




# Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels

Wei Zhou<sup>1</sup>  | Lars G. Fritsche<sup>2,3</sup> | Sayantan Das<sup>3</sup> | He Zhang<sup>4</sup> | Jonas B. Nielsen<sup>4</sup>  |  
Oddgeir L. Holmen<sup>2,5</sup> | Jin Chen<sup>4</sup> | Maoxuan Lin<sup>4</sup> | Maiken B. Elvestad<sup>2</sup> |  
Kristian Hveem<sup>6,7</sup> | Goncalo R. Abecasis<sup>3</sup> | Hyun Min Kang<sup>3†,\*</sup> | Cristen J. Willer<sup>1,4,8†</sup> 

<sup>1</sup>Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America

<sup>2</sup>K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway

<sup>3</sup>Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America

<sup>4</sup>Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

<sup>5</sup>St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway

<sup>6</sup>HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway

<sup>7</sup>Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, Levanger, Norway

<sup>8</sup>Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

## Correspondence

Cristen J. Willer, Department of Computational Medicine and Bioinformatics, Department of Internal Medicine, Division of Cardiology, and Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA.

Email: cristen@umich.edu

\* Additional corresponding author

Hyun Min Kang, Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA.

Email: hmkang@umich.edu

† These authors contributed equally to this work.

## ABSTRACT

The accuracy of genotype imputation depends upon two factors: the sample size of the reference panel and the genetic similarity between the reference panel and the target samples. When multiple reference panels are not consented to combine together, it is unclear how to combine the imputation results to optimize the power of genetic association studies. We compared the accuracy of 9,265 Norwegian genomes imputed from three reference panels—1000 Genomes phase 3 (1000G), Haplotype Reference Consortium (HRC), and a reference panel containing 2,201 Norwegian participants from the population-based Nord Trøndelag Health Study (HUNT) from low-pass genome sequencing. We observed that the population-matched reference panel allowed for imputation of more population-specific variants with lower frequency (minor allele frequency (MAF) between 0.05% and 0.5%). The overall imputation accuracy from the population-specific panel was substantially higher than 1000G and was comparable with HRC, despite HRC being 15-fold larger. These results recapitulate the value of population-specific reference panels for genotype imputation. We also evaluated different strategies to utilize multiple sets of imputed genotypes to increase the power of association studies. We observed that testing association for all variants imputed from any panel results in higher power to detect association than the alternative strategy of including only one version of each genetic variant, selected for having the highest imputation quality metric. This was particularly true for lower frequency variants (MAF < 1%), even after adjusting for the additional multiple testing burden.

## KEYWORDS

genotype imputation, GWAS, multiple reference panels, population-specific, study power

## 1 | INTRODUCTION

Many novel disease-associated signals for a wide variety of diseases and traits have been successfully identified using imputation-based meta-analyses (Cheng et al., 2016; Cooper et al., 2008; De Jager et al., 2009; Ge et al., 2016; Horikoshi et al., 2015; Houlston et al., 2008; Jin et al., 2016; Loos et al., 2008; Ruth et al., 2015; Zeggini et al., 2007, 2008). Genotype imputation is the process of inferring missing genotypes in study samples using a reference panel of high-density haplotypes (Li, Willer, Sanna, & Abecasis, 2009). Imputation allows variants that are not directly genotyped to be studied without other costs than computation. Previous simulations showed that imputation substantially increases the power of association studies to detect causal loci (Marchini & Howie, 2010; Spencer, Su, Donnelly, & Marchini, 2009). Imputation-based genome-wide association studies (GWAS) have successfully identified novel signals that were undetected in chip-based studies. For example, two disease-associated signals were detected in the 1000G-based imputation (Auton et al., 2015) for the Wellcome Trust Case Control Consortium phase 1 Data (WTCCC), which were missed in the original WTCCC GWAS study that was performed 4 years before (Burton et al., 2007; Huang, Ellinghaus, Franke, Howie, & Li, 2012). Imputation also facilitates fine-mapping studies by allowing most polymorphic variants, including causative ones, to be tested in known disease-associated loci. For example, the strongest association signal, observed at the imputed variant rs7903146 of the *TCF7L2* locus in the WTCCC type 2 diabetes scan, is suggested to be the (causal association) in the locus (Mahajan et al., 2014; Marchini, Howie, Myers, McVean, & Donnelly, 2007). Furthermore, imputation allows for meta-analysis between samples that have been genotyped using different arrays, increasing power.

For studies that have access to population-matched genome sequenced individuals, there is uncertainty in deciding between a smaller, ancestry-matched reference panel and a larger publicly available cosmopolitan reference panel. An ideal reference panel is expected to have closely matched ancestry to study samples because the genetic similarity increases the accuracy of imputation (Deelen et al., 2014; Huang et al., 2015; Huang & Tseng, 2014; Low-Kam et al., 2016; Mitt et al., 2017; Okada, Momozawa, Ashikawa, Kanai, & Matsuda, 2015; Pistis et al., 2015; Roshyara & Scholz, 2015; Walter et al., 2015). On the other hand, the imputation accuracy increases when larger reference panels are used, especially for lower frequency variants (Browning

& Browning, 2009; Howie, Donnelly, & Marchini, 2009; Huang et al., 2009; Li et al., 2009; Roshyara & Scholz, 2015).

Furthermore, different whole-genome reference panels may generate discordant imputed genotypes for the same variants in the same study samples. This brings in challenges for the follow-up association tests. The optimal strategy to perform association tests using genotypes imputed by different reference panels remains unclear. IMPUTE2 provides one possible approach to merge all reference panels to a single larger panel for genotype imputation when multiple reference panels are available (Howie et al., 2009), which may avoid the problem that different versions of genotypes are imputed for the same variants. The Genome of the Netherlands Consortium and the UK10K study has further shown that the combined reference panel of 1000G and the population-specific reference resulted in better imputation results compared to the two individual panels for rare variants (Deelen et al., 2014; Huang et al., 2015). However, this approach is not feasible when individual-level haplotypes within the reference panel are not accessible, as is the case with the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016), primarily due to ethical issues surrounding sharing of individual-level genetic data (McCarthy et al., 2016).

