Finding the Missing Heritability of Genome-wide Association Study Using Genotype Imputation

Yanhui Fan, You-Qiang Song
Centre for Genomic Sciences, The University of Hong Kong; School of Biological Sciences, The University of Hong Kong

Genome-wide association studies (GWAS) have identified thousands of genetic risk variants. However, these variants have explained relatively little of estimated heritability for most complex diseases. The 1000 Genomes Project is a good source to impute missing genotypes for previous GWAS data. Imputation based GWAS can identify more associated signals on a genomewide scale. These new markers can be potential sources of missing heritability. In this study, we did the genotype imputation on the Wellcome Trust Case Control Consortium Phase I genotype data using 1000 genomes as reference. Then we estimated the phenotypic variance explained by all significant association signals. The results suggested that the proportions of phenotypic variance explained by genetic variants increased significantly when the new association variants identified through 1000 genome based imputation were included. These results were consistent with the hypothesis that larger number of variants that are yet to be identified as potential sources of missing heritability.

**Objective**

We hypothesize that the 1000 Genomes-based imputation will increase the density of GWAS marks and will facilitate the investigation of missing heritability without the need for additional genotyping or sequencing.

**Introduction**

Although genome-wide association studies (GWAS) have identified thousands of genetic variants that associated with different complex diseases, a wide gap exists between the estimates of heritability and the heritability that are explained by the genetic variants via GWAS\(^\text{[1]}\). The potential reasons for the missing heritability include myriads of common variants with small effects yet to be found, rare variants and structure variants (insertions, deletions, duplications, inversions, translocations, and copy-number variants) that are poorly detected by available genotyping arrays, and insufficient capability to detect epistasis effects, parental age effects, epigenetic effects, and gene-environment (G×E) interactions\(^\text{[2],[3],[4],[5],[6],[7],[8],[9],[10]}\).

Yang et al. reported a joint estimate of all SNPs, and found that their method (GCTA) can explain a large proportion of the heritability for human height\(^\text{[11]}\). Park et al. re-examined existing GWAS to estimate the number of susceptible loci and the distribution of their effect sizes. They used such estimates to ascertain power and sample-size requirements for future new GWAS or meta-analyses\(^\text{[12]}\). Heritability on the liability scale estimated by GCTA ranged from 0.05 to 0.38 across 13 cancer types\(^\text{[13]}\). These studies argued that a large proportion of the missing heritability can be explained by common variants.

Previous study have demonstrated that 1000 Genomes-based imputation could identify both novel and refined association loci due to the increased density of marks\(^\text{[14],[15]}\). We hypothesize that the increased density of GWAS marks will also facilitate the investigation of missing heritability without the need for additional genotyping or sequencing. We use IMPUTE2\(^\text{[16]}\) for genotype imputation and then apply GCTA\(^\text{[17]}\) to the association results before and after imputation to estimate the heritability of each disease.
BD: bipolar disorder; CAD: coronary artery disease; CD: Crohn’s disease; RA: rheumatoid arthritis; T1D: type 1 diabetes; T2D: type 2 diabetes; K: prevalence; VE: explained variance; SE: standard error. a: Genome-wide association analysis without imputation. b: Genome-wide association analysis with imputation using 1000 Genomes data as reference panel. c: The number of SNPs with p-value less than 1 × 10^{-8}. d: The estimate of phenotypic variance explained by SNPs with p-value less than 1 × 10^{-8}. The values in the parentheses are the standard error of the explained phenotypic variance.
Results & Discussion
After quality control, a total of 444,167 SNPs for 16,179 individuals were retained for the initial association analysis. These SNPs were used as the input genotype data for imputation. Approximately 2.7 million SNPs for each trait were used for association analysis after imputation. The estimation of the phenotypic variance explained by all SNPs with p-value less than $1 \times 10^{-8}$ was performed using the restricted maximum likelihood (REML) analysis, which was implemented in GCTA [Yang 2011](#17). Figure 1 shows the number of SNPs and the estimate of phenotypic variance explained by these SNPs for the six traits. The numbers of SNPs that passed the significant threshold increased more than ten times after imputation compared with the number before imputation. Before imputation, only several to 12.65 percent of the phenotypic variance was explained by the significant SNPs. After imputation, 25.52% to 56.28% of the phenotypic variance was explained by the significant SNPs. SNPs with p-value less than $1 \times 10^{-8}$ after imputation can explain 33.91% to 40.40% of BD phenotypic variance when different prevalence was used. The proportion of phenotypic variance explained by genetic variants in T1D was almost tripled in the 1000 Genomes imputation-based association analysis than in the association analysis without imputation. The explained proportion of phenotypic variance were increased approximately 14 and 17 times in RA and CD, respectively. The proportion were increased even higher in CAD and T2D, about 62 and 95 times, respectively.

