). Therefore, based on our results, a rare variant association research considering the study data and populations of study samples can be designed more cost-effectively.

Despite providing a new imputation strategy and valuable insights for imputing rare variants, the current study has several limitations that should be noted. First, we removed substantial amount of variants from our analysis. Poorly-imputable variants (MAF < 0.3 %) were 43.1 % of variants in the initial reference panel. If additional true genotype data is available (e.g. exome sequencing data), dosage r^2 can be calculated instead of estimated r^2, then thorough analysis would be possible for all very rare variants. Second, the case samples of the reference panel may influence imputation results. Generally, public reference panels are consists of normal samples [8]. In this study, the 848 samples of the reference panel are consists of 415 type 2 diabetes and 433 controls. To analyze the impact of samples with disease status in the reference panel on imputation performance, we performed imputation on chromosome 1 of combined genotype panel using the reference panel consisting of 415 cases and the reference panel consisting of 433 controls. After that, we calculated the Pearson correlation coefficient between imputation qualities of two reference panels. The Pearson correlation coefficient was 0.96. Also Duan et al. previously reported that there was no loss of imputation quality using the reference panel consisting of samples with phenotypic extremes or disease status [17]. Thus, imputation results of our study may not be influenced by cases of the reference panel. Third, the comparison analysis of this study may not be exhaustive. We compared the imputation performance of the reference panels; however, 1KG phase 3 data and the merged panel approach were not included in the analysis. Further study is warranted to compare the reference panels and related approaches exhaustively. Lastly, our study did not report explicit cost benefits by imputation strategies considering costs of bioinformatics and whole genome sequencing. In practice, sequencing data analysis and imputation are not free of charge in considering compute-intensive analysis. Therefore, explicit cost-benefits analyses that argue for optimal designs in light of difference cost structures are required for a further study.

Next generation sequencing (NGS) provides base-pair resolution data and generated a near complete catalogue of genetic variants in human genome [8]. Unlike previous genome studies that focused on using initial genomic technologies such as chip-based genotyping on common variants, NGS is expected to uncover the role of less common and rare variants in various diseases. However, due to high cost and computation-intensive analysis, large-scale, population based genomic sequencing studies are not yet feasible [10, 11]. Although NGS is not yet cost-effective for large-scale genome studies, it will soon become essential, as its cost rapidly decreases. Meanwhile, imputation-based research strategies would be cost-effective for identifying associations between diseases and variants, including less common and rare variants.

Conclusions

Here, we proposed a combined approach that imputes rare variants using the genotype panel of combined data including SNP chip and exome chip. We evaluated the performance of the combined approach using 848 samples from a study-specific reference panel and 5,349 samples of genotype panels consisting of exome chips only, SNP chips only, and combined data from an exome chip and a SNP chip. For rare variants (MAF < 1 %), the combined approach greatly increased imputation quality approximately 11 % compared to that of the exome chip only and showed up to a 117.7 % increase in genomic coverage. The proposed combined approach would be a cost-effective strategy to obtain better imputation quality and enhanced genomic coverage for rare variants.

We also systematically investigated the effect of various combinations of reference panels and genotype panels. The best performing approach combined data from a study-specific reference panel and a genotype panel.

Methods

Study samples

As part of the Korean Genome Analysis Project, Korea Association REsource (KARE) study was initiated in 2007 to conduct a large-scale genome-wide association study aiming to discover variants associated with Type 2 diabetes and numerous complex traits. The detailed information has been described elsewhere [32]. Briefly, a total of 10,038 participants aged 40 to 69 were recruited from two population-based cohorts comprising the Ansung ($n = 5{,}018$) and Ansan (5,020) cohorts. In this study, we used exome sequencing data and genotyping data from KARE samples. All participants of KARE provided written informed consent. 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