Here, we genotyped 9,265 Norwegian participants from the HUNT study (Krokstad et al., 2013) for 350,270 polymorphic autosomal variants using the Illumina Human CoreExome array with approximately 240,000 GWAS tagging markers. We created a population-matched reference panel by whole-genome sequencing (WGS) 2,021 individuals from the HUNT study to a mean depth of 5 $\times$ . We imputed variants from the HUNT WGS reference panel as our ethnically matched panel. We also performed imputation with two additional imputation reference panels: the HRC (McCarthy et al., 2016) and 1000G phase 3 (Auton et al., 2015). First, we systematically evaluated and compared the imputation results from the three reference panels, including the number of successfully imputed variants as well as the imputation accuracy. Next, we evaluated and compared the power of association tests between two approaches to incorporate multiple versions of imputed genotypes. First is the “best  $R_{sq}$ ” approach, which retains imputed genotypes only from the panel with highest imputation quality metrics for each variant. Second is the “best  $P$ -value” approach that tests association with all imputed genotypes and uses the most significant association  $P$  value, adjusting for the additional variants tested.

## 2 | MATERIALS AND METHODS

### 2.1 | Array-based genotyping

9,265 samples from the HUNT Biobank in Norway were genotyped at 350,270 polymorphic autosomal variants using an Exome + GWAS chip array (HumanCoreExome-12 v1.0, Illumina, Inc., San Diego, CA). Genotype calling was performed using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina, Inc., San Diego, CA). Samples with <98% genotype calls ( $N = 37$ ), evidence of gender discrepancy ( $N = 21$ ), duplicates ( $N = 66$ ) as well as individuals with non-Norwegian ancestry identified by plotting the first 10 genotype-driven principal components (Jolliffe, 1986) ( $N = 7$ ) were excluded from further analysis ( $N = 131$ , 1.19%). As Supporting Information Fig. S1 shows, the HUNT GWAS samples have similar ancestry to the samples in the HUNT WGS reference panel. All HUNT research subjects provided informed written consent and IRB approval was obtained for genetic studies.

Relatedness was evaluated based on the estimation of the proportion of identity by descent (IBD) by PLINK (Purcell et al., 2007). We excluded 1,644 samples from the HUNT GWAS sample due to first- or second-degree relatedness to samples in HUNT WGS, defined as  $IBD \geq 0.25$ . We excluded samples that were related to samples within the reference panel to avoid inflating imputation statistics for regions inherited IBD. We performed variant-level quality control by excluding 19,872 variants that met any of the following criteria; variants with a cluster separation score < 0.3 reported by GenomeStudio V2011.1 (Illumina, Inc., San Diego, CA), <95% genotype call rate, or deviation from Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-5}$ ).

### 2.2 | Genotype imputation

Genotype imputation with the 1000G phase 3 (Auton et al., 2015) and the HRC (McCarthy et al., 2016) reference panels was conducted using the Michigan Imputation Server (Das et al., 2016) and imputation with the HUNT WGS reference panel was conducted using a local server. The study samples were phased using SHAPEIT2 (v2.r790) (Delaneau, Zagury, & Marchini, 2013) followed by imputation using minimac3 (v2.0.1) (Fuchsberger, Abecasis, & Hinds, 2015; Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). Two imputation metrics output by minimac3 were used for evaluating the imputation quality: ImpRsq and EmpRsq. ImpRsq is previously known as  $\hat{r}^2$  in different versions of MaCH/minimac (Fuchsberger et al., 2015; Howie et al., 2012; Li, Willer, Ding, Scheet, & Abecasis, 2010). ImpRsq is defined for both genotyped and ungenotyped variants in the chip array as an estimate of the squared correlation between imputed dosages and true, unobserved

genotypes, calculated as the observed variance over the expected variance. EmpRsq is defined only for genotyped variants in the chip array as the squared correlation between leave-one-out imputed dosages and the true, observed genotypes (see “Estimated Imputation Accuracy” section at [https://genome.sph.umich.edu/wiki/Minimac\\_Diagnostics](https://genome.sph.umich.edu/wiki/Minimac_Diagnostics) for details).

### 2.3 | Reference panels

The HUNT WGS reference panel contains 1,101 earliest onset cases with myocardial infarction and 1,100 age and sex matched controls that were selected from the HUNT study (Krokstad et al., 2013). WGS to  $\sim 5\times$  depth was performed on either Illumina HiSeq 2000 or 2500. We followed the Got-Cloud SNP calling pipeline to process the WGS data (Jun, Wing, Abecasis, & Kang, 2015). The variant sites and genotype likelihood were called using SAMtools (Li et al., 2009) and the genotypes for SNPs were refined and phased using Beagle v4 (Browning & Browning, 2013). After quality control, 20.2 million single nucleotide variants were retained in 2,201 samples, of which four million were unique to our study; not observed in dbSNP 144 (Sherry et al., 2001), 1000 Genomes phase 3 (Auton et al., 2015), UK10K (Walter et al., 2015), ESP6500 (*NHLBI GO Exome Sequencing Project (ESP)*, August 2016 accessed), or ExAC.r0.3 (Lek et al., 2016) (Table 1). The individuals in the HUNT WGS panel have similar ancestry to the HUNT study samples (Supporting Information Fig. S1) and are from the same geographic region, although we excluded in the genotyped samples any first- or second-degree relatives of the sequenced samples to avoid biased estimates of the accuracy of imputation. Additionally, there were no close relatives within the sequenced samples. The other two reference panels that we used for genotype imputation are the 1000 Genomes phase 3 (1000G) (Auton et al., 2015) and the HRC release 1 (McCarthy et al., 2016) containing 32,488 individuals, both of which are pre-loaded in the Michigan Imputation Server (Das et al., 2016) (Table 2). The HUNT cohort contributed an early freeze of WGS data consisting of 1,023 samples to the HRC consortium. Thus, the HUNT WGS and the HRC reference panels have 1,023 samples in common. Variants with minor allele counts (MAC) less than five were excluded from HRC (McCarthy et al., 2016).

### 2.4 | Permutation test

To determine the genome-wide significance thresholds for association tests using the two approaches to incorporate imputed genotypes, we performed permutation tests. The measurements of the high-density lipoprotein (HDL) cholesterol for the study samples were permuted 1,000 times. Each permutation was followed by a genome-wide association test (GWAS) using the permuted phenotypes. The most significant

**TABLE 1** Summary of the variants in the HUNT WGS reference panel containing 2,201 individuals with average sequencing depth 5×

Variant type	Total number of variants	Mean number of variants per individual (SD)	Mean number of unique variants per individual (SD)	1000 Genomes (%)	Number of novel variants <sup>a</sup>
Splice	1,265	71.5(4.6)	0.2(0.47)	36.6	355
Nonsense	2,432	71.5(6)	0.43(0.74)	36.6	585
Missense	113,576	9,480(113)	13.8(13.6)	56.3	13,927
Synonymous	77,699	10,707(100)	7.1(7.5)	68.5	5,935
Noncoding	20,050,237	3,342,839(15,415)	1531(906)	68.7	4,030,199
Total	20,245,209	3,363,168(15,522)	1,552(919)	68.6	4,051,001