We then grouped SNPs with association p-value reached the genome-wide significant level ($1 \times 10^{-8}$) after imputation but were not in LD ($r^2 > 0.8$) with any SNP with association p-value less than $1 \times 10^{-5}$ before imputation as “novel” SNPs. The number of “novel” SNPs and the estimate of phenotypic variance explained by them for the six traits were listed in Figure 2. The results suggested that the novel SNPs are the main reasons behind the increasing of heritability estimate. Since most variants have relatively small effect size, sample size of most studies were not big enough, and the limitation of current genotyping technology, more common variants with intermediate effect and rare variants may be with large effect are yet to be identified. These variants should be tractable through large meta-analysis and imputation-based association analysis. This is the first study that comprehensively examined the utility of 1000 Genomes-based imputation for finding missing heritability. The proportion of phenotypic variance that was explained by genetic variants increased when the contribution of these new variants was included. These findings support that a larger number of variants are yet to be found. These variants are potential sources of missing heritability.

Conclusions
The new additional identified trait-associated variants identified through 1000 genome based imputation can explain part of missing heritability.

Limitations
One potential problem is that the heritability estimates produced by GCTA is sensitive to the chosen sample and may be biased [Krishna Kumar 2015](#18). Although the 1000 Genomes-based imputation increased the proportion of phenotypic variance explained by genetic variants, a substantial proportion of heritability remains unexplained for these diseases. The next-generation sequencing data will accelerate the process of exploring missing heritability. With the rapid increase of the implementation of next-generation sequencing technology, large scale next-generation sequence data from well-phenotyped individuals will be available. It will be a great opportunity to unveil the missing heritability unexplained by common variants that were not covered by current genome-wide association studies.
Additional Information

Methods and Supplementary Material
Please see https://sciencematters.io/articles/201604000013.

Funding Statement
This work was funded by grants from NSFC (No. 81271226), the Research Grant Council of Hong Kong (HKU775208M, HKU 777212M), the Research Fund for the Control of Infectious Diseases of Hong Kong (No.11101032), and the Health and Medical Research Fund of Hong Kong Government (HMRF) (No: 01121726)

Acknowledgements
We acknowledge the WTCCC for making the data available.

Ethics Statement
Not Applicable.

Citations
[9] Greg Gibson. "Hunts of hidden heritability in GWAS". In: Nature Genetics 42.7 (July 2010), pp. 558–560. doi: 10.1038/ng0710-558. URL: http://dx.doi.org/10.1038/ng0710-558
[11] Jian Yang et al. "Common SNPs explain a large proportion of the heritability for human height". In: Nature Genetics 42.7 (June 2010), pp. 565–569. doi: 10.1038/ng.608. URL: http://dx.doi.org/10.1038/ng.608
[12] Ju-Hyun Park et al. "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries". In: Nature Genetics 42.7 (June 2010), pp. 570–575. doi: 10.1038/ng.610. URL: http://dx.doi.org/10.1038/ng.610
[10] Paul R. Burton et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: Nature 447:7145 (June 2007), pp. 661–678. doi:10.1038/nature05911 url: http://dx.doi.org/10.1038/nature05911