By the Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Ethnic Samples (T2D-GENES) Consortium, about 10,000 exomes from five ethnic groups were sequenced using Agilent Human Exon v2 capture (~18,000 genes) at the Broad Sequencing Center. Among these, part of the samples are from the KARE project [32], including unrelated 538 type 2 diabetes samples and 579 control samples, were included, and 1,087 samples were used for further analysis after sample quality control. The reference genome hg19 was used for alignment and variant calling performed using the Genome Analysis Toolkit v2 [33]. Among 1,087 exome sequenced samples, only 848 samples had data of both SNP chip and exome chip. As a result, 500,821 autosomal variants from 848 Korean samples were used for constructing reference panels. The accuracy of the called variants was calculated by comparing genotypes from sequencing data with genotypes of genotyping chip data, showing overall concordances of 99.76 % and 99.96 % for the Affymetrix 5.0 and exome chips, respectively.

SNP chip and exome chip data

Previously, 8,842 samples of KARE project were genotyped using the Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix Inc., San Diego, CA, USA) [32]. Among them, 6,197 identical samples were genotyped using the Illumina HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA, USA) exome chip. For the two platforms, standard sample quality controls were conducted, excluding those with a high missing rate (>4 %), gender discrepancy, excessive heterozygosity, or cryptic first-degree relatives. Exclusion criteria for the Affymetrix SNP chip were: Hardy-Weinberg equilibrium p-values < 10^{-6}, genotype call rates < 95 %, and MAF < 0.01. All SNP chromosomal positions were updated to hg19 using the Affymetrix annotation file. Quality control of exome chip variants was similar to that of the SNP chip, except for the threshold for filtering out variants with low allele frequency. Only monomorphic variants were excluded for further analysis. From quality controlled data, we used 6,197 samples that were common between sets of the Affymetrix SNP chip and exome chips. Variants included in the analysis were 344,366, and 66,196 for the SNP chip and exome chip, respectively. Among 6,197 samples, 848 samples were used for constructing the reference panel, and the remaining 5,349 samples were used for genotype panels.

Construction of reference panels

We constructed the reference panel by merging whole exome sequencing, exome chip, and SNP chip of 848 identical samples. Prior to merging, overlapping variants between the WES and chip data were removed from chip datasets. For overlapping variants between GWAS and EXOME, variants from EXOME were used to remove overlapping variants from GWAS. After merging all the data, the initial reference panel contained 856,690 variants. For comparisons, we excluded poorly imputable variants (MAFs < 0.3 %) for further analysis. The final WES + GWAS + EXOME reference panel contained 487,381 variants and was phased using the ShapeIT v2 program [34]. After phasing, a subset of variants from the WES + GWAS + EXOME reference panel was selected for constructing three reference panels: WES, WES + GWAS, and WES + EXOME. We also downloaded 1,000 genomes phase I Shapeit2 reference with no monomorphic and no singleton sites from MACH website (http://www.sph.umich.edu/csg/abecasis/MACH).

Construction of genotype panels

Among 6,197 samples, 5,349 samples remained after excluding 848 samples were used for constructing the reference panel. The genotype panel consisting of exome chip of 5,349 samples was phased using the ShapeIT v2 progam. As the genotype panel of SNP chip only, SNP chip data of 5,349 samples were phased using the ShapeIT v2 program. For the genotype panel of combined data, the SNP chip and exome chip of 5,349 identical samples were merged and phased using the ShapeIT v2 program.

Statistical analysis

In this study, we performed typical pre-phasing-based imputation on genotype panels [35]. For imputation, we used minimac software, a low memory and computationally efficient implementation of the MaCH algorithm [25]. The dosage r^2 was accessed by calculating squared Pearson correlations between imputed dosages and true genotypes from exome chip. For comparison analysis of imputation performance, we used $\hat{r}^2$ provided by minimac as an imputation quality measure. To compare the imputation results between pairs of genotype panels, Wilcoxon signed-rank tests were performed for $\hat{r}^2$ values of imputed variants. Statistical analyses and visualization of the results were performed using the R program (http://www.r-project.org).

Availability of data

Exome sequencing data will be available on dbGAP. The genotype data of KARE samples are available by sending a request to the Distribution desk of Korea Biobank Network, National Institute of Health, Korea.