<sup>a</sup>Novel: not reported in dbSNP 144 (Sherry et al., 2001), 1000 Genomes phase 3 (Auton et al., 2015), UK10K (Walter et al., 2015), ESP6500 (*NHLBI GO Exome Sequencing Project (ESP)*, August 2016 accessed), or ExAC.r0.3 (Lek et al., 2016). SD, standard deviation

**TABLE 2** Reference panels used for genotype imputation

Reference panels	Variants	Sample size	Population
Haplotype Reference Consortium (McCarthy et al., 2016) (HRC)	39 million SNPs (MAC $\geq$ 5)	32,488 <sup>a</sup>	Cosmopolitan (mostly European)
1000 Genomes phase 3 version 5 (Auton et al., 2015) (mean depth < 8×)	81 million biallelic SNPs, indels, deletions, complex short substitutions, and other structural variant classes (MAC $\geq$ 2)	2,504	Cosmopolitan
HUNT whole-genome sequencing (HUNT WGS) (mean depth $\sim$ 5×)	20 million SNPs	2,201 <sup>a</sup>	Norwegian

<sup>a</sup>HRC and HUNT WGS data set have 1,023 samples in overlap. MAC, minor allele count.

*P*-values from each of the 1,000 GWAS were ranked. The significance threshold with family-wise error rate (FWER)  $n/1000$  equals to the  $n$ th smallest *P*-value. Because the “best *P*-value” approach tests more variants, it will have a more stringent significance threshold than the “best Rsq” approach.

## 2.5 | Power estimation

In order to estimate the power to detect association under the two approaches to incorporate imputed genotypes from multiple reference panels, we considered directly genotyped variants as causal variants, and used multiple sets of imputed genotypes to evaluate the power. First, we obtained the leave-one-variant-out imputed dosages for those directly genotyped variants. The official release of minimac3 performs leave-one-out hidden Markov model (HMM) calculation internally to calculate leave-one-out Rsq summary statistics, but does not output individual dosages (Fuchsberger et al., 2015; Howie et al., 2012). We modified minimac3 to include the individual leave-one-out dosages in the output VCF for the genotyped variants. Second, we simulated phenotypes based on the genotypes obtained by the chip array. Finally, we evaluated the power of the two approaches by performing association tests between the simulated phenotypes and the imputed dosages based on either “best Rsq” or “best *P*-value” approaches.

The details of simulation follow the steps described below:

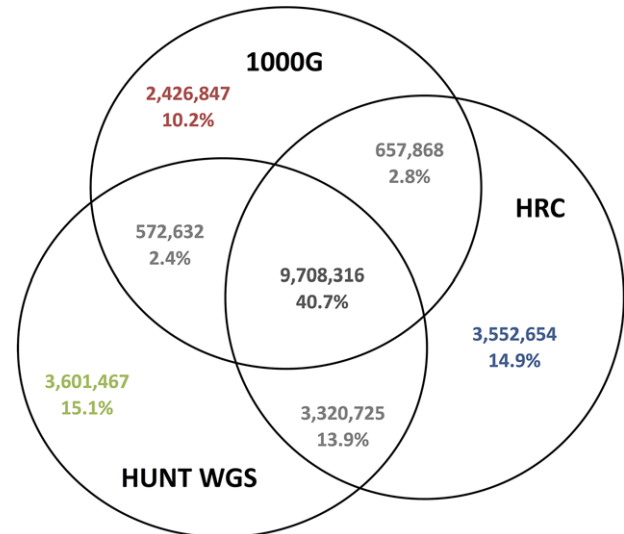
1. Select the noncentrality parameter corresponding to the association test *P*-value  $p_t$ . We calculate the noncentrality parameter  $Nr^2$  as a chi-square statistics corresponding to the upper-tail probability  $p_t$ , where  $N$  is the total number of study subjects. This ensures that the median *P*-value is  $p_t$  when the true phenotypic variance explained by the genotype is  $r^2$ .
2. For each variant, we randomly draw  $\epsilon$  from the normal distribution with mean 0 and standard deviation  $\sqrt{1 - r^2}$ . We calculate the effect size  $\beta$  as  $\sqrt{r^2} / 2f(1 - f)$ , where  $f$  is the minor allele frequency (MAF) estimated using the chip genotypes of the variant. The phenotype value  $y$  is then calculated as  $G\beta + \epsilon$ , where the chip genotypes  $G$  is 0, 1, or 2. The phenotypic variance explained by  $G$  and  $\epsilon$  will be  $r^2$  and  $1 - r^2$ , respectively.
3. We perform the linear regression using the leave-one-variant-out dosages for this variant, which were imputed using the three different reference panels, respectively, and the phenotype  $y$ .

- For the “best  $P$ -value” approach, the final association  $P$  value equals to the most significant one among the three  $P$  values associated with the three different versions of imputed dosages. With the “best  $R_{sq}$ ” approach, the final  $P$  value equals to the one corresponding to the reference panel with the highest imputation quality (ImpRsq), an estimated value for the correlation between imputed genotypes and true, unobserved genotypes.
- The power to detect association signals equals to the percentage of final  $P$  values exceeding the genome-wide significance threshold determined for each approach by the permutation tests described above.

We performed linkage disequilibrium (LD) based variant pruning for the 289,376 directly genotyped variants that were found by all three reference panels using PLINK (Purcell et al., 2007) and obtained 132,183 variants with LD  $r^2 < 0.2$  among each other. Then, we randomly selected 3,000 variants for each of the MAF categories:  $MAF \leq 0.001$ ,  $MAF > 0.001$  and  $\leq 0.01$ ,  $MAF > 0.01$  and  $\leq 0.05$ , and  $MAF > 0.05$ . We applied  $ImpRsq > 0.3$ ,  $0.5$ , and  $0.8$  to remove poorly imputed genotypes. Variants that were successfully imputed from at least two references were used for this simulation study. All five steps above were repeated given different  $p_t$ s ranging from  $5 \times 10^{-8}$  to  $1 \times 10^{-13}$ . Additionally, the entire process was repeated five times across the selected variants to average power.