Abbreviations

GWAS: Genome-wide association study; MAF: Minor allele frequency; SNP: Single nucleotide polymorphism; WES: Whole exome sequencing data; EXOME: Exome chip data; 1KG: 1,000 genomes phase 1 dataset; NGS: Next generation sequencing; KARE: Korea Association REsource; T2D-GENES: Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Ethnic Samples.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TP conceived and designed the study. YJK analyzed the data. YJK, JL, BJK, and TP are contributed to interpretation of the data and writing the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant (2012R1A3A2026438), by the Bio & Medical Technology Development Program of the NRF grant (2013M3A9C4078158), by grants from Korea Centers for Disease Control and Prevention (4845–301, 4851–302, 4851–307) and an intramural grant from the Korea National Institute of Health (2014-NI73001-00), the Republic of Korea. Sequencing data from the T2D-GENES Consortium was supported by NIH/NIDDK U01's DK085501, DK085524, DK085526, DK085545 and DK085584.

Membership of the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium

Gonçalo Abecasis[1], Marcio Almeida[2], David Altshuler[3,4,5,6,7,8,9], Jennifer L Asimit[10], Gil Atzmon[11], Mathew Barber[12], Nir Barzilai[13], Nicola L Beer[14],

Graeme I Bell[12,15], Jennifer Below[16], Tom Blackwell[1], John Blangero[2], Michael Boehnke[1], Donald W Bowden[17,18,19,20], Noël Burtt[3], John Chambers[21,22,23], Han Chen[24], Peng Chen[25], Peter S Chines[26], Sungkyoung Choi[27], Claire Churchhouse[3], Pablo Cingolani[28], Belinda K Cornes[29], Nancy Cox[12,15], Aaron G Day-Williams[10], Ravindranath Duggirala[2], Josée Dupuis[24], Thomas Dyer[2], Shuang Feng[1], Juan Fernandez-Tajes[30], Teresa Ferreira[30], Tasha E Fingerlin[31], Jason Flannick[3,5], Jose Florez[3,5,6], Pierre Fontanillas[3], Timothy M Frayling[32], Christian Fuchsberger[1], Eric R Gamazon[14], Kyle Gaulton[30], Saurabh Ghosh[33], Benjamin Glaser[34], Anna Gloyn[14], Robert L Grossman[15,35], Jason Grundstad[35], Craig Hanis[16], Allison Heath[35], Heather Highland[16], Momoko Horikoshi[30], Ik-Soo Huh[27], Jeroen R Huyghe[1], Kamran Ikram[29,36,37,38], Kathleen A Jablonski[39], Goo Jun[1], Norihiro Kato[40], Jayoun Kim[27], Young Jin Kim[41], Bong-Jo Kim[41], Juyoung Lee[41], C Ryan King[42], Jaspal Kooner[22,23,43], Min-Seok Kwon[27], Hae Kyung Im[42], Markku Laakso[44], Kevin Koi-Yau Lam[25], Jaehoon Lee[27], Selyeong Lee[27], Sungyoung Lee[27], Donna M Lehman[45], Heng Li[3], Cecilia M Lindgren[30], Xuanyao Liu[25,46], Oren E Livne[12], Adam E Locke[1], Anubha Mahajan[30], Julian B Maller[30,47], Alisa K Manning[3], Taylor J Maxwell[16], Alexander Mazoure[48], Mark I McCarthy[14,30,49], James B Meigs[6,50], Byungju Min[27], Karen L Mohlke[51], Andrew P Morris[30], Solomon Musani[52], Yoshihiko Nagai[48], Maggie C Y Ng[17,18], Dan Nicolae[12,15,53], Sohee Oh[27], Nicholette Palmer[17,18,19], Taesung Park[27], Toni I Pollin[54], Inga Prokopenko[30,55], David Reich[3,4], Manuel A Rivas[30], Laura J Scott[1], Mark Seielstad[56], Yoon Shin Cho[57], Xueling Sim[1], Robert Sladek[48,58], Philip Smith[59], Ioanna Tachmazidou[10], E Shyong Tai[25,36,60], Yik Ying Teo[61,62,63,64,65], Tanya M Teslovich[1], Jason Torres[12,15], Vasily Trubetskoy[12,15], Sara M Willems[66], Amy L Williams[3,4], James G Wilson[67], Steven Wiltshire[30], Sungho Won[68], Andrew R Wood[32], Wang Xu[60], Joon Yoon[27], Matthew Zawistowski[1], Eleftheria Zeggini[10], Weihua Zhang[21], and Sebastian Zöllner[1,69]

1. Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.
2. Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas 78227, USA.
3. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.
4. Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
5. Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston 02114,Massachusetts, USA.
6. Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA.
7. Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
8. Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02114, USA.
9. Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
10. Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK.
11. Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York10461, USA.
12. Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
13. Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, NY, USA
14. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, OX3 7LJ, UK.
15. Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA.
16. Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA.
17. Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157,USA.
18. Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
19. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
20. Internal Medicine-Endocrinology, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.
21. Department of Epidemiology and Biostatistics, Imperial College London, London SW7 2AZ, UK.
22. Imperial College Healthcare NHS Trust, London W2 1NY, UK.
23. Ealing Hospital National Health Service (NHS) Trust, Middlesex UB1 3HW, UK.
24. Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02115, USA.
25. Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore.
26. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
27. Seoul National University, Seoul 110-799, South Korea.
28. McGill Centre for Bioinformatics, McGill University, Montréal, Quebec, H3G 0B1, Canada.
29. Singapore Eye Research Institute, Singapore National Eye Centre, Singapore 168751, Singapore.
30. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.
31. Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado 80045, USA.
32. Genetics of Complex Traits, University of Exeter Medical School, Exeter, EX4 4SB, UK.
33. Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700018, India
34. Department of Endocrinology, Hadassah Medical Center, Kiryat Hadassah, Jerusalem, Israel
35. Institute for Genomics and Systems Biology,University of Chicago, Chicago, Illinois 60637, USA.
36. Duke National University of Singapore Graduate Medical School, Singapore 169857, Singapore.
37. Department of Ophthalmology, National University of Singapore and National University Health System, Singapore 119228, Singapore.
38. Department of Ophthalmology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.
39. The Biostatistics Center, GeorgeWashington University, Rockville, Maryland 20852, USA.
40. Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo162-8655, Japan.
41. Center for Genome Science, Korea National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do, 363-951, South Korea.
42. Department of Health Studies, University of Chicago, Chicago, Illinois60637, USA.
43. National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London W12 0HS, UK.
44. Department of Medicine, University of Eastern Finland, Kuopio Campus and Kuopio University Hospital, FI-70211 Kuopio, Finland.
45. Division of Clinical Epidemiology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA.
46. Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.
47. Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK.
48. McGill University, Montréal, Québec H3A 0G4, Canada.
49. Oxford NIHR Biomedical Research Centre, Churchill Hospital, Headington, OX3 7LE, UK.
50. General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
51. Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina 27599, USA.
52. Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi 39126, USA.
53. Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA.
54. Department of Medicine, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, Maryland21201, USA.
55. Genomics of Common Disease, Imperial College London, Hammersmith Hospital, London, W12 0NN, UK.
56. University of California San Francisco, San Francisco, California 94143, USA.
57. Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do, 200-702 South Korea.
58. Department of Medicine, Royal Victoria Hospital, Montréal, Québec H3A1A1, Canada.
59. National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Bethesda, MD 20817, USA.
60. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore117597, Singapore.
61. Department of Epidemiology and Public Health, National University of

Singapore, Singapore 117597, Singapore.
62. Centre for Molecular Epidemiology, National University of Singapore, Singapore 117456, Singapore.
63. Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore.
64. Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.
65. Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore.
66. Department of Genetic Epidemiology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.
67. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA.
68. Department of Epidemiology and Biostatistics, School of Public Health & Institute of Health and Environment, Seoul National University, Seoul 151-742, Republic of Korea
69. Department of Psychiatry, University of Michigan, Ann Arbor, Michigan 48109, USA.