## 2.6 | Partial correlation estimation

To quantify the net gain of imputation accuracy obtained by including another reference panel on top of an existing panel, we estimated the partial correlation between the leave-one-out imputed dosages from the additional panel and the chip genotypes, conditioned on the leave-one-out imputed dosages from the existing panel. The correlation has been estimated for every pair of reference panels among the three on each of the 289,376 genotyped variants that were found in all three panels. For example, to estimate the net gain of including 1000G panel on top of HUNT panel (PartialRsq [1000G,Chip | HUNT]), we first obtained the leave-one-out dosages based on 1000G and HUNT WGS (details described in Section 2.5). Secondly, for each variant, we performed three linear regressions on the chip genotypes: the first one has the imputed dosages from 1000G and HUNT WGS as covariates (model 1), the second one has the imputed dosages from HUNT WGS only as a covariate (model 2), and the third one does not have any other covariate except for the intercept (model 3). Lastly, we obtained sum of squared residuals (SSR) for the three linear regressions and calculated the partial correlation (partial Rsq) as  $\frac{SSR_{model2} - SSR_{model1}}{SSR_{model3}}$ . In a similar notation, the EmpRsq is equivalent to  $\frac{SSR_{model3} - SSR_{model2}}{SSR_{model3}}$ ,



**FIGURE 1** Number of variants that were imputed by different reference panels. The corresponding percentage is the variants number out of all 23.8 million variants that were successfully imputed by any of the three reference panels

and their sum should be equivalent to the proportion of explained variance by both sets of imputed dosages. Our intuition is that the more extra information the additional reference panel provides, the higher the partial correlation will be.

## 3 | RESULTS

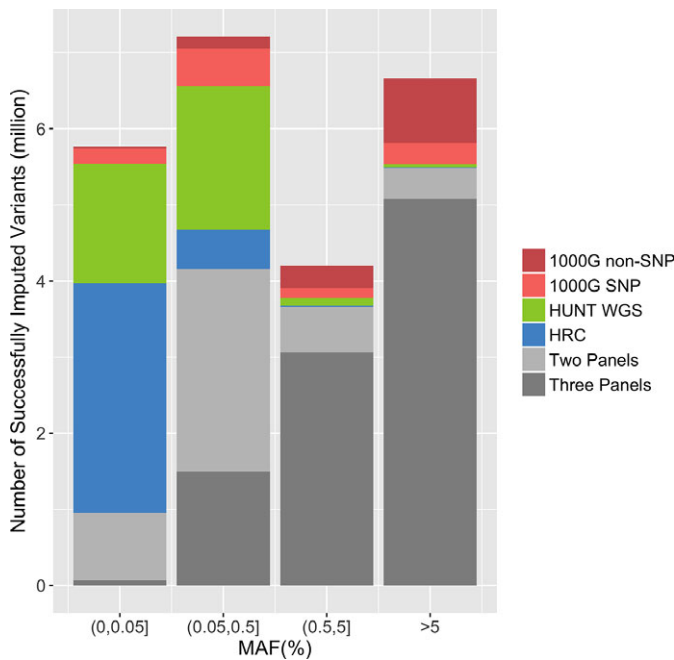
### 3.1 | Evaluating successfully imputed variants using different reference panels

In total, ~23.8 million variants were successfully imputed using minimac3 (Fuchsberger et al., 2015; Howie et al., 2012) from at least one of the three reference panels and exceeded the threshold of estimated imputation quality ( $ImpRsq \geq 0.3$ ) (Figure 1). The three reference panels yielded roughly equal number of SNPs with MAF more than 1%, but the 1000G uncovered more unique variants; approximately 75.3% (1,068,228 out of 1,418,417) that were uniquely imputed from 1000G are indels or structural variants, a category of variation that is not available in the other two reference panels. We observed that imputation from the HRC panel resulted in more extremely rare variants (MAF less than 0.05%) than from HUNT WGS and 1000G. Imputation from the HUNT WGS panel uncovered more variants with MAF between 0.05% and 0.5% than the other two reference panels (Table 3). Approximately, 3.6 million variants were uniquely imputed by the HUNT WGS panel (Figure 1) and the majority of them have MAF less than or equal to 0.05% (Figure 2). A threshold  $\geq 0.3$  for ImpRsq was applied as recommended to remove most of poorly imputed variants while retaining the vast majority of

**TABLE 3** Numbers of imputed variants contributed by each reference panel categorized by MAF

MAF	HRC release 1 (39.2 M SNPs, 32,488 samples including 1,203 HUNT samples)			1000G phase3 v5 (81.2 M markers, 2,504 samples)		HUNT 5 × WGS (20.2 M SNPs, 2,201 samples)			
	Number of passed variants	Number of passed variants	Number of uniquely imputed variants	Number of passed variants	Percent of passed variants	Number of uniquely imputed variants	Number of passed variants	Percent of passed variants	Number of uniquely imputed variants
(0, 0.0005]	4,337,138	23.9%	<b>3,009,729</b>	567,481	2.4%	230,186	2,291,216	50.6	1,570,259
(0.0005, 0.001]	1,339,096	91.1%	373,964	501,248	11.4%	176,252	1,668,837	94.4	<b>901,106</b>
(0.001, 0.005]	2,964,988	97.5%	140,318	2,119,956	33.6%	475,376	3,917,801	98.0	<b>982,320</b>
(0.005, 0.01]	1,125,181	99.2%	7,426	1,074,885	68.9%	<b>126,616</b>	1,279,200	98.6	47,426
(0.01, 0.05]	2,314,490	99.6%	10,525	2,554,206	89.2%	<b>295,991</b>	2,538,140	99.1	55,490
> 0.05	5,158,670	99.8%	10,692	6,547,887	98.1%	<b>1,122,426</b>	5,507,946	99.6	44,866
Total	17,239,563	55.1%	3,552,654	13,365,663	29.5%	2,426,847	17,203,140	87.4	3,601,467

The threshold  $\text{ImpRs}q \geq 0.3$  was applied. Each reference panel contributed uniquely imputed variants. The greatest number of the uniquely imputed variants among the three reference panels for variants in each MAF category is highlighted in bold. MAF, minor allele frequency;  $\text{ImpRs}q$ , imputation quality metric  $R^2$ .

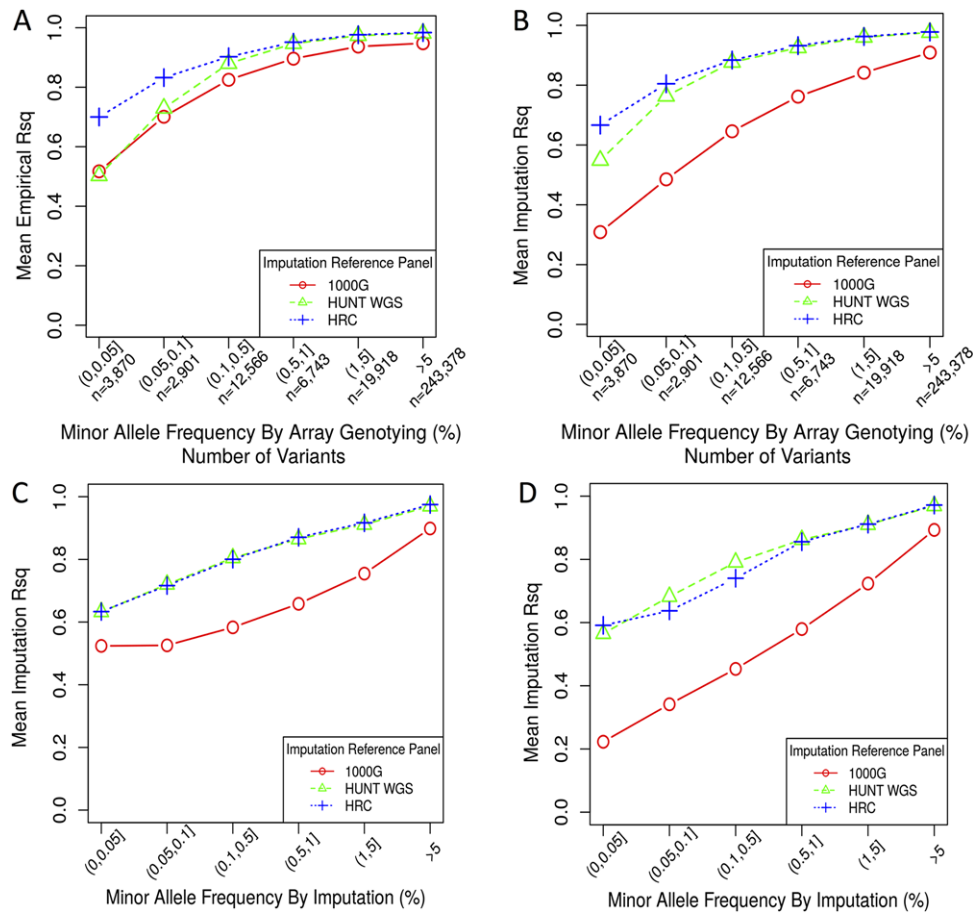


**FIGURE 2** Distribution of numbers of variants that were imputed from only one reference panel or from multiple reference panels in different MAF categories. Variants that were imputed by 1000G only are categorized as SNPs and non-SNP variants, including indels, deletions, complex short substitutions, and other structural variant classes. 1000G, 1000 Genomes phase 3; WGS, whole-genome sequencing; HRC, Haplotype Reference Consortium; MAF, minor allele frequency

well imputed SNPs (Li et al., 2009). We observed that the average  $\text{EmpRs}q$  remained above 0.6 for all MAF categories from all three reference panels when the  $\text{ImpRs}q \geq 0.3$  threshold was applied (Supporting Information Fig. S2).

### 3.2 | Comparing imputation accuracy from different reference panels

To compare the imputation accuracy across the three reference panels, we examined all 289,376 variants that were directly genotyped by the chip array and available in all three reference panels. “Leave-one-variant-out” imputation results were used for these directly genotyped variants, meaning that one by one, each genotyped variant was masked, imputed, and then compared to the directly genotyped calls. The  $\text{EmpRs}q$  was estimated for each genotyped variant from each panel, which is the squared Pearson correlation between the imputed allele dosages and the genotypes called by direct genotyping. Figure 3A compares the average  $\text{EmpRs}q$  for all genotyped variants categorized by MAF among different reference panels. The MAF is estimated using the genotypes called by the chip array. Imputation from HRC has higher imputation accuracy for rare variants with  $\text{MAF} < 0.5\%$  than the other two reference panels, which is expected because the number of samples available in HRC is much larger than the other two panels and the imputation accuracy for extremely rare variants depends on the number of copies of alternate alleles (Roshyara & Scholz, 2015). What is unexpected is that for variants with  $\text{MAF} \geq 0.5\%$ , HRC and HUNT WGS panels show comparable imputation accuracy, even though the size of the HUNT WGS panel is 15 times smaller than HRC. Consistent to previous studies, this result demonstrated the value of WGS for ancestry-matched samples as a reference panel for genotype imputation (Deelen et al., 2014; Huang et al., 2015; Huang & Tseng, 2014; Low-Kam et al., 2016; Okada et al., 2015; Pistis et al., 2015; Roshyara & Scholz, 2015; Walter et al., 2015). It is also noticed that imputation from 1000G has lower average  $\text{ImpRs}q$  than the other two reference panels (Figure 3B–D), which is consistent to the lower proportion of variants passing



**FIGURE 3** HRC and HUNT WGS panels show comparable imputation quality. (A) Comparing the mean empirical  $R^2$  (y-axis) reported by different reference panels for variants that were directly genotyped categorized by the MAF (x-axis) without any  $\text{ImpRsq}$  threshold applied. (B) Comparing the mean Imputation  $R^2$  (y-axis) reported by different reference panels for variants that were directly genotyped categorized by the MAF (x-axis) without any  $\text{ImpRsq}$  threshold applied. (C) Comparing the mean Imputation  $R^2$  (y-axis) reported by different reference panels for all imputed variants ( $\text{ImpRsq} > 0.3$ ) by the MAF (x-axis). (D) Comparing the mean Imputation  $R^2$  (y-axis) reported by different reference panels for all imputed variants by the MAF (x-axis) without any  $\text{ImpRsq}$  threshold applied. 1000G, 1000 Genomes phase 3; WGS, whole-genome sequencing; HRC, Haplotype Reference Consortium; MAF, minor allele frequency;  $\text{ImpRsq}$ , imputation quality metric  $R^2$

the various  $\text{ImpRsq}$  thresholds in 1000G (Supporting Information Fig. S2).

To further evaluate the impact of the sample size of the HUNT WGS panel on the imputation accuracy, we have randomly drawn 500, 1,000, and 1,500 samples from the original HUNT reference panel for imputation. Figure S3 shows the comparison of the average  $\text{EmpRsq}$  for all genotyped variants categorized by MAF among the target samples, across all reference panels. As expected, increases in the sample size of the HUNT WGS reference panels resulted in higher imputation accuracy, particularly for less frequent variants with  $\text{MAF} < 0.5\%$ . Interestingly, we observed that the HUNT WGS with 500 samples outperforms 1000G (Auton et al., 2015) for variants with  $\text{MAF} > 0.5\%$ . These results are consistent with other studies with population-specific reference panels (Mitt et al., 2017; Pistis et al., 2015). The subset of 1,000 samples provides better imputation accuracy than 1000G (Auton et al., 2015) even for variants with  $\text{MAF}$  as low as 0.1% and

comparable imputation accuracy to HRC (McCarthy et al., 2016) for variants with  $\text{MAF} > 0.5\%$ .