Author details

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, South Korea. [2]Division of Structural and Functional Genomics, Center for Genome Science, Korean National Institute of Health, Osong, Chungchungbuk-do 363-951, South Korea. [3]Department of Statistics, Seoul National University, San 56-1, Shilim-dong, Kwanak-gu, Seoul 151-742, South Korea.

Received: 18 March 2015 Accepted: 3 November 2015
Published online: 29 December 2015

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. 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. doi:10.1073/pnas.0903103106.
2. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111(4):E455–64. doi:10.1073/pnas.1322563111.
3. 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. doi:10.1038/nrg2867.
4. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet. 2008;82(1):100–12. doi:10.1016/j.ajhg.2007.09.006.
5. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23. doi:10.1016/j.ajhg.2014.06.009.
6. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20. doi:10.1038/nature11690.
7. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337(6090):100–4. doi:10.1126/science.1217876.
8. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. doi:10.1038/nature11632.
9. Saint Pierre A, Genin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13(5):353–61. doi:10.1093/bfgp/elu025.
10. Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, Nalls MA, 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. doi:10.1016/j.ajhg.2012.08.031.
11. 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. doi:10.1002/gepi.21675
12. Du M, Auer PL, Jiao S, Haessler J, Altshuler D, Boerwinkle E, et al. Whole-exome imputation of sequence variants identified two novel alleles associated with adult body height in African Americans. Hum Mol Genet. 2014;23(24):6607–15. doi:10.1093/hmg/ddu361.
13. Sveinbjornsson G, Mikaelsdottir E, Palsson R, Indridason OS, Holm H, Jonasdottir A, et al. Rare mutations associating with serum creatinine and chronic kidney disease. Hum Mol Genet. 2014;23(25):6935–43. doi:10.1093/hmg/ddu399.
14. Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One. 2011;6(9):e24945. doi:10.1371/journal.pone.0024945.
15. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511. doi:10.1038/nrg2796.
16. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur J Hum Genet. 2014. doi: 10.1038/ejhg.2014.19.
17. Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, Jun G, et al. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. Bioinformatics. 2013;29(21):2744–9. doi:10.1093/bioinformatics/btt477.
18. Joshi PK, Prendergast J, Fraser RM, Huffman JE, Vitart V, Hayward C, 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. doi:10.1371/journal.pone.0068604.
19. Kreiner-Moller E, Medina-Gomez C, Uitterlinden AG, Rivadeneira F, Estrada K. Improving accuracy of rare variant imputation with a two-step imputation approach. Eur J Hum Genet. 2015. doi: 10.1038/ejhg.2014.91.
20. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, Stringham HM, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet. 2013;45(2):197–201. doi:10.1038/ng.2507.
21. Martin AR, Tse G, Bustamante CD, Kenny EE. Imputation-based assessment of next generation rare exome variant arrays. Pac Symp Biocomput. 2014:241–52
22. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529. doi:10.1371/journal.pgen.1000529.
23. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013;92(1):52–66. doi:10.1016/j.ajhg.2012.12.005.
24. Nelson SC, Doheny KF, Pugh EW, Romm JM, Ling H, Laurie CA, et al. Imputation-based genomic coverage assessments of current human genotyping arrays. G3 (Bethesda). 2013;3(10):1795–807. doi:10.1534/g3.113.007161.
25. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34. doi:10.1002/gepi.20533.
26. Huang L, Jakobsson M, Pemberton TJ, Ibrahim M, Nyambo T, Omar S, et al. Haplotype variation and genotype imputation in African populations. Genet Epidemiol. 2011;35(8):766–80. doi:10.1002/gepi.20626.
27. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7(1):16. doi:10.1186/s13073-015-0138-2.
28. Francioli L, Menelaou A, Pulit S, van D F, Palamara P, Elbers C, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46(8):818–25. doi:10.1038/ng.3021.
29. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.
30. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695–701. doi:10.1038/ng.f.136.
31. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779.
32. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, 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. doi:10.1038/ng.357.
33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. doi:10.1101/gr.107524.110.
34. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–81. doi:10.1038/nmeth.1785.
35. 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. doi:10.1038/ng.2354.