We examined whether our evaluation of imputation accuracy is biased in favor of HUNT WGS due to relatedness. Previous studies have shown that the relatedness between study samples and reference samples increases genotype imputation efficiency because related individuals tend to share longer haplotype stretches than unrelated ones (Huang & Tseng, 2014). To avoid the bias of imputation accuracy due to the relatedness between our study samples and the samples in the HUNT WGS reference panel, we excluded 1,644 study samples who are up to second-degree relatives of HUNT WGS samples. Relatedness was based on the estimation of the proportion of IBD by PLINK (Purcell et al., 2007). We observed that excluding these study samples did not affect the imputation accuracy except causing a slight decrease of the imputation accuracy for those very rare variants with  $\text{MAF} < 0.05\%$  (Supporting Information Fig. S4).

### 3.3 | Evaluating two possible association test strategies to use multiple sets of imputed genotypes

As Figure 1 shows, approximately 60% of all successfully imputed variants were imputed from more than one reference panel, which makes it unclear how to perform downstream association tests. We compared two possible strategies: the “best  $P$ -value” and the “best Rsq” approaches. The “best  $P$ -value” approach uses each version of imputed genotypes to choose the lowest association  $P$ -value, thereby increasing the burden of adjusting for multiple hypothesis testing. The “best Rsq” approach selects the imputed variant with the highest estimated imputation quality ImpRsq, which is expected to be a reasonable approximation of the association between imputed and true genotypes, especially for common variants (Supporting Information Fig. S5). We have compared the power of the two approaches to detect association signals accounting for the fact that the “best  $P$ -value” approach needs adjusting for the additional variants tested. To determine the significance thresholds for association tests with a FWER 0.05, we estimated the number of independent tests using 1,000 permutations. For the “best Rsq” approach, where fewer “variants” are analyzed, the significance threshold is  $4.69 \times 10^{-9}$  ( $2.10 \times 10^{-9}$  with a Bonferroni correction) and for the best  $P$ -value approach, it is  $2.53 \times 10^{-9}$  ( $1.05 \times 10^{-9}$  with a Bonferroni correction).

Using the permutation-derived significance thresholds above, we evaluated the power of the two approaches for association tests with quantitative traits through a simulation study (details described in Section 2). Our results indicated that the “best  $P$ -value” approach has more power to detect association signals than the “best Rsq” approach, particularly for rare variants with  $MAF < 1\%$ , no matter how stringent the ImpRsq threshold was used for filtering out the poorly imputed genotypes (Figure 4, Supporting Information Figure S6 and Supporting Information Table S1). This is probably because the estimated imputation quality ImpRsq does not always agree with empirical imputation quality EmpRsq especially for rare variants (Supporting Information Figure S5), resulting in loss of variants with highest empirical imputation quality when selecting the “best Rsq” strategy. In addition, the distributions of the ImpRsq are quite different from different panels. Notably, from 1000G the ImpRsq and EmpRsq were substantially lower for low-frequency variants ( $0.5\% < MAF < 5\%$ ), and ImpRsq tends to underestimate EmpRsq (Supporting Information Fig. S5). The two approaches have comparable association power for variants with  $MAF \geq 1\%$ , where estimated and empirical imputation qualities highly agree with each other (Supporting Information Fig. S5). Our observation suggests that the inaccurate prediction of imputation quality have a higher impact than increased burden of multiple testing in association test with rare variants.

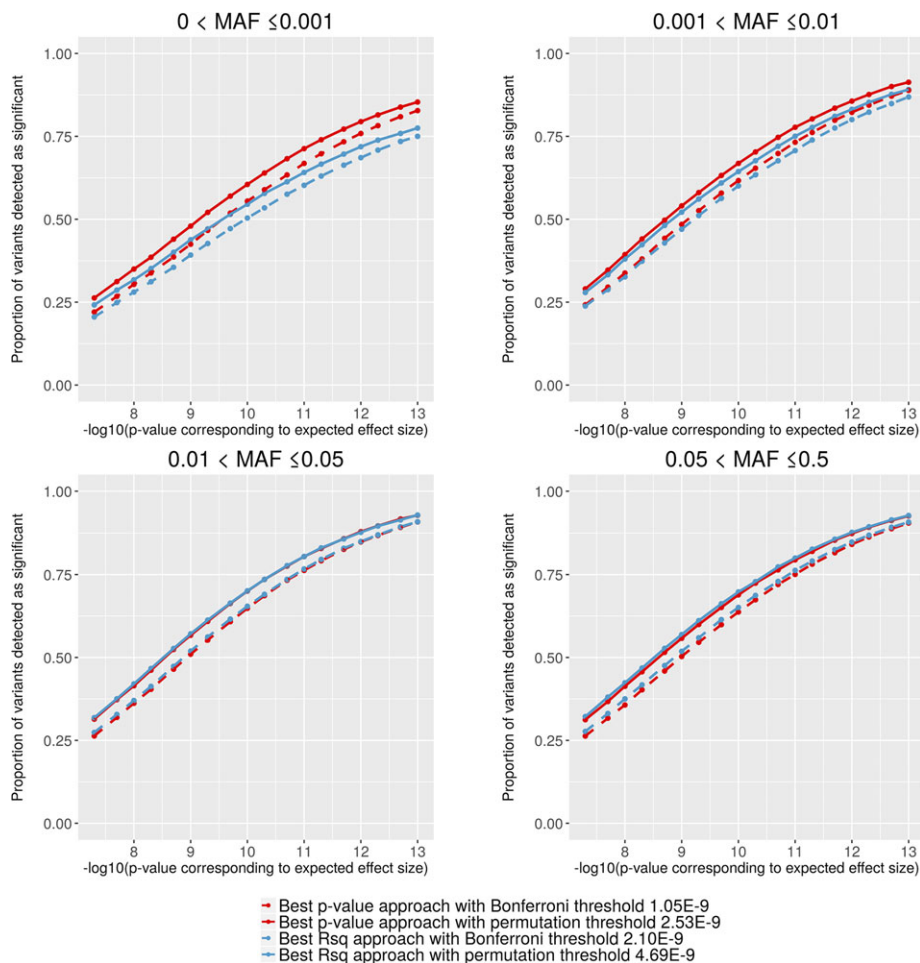
### 3.4 | Evaluating net gain of imputation accuracy by including an additional reference panel

Finally, we quantified the net gain of imputation accuracy by including an additional reference panel as a “partial Rsq” conditioned on the imputed genotypes from an existing reference panel (see Section 2 for details). Intuitively, this represents the difference between the “optimal EmpRsq” linearly combined between two sets of imputed genotypes and the EmpRsq from the original imputed genotypes. The 289,376 genotyped variants that were found in all three panels were used to evaluate the additional information that were gained from one reference panel given imputed dosages based on another panel. As Supporting Information Fig. S7 presents, each reference panel is able to provide additional information to improve imputation accuracy. However, relatively less information could be gained by including 1000G on top of HRC across all MAF categories. This is expected because 1000G samples are included in the HRC panel, with the caveat that only single nucleotide variants with  $MAC \geq 5$  were retained. Note that evaluation of indels and structural variants absent in HRC were not included in this experiment. In contrast, given the imputed dosages from 1000G, both HUNT WGS and HRC provide substantial net gain of imputation accuracy, which is consistent to our observations. Furthermore, HUNT WGS and HRC provide additional information conditional on each other. More specifically, more extra information was obtained from HRC given HUNT WGS than those were obtained from HUNT WGS given HRC for these genotyped variants, which is also consistent to our observations in Figure 3.

## 4 | DISCUSSION

Many studies have performed WGS of a subset of samples followed by imputation into samples with GWAS data (Gudbjartsson et al., 2015; Lane, Vlasac, & Anderson, 2016; Nalls et al., 2014; van Leeuwen et al., 2016). However, the trade-offs between the panel size, imputable variant types, and population specificity across different reference panels make it challenging to decide on the optimal strategy for imputation and downstream association analysis. We evaluated methods for genotype imputation when different reference panels are available. Our findings have demonstrated the benefits of uncovering novel variants with low frequency by using population-specific reference panels as has been reported by previous studies (Huang et al., 2015). Because the population-specific HUNT panel shared 1,023 samples with HRC (McCarthy et al., 2016), we expect to see an even bigger advantage in the number of novel low-frequency variants imputed by the population-specific panel if there were no overlap between the two reference panels.





**FIGURE 4** Comparison of power to detect true associations between best  $P$ -value and best  $R_{sq}$  approaches via simulation studies. For each MAF category, 3,000 directly genotyped variants were randomly selected based on their MAF estimated with genotypes obtained from the chip array to estimate the power. The power was calculated as the proportion of significantly associated variants across three imputed panels based on each strategy given the corresponding significance threshold.  $\text{ImpRsq} \geq 0.3$  was applied to remove poorly imputed genotypes. The numbers of variants that were successfully imputed from at least two reference panels and used in the simulation studies are: 2,513 with  $\text{MAF} > 0$  and  $\leq 0.001$ ; 2,989 with  $\text{MAF} > 0.001$  and  $\leq 0.01$ ; 3,000 with  $\text{MAF} > 0.01$  and  $\leq 0.05$ ; and 3,000 with  $\text{MAF} > 0.05$ . MAF, minor allele frequency;  $\text{ImpRsq}$ , imputation quality metric  $R^2$

We have also observed that large-scale publicly available reference panels, as exemplified by HRC (McCarthy et al., 2016) and 1000G (Auton et al., 2015), contribute a large number of variants that are not captured by population-specific reference panels. More specifically, HRC (McCarthy et al., 2016), which has much larger sample size and contains more general European populations, contributes 3.5 million variants that could not be imputed by the other two panels. Because 1000G (Auton et al., 2015) has additional advantages that indels and structural variants are comprehensively detected and genotyped, 1.3 million non-SNP variants have only been imputed by 1000G. Furthermore, each reference panel may provide additional information to improve imputation accuracy. Therefore, to increase the variant coverage and imputation accuracy as much as possible, we recommend using all three reference panels for imputation if available. If a single panel has to be chosen, each option will have different

advantages and disadvantages. We have shown that imputation from population-specific reference panels provides comparable imputation accuracy for variants with  $\text{MAF} > 0.1\%$  as using reference panels with 15 times larger sample size with only broad ancestry matching (i.e., European). Although panel sizes are similar, the population-specific reference panel results in higher imputation accuracy than the mixed-ancestry 1000G panel (Auton et al., 2015) for variants with  $\text{MAF} \geq 0.05\%$ . This has also been observed by a recently published study on Estonians (Mitt et al., 2017).

To address the issue of imputing different versions of the same variant from different reference panels, we propose the “best  $P$ -value” approach, which analyzes all versions of each imputed variant and accounts for the multiple testing. Our simulation study demonstrated that this approach has higher power for detecting association signals than selecting the imputed variant with highest imputation quality given the

distributions of the imputation quality metrics from different reference panels may be quite different, even adjusting for additional variants tested.

The UK10K study and the Genome of the Netherlands (GoNL) Consortium suggested that merging multiple reference panels to a larger reference panel would improve imputation performance, especially for less frequent variants (Deelen et al., 2014; Huang et al., 2015). Compared to this approach, our “best *P*-value” approach does not require access to all reference panels and is feasible even if not all reference panel haplotypes are directly accessible. If large imputation reference panels, such as the HRC (McCarthy et al., 2016), are not directly accessible, conducting association tests for all imputed versions of genotype with slightly higher computational cost will be an effective strategy.

In summary, we recommend creating a small size ancestry-matched reference panel using WGS to allow for improved imputation of low-frequency variants that may be enriched in that ancestral group, performing genotype imputation using the ancestry-matched reference panel and other large publicly available databases, and analyzing all versions of imputed variants in downstream association testing.

## ACKNOWLEDGMENTS

The whole genome sequencing of 2,201 HUNT samples was supported by NHLBI HL109964 and HL135824 (C.J.W.). Genotyping services were supported by The Liaison Committee between the Central Norway Regional Health Authority, the Norwegian University of Science and Technology, the Research Council of Norway, and the University of Michigan. W.Z. was supported by the University of Michigan Rackham Predoctoral Fellowship. We wish to thank all HUNT study participants who contributed to scientific research. We also appreciate the reviewers and editors for their thoughtful and constructive comments that helped improve the manuscript substantially.

## CONFLICTS OF INTEREST

The authors have no conflict of interest to declare.

## ORCID

Wei Zhou  <http://orcid.org/0000-0001-7719-0859>

Jonas B. Nielsen  <http://orcid.org/0000-0002-6654-2852>

Cristen J. Willer  <http://orcid.org/0000-0001-5645-4966>

## REFERENCES

- Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., ... Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223. <https://doi.org/10.1016/j.ajhg.2009.01.005>
- Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*(2), 459–471. <https://doi.org/10.1534/genetics.113.150029>
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., ... Samani, N. J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.
- Cheng, T. H., Thompson, D. J., O'Mara, T. A., Painter, J. N., Glubb, D. M., Flach, S., ... Spurdle, A. B. (2016). Five endometrial cancer risk loci identified through genome-wide association analysis. *Nat Genet*, *48*(6), 667–674. <https://doi.org/10.1038/ng.3562>
- Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V., Walker, N. M., Allen, J. E., ... Todd, J. A. (2008). Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, *40*(12), 1399–1401. <https://doi.org/10.1038/ng.249>
- Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet*, *48*(10), 1284–1287. <https://doi.org/10.1038/ng.3656>
- De Jager, P. L., Jia, X., Wang, J., de Bakker, P. I., Ottoboni, L., Aggarwal, N. T., ... Oksenberg, J. R. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, *41*(7), 776–782. <https://doi.org/10.1038/ng.401>
- Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., ... Kreiner-Moller, E. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the “Genome of The Netherlands.” *European Journal of Human Genetics*, *22*(11), 1321–1326. <https://doi.org/10.1038/ejhg.2014.19>
- Delaneau, O., Zagury, J. F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. <https://doi.org/10.1038/nmeth.2307>
- Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). mini-mac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. <https://doi.org/10.1093/bioinformatics/btu704>
- Ge, Y., Wang, Y., Shao, W., Jin, J., Du, M., Ma, G., ... Zhang, Z. (2016). Rare variants in BRCA2 and CHEK2 are associated with the risk of urinary tract cancers. *Scientific Reports*, *6*, 33542–33548. <https://doi.org/10.1038/srep33542>
- Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddsson, A., Gylfason, A., & Besenbacher, S. (2015). Large-scale whole-genome sequencing of the Icelandic population. *47*(5), 435–444. <https://doi.org/10.1038/ng.3247>
- Horikoshi, M., Mgi, R., van de Bunt, M., Surakka, I., Sarin, A. P., Mahajan, A., ... Morris, A. P. (2015). Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS Genetics*, *11*(7), e1005230. <https://doi.org/10.1371/journal.pgen.1005230>
- Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., ... Dunlop, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics*, *40*(12), 1426–1435. <https://doi.org/10.1038/ng.262>
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide

- association studies through pre-phasing. *Nature Genetics*, 44(8), 955–959. <https://doi.org/10.1038/ng.2354>
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6), e1000529. <https://doi.org/10.1371/journal.pgen.1000529>
- Huang, G. H., & Tseng, Y. C. (2014). Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proceedings*, 8(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo), S64. <https://doi.org/10.1186/1753-6561-8-s1-s64>
- Huang, J., Ellinghaus, D., Franke, A., Howie, B., & Li, Y. (2012). 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *European Journal of Human Genetics*, 20(7), 801–805. <https://doi.org/10.1038/ejhg.2012.3>
- Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., ... Durbin, R. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. 6, 8111–8119. <https://doi.org/10.1038/ncomms9111>
- Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*, 84(2), 235–250. <https://doi.org/10.1016/j.ajhg.2009.01.013>
- Jin, Y., Andersen, G., Yorgov, D., Ferrara, T. M., Ben, S., Brownson, K. M., ... Koks, S. (2016). Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. <https://doi.org/10.1038/ng.3680>
- Jolliffe I. T. (1986). Principal Component Analysis and Factor Analysis. In *Principal component analysis* (pp. 115–128). Springer: New York.
- Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, 25(6), 918–925. <https://doi.org/10.1101/gr.176552.114>
- Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., ... Holmen, J. (2013). Cohort Profile: The HUNT Study, Norway. *International Journal of Epidemiology*, 42(4), 968–977. <https://doi.org/10.1093/ije/dys095>
- Lane, J. M., Vlasac, I., & Anderson, S. G. (2016). Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nature communications*, 7, 10889–10898. <https://doi.org/10.1038/ncomms10889>
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. <https://doi.org/10.1038/nature19057>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Reviews of Genomics and Human Genetics*, 10, 387–406.
- Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8), 816–834. <https://doi.org/10.1002/gepi.20533>
- Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., ... Mohlke, K. L. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6), 768–775. <https://doi.org/10.1038/ng.140>
- Low-Kam, C., Rhainds, D., Lo, K. S., Provost, S., Mongrain, I., Dubois, A., ... Lettre, G. (2016). Whole-genome sequencing in French Canadians from Quebec. *Human Genetics*, 135(11), 1213–1221. <https://doi.org/10.1007/s00439-016-1702-6>
- Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., ... Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3), 234–244. <https://doi.org/10.1038/ng.2897>
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), 499–511. <https://doi.org/10.1038/nrg2796>
- Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7), 906–913. <https://doi.org/10.1038/ng2088>
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... Durbin, R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10), 1279–1283. <https://doi.org/10.1038/ng.3643>
- Mitt, M., Kals, M., Parn, K., Gabriel, S. B., Lander, E. S., Palotie, A., ... Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, 25(7), 869–876. <https://doi.org/10.1038/ejhg.2017.51>
- Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., ... Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics*, 46(9), 989–993. <https://doi.org/10.1038/ng.3043>
- Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., & Matsuda, K. (2015). Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. 47(7), 798–802. <https://doi.org/10.1038/ng.3310>
- Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., ... Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. *European Journal of Human Genetics*, 23(7), 975–983. <https://doi.org/10.1038/ejhg.2014.216>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Roshyara, N. R., & Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genetics*, 16, 90–105. <https://doi.org/10.1186/s12863-015-0248-2>
- Ruth, K. S., Campbell, P. J., Chew, S., Lim, E. M., Hadlow, N., Stuckey, B. G., ... Perry, J. R. (2015). Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *European Journal of Human Genetics*, 24(2), 284–290. <https://doi.org/10.1038/ejhg.2015.102>

- Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.
- Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477. <https://doi.org/10.1371/journal.pgen.1000477>
- van Leeuwen, E. M., Sabo, A., Bis, J. C., Huffman, J. E., Manichaikul, A., Smith, A. V., ... van Duijn, C. M. (2016). Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *Journal of Medical Genetics*, *53*(7), 441–449. <https://doi.org/10.1136/jmedgenet-2015-103439>
- Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: <http://evs.gs.washington.edu/EVS/>) [August 2016 accessed].
- Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., ... Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–90. <https://doi.org/10.1038/nature14962>
- Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., ... Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, *40*(5), 638–645. <https://doi.org/10.1038/ng.120>
- Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., ... Hattersley, A. T. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, *316*(5829), 1336–1341. <https://doi.org/10.1126/science.1142364>

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Zhou W, Fritsche LG, Das S, et al. Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet Epidemiol.* 2017;41:744–755. <https://doi.org/10.1002/gepi.22067>