matters

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

Correspondence
felixfanyh@gmail.com
songy@hku.hk

Disciplines
Genetics

Keywords
GWAS
Finding Missing Heritability
Genotype Imputation

Type of Observation
Standalone

Type of Link
Standard Data

Submitted Jan 21, 2016
Published May 4, 2016

Abstract
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 genome-wide 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 Genomes-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.

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 [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 [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 [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 [12]. Heritability on the liability scale estimated by GCTA ranged from 0.05 to 0.38 across 13 cancer types [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 [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 [16] for genotype imputation and then apply GCTA [17] to the association results before and after imputation to estimate the heritability of each disease.

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.

Figure 1. Estimation of the phenotypic variance explained by SNPs.

Disease	K	GWAS[a]		Imputation-based GWAS[b]	
		No. Sig SNP[c]	VE%(SE%)[d]	No. Sig SNP[c]	VE%(SE%)[d]
BD	0.0045	0	NA	437	33.91(1.28)
	0.01	0	NA	437	40.40(1.52)
CAD	0.056	13	0.91(0.86)	278	56.28(4.34)
CD	0.0005	26	1.48(0.63)	395	25.52(1.07)
	0.001	26	1.64(0.70)	395	28.31(1.19)
RA	0.0075	3	2.05(1.67)	138	27.85(3.27)
	0.01	3	2.19(1.78)	138	29.74(3.50)
T1D	0.0054	43	12.65(2.77)	582	36.74(1.45)
T2D	0.028	10	0.51(0.49)	175	48.24(3.80)

a

Figure Legend

Figure 1.

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.

Supplementary Figure 1. Estimate of the phenotypic variance explained by SNPs.

Supplementary Figure 2. Estimation of the phenotypic variance explained by novel SNPs.

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 1×10^{-8} was performed using the restricted maximum likelihood (REML) analysis, which was implemented in GCTA [17].

Figure 1 shows the number of SNPs and the estimate of phenotypic variance was explained by these SNPs for the 6 traits. The numbers of SNPs that passed the significant threshold increased more than 10 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 1×10^{-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 (1×10^{-8}) after imputation but were not in LD (r^2 >0.8) with any SNP with association p-value less than 1×10^{-5} before imputation as "novel" SNPs. The number of "novel" SNPs and the estimate of phenotypic variance explained by them for the 6 traits were listed in supplementary 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 Genomes-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 [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

[1] Paolo Vineis and Neil Pearce and. "Missing heritability in genome-wide association study research". In: *Nature Reviews Genetics* 11.8 (Aug. 2010), pp. 589–589. DOI: 10.1038/nrg2809-c2. URL: http://dx.doi.org/10.1038/nrg2809-c2.

[2] Anne Goriely and Andrew O. M. Wilkie. "Missing heritability: paternal age effect mutations and selfish spermatogonia". In: *Nature Reviews Genetics* 11.8 (Aug. 2010), pp. 589–589. DOI: 10.1038/nrg2809-c1. URL: http://dx.doi.org/10.1038/nrg2809-c1.

[3] Evan E. Eichler et al. "Missing heritability and strategies for finding the underlying causes of complex disease". In: *Nature Reviews Genetics* 11.6 (June 2010), pp. 446–450. DOI: 10.1038/nrg2809. URL: http://dx.doi.org/10.1038/nrg2809.

[4] Angus J Clarke and David N Cooper and. "GWAS: heritability missing in action?" In: *European Journal of Human Genetics* 18.8 (Mar. 2010), pp. 859–861. DOI: 10.1038/ejhg.2010.35. URL: http://dx.doi.org/10.1038/ejhg.2010.35.

[5] Teri A. Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (Oct. 2009), pp. 747–753. DOI: 10.1038/nature08494. URL: http://dx.doi.org/10.1038/nature08494.

[6] Ali J. Marian and. "Elements of 'missing heritability'". In: *Current Opinion in Cardiology* 27.3 (May 2012), pp. 197–201. DOI: 10.1097/hco.0b013e328352707d. URL: http://dx.doi.org/10.1097/hco.0b013e328352707d.

[7] Eamonn MM Quigley and. "Epigenetics: filling in the 'heritability gap' and identifying gene-environment interactions in ulcerative colitis". In: *Genome Medicine* 4.9 (2012), p. 72. DOI: 10.1186/gm373. URL: http://dx.doi.org/10.1186/gm373.

[8] Noah Zaitlen and Peter Kraft and. "Heritability in the genome-wide association era". In: *Human Genetics* 131.10 (July 2012), pp. 1655–1664. DOI: 10.1007/s00439-012-1199-6. URL: http://dx.doi.org/10.1007/s00439-012-1199-6.

[9] Greg Gibson and. "Hints 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.

[10] O. Zuk et al. "The mystery of missing heritability: Genetic interactions create phantom heritability". In: *Proceedings of the National Academy of Sciences* 109.4 (Jan. 2012), pp. 1193–1198. DOI: 10.1073/pnas.1119675109. URL: http://dx.doi.org/10.1073/pnas.1119675109.

[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.

[13] Joshua N. Sampson et al. "Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types". In: *Journal of the National Cancer Institute* 107.12 (Oct. 2015), djv279. DOI: 10.1093/jnci/djv279. URL: http://dx.doi.org/10.1093/jnci/djv279.

[14] Jie Huang et al. "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data". In: *European Journal of Human Genetics* 20.7 (Feb. 2012), pp. 801–805. DOI: 10.1038/ejhg.2012.3. URL: http://dx.doi.org/10.1038/ejhg.2012.3.

[15] C Herold et al. "Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3". In: *Molecular Psychiatry* (Feb. 2016). DOI: 10.1038/mp.2015.218. URL: http://dx.doi.org/10.1038/mp.2015.218.

[16] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini and. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies". In: *PLOS Genetics* 5.6 (June 2009), e1000529. DOI: 10.1371/journal.pgen.1000529. URL: http://dx.doi.org/10.1371/journal.pgen.1000529.

[17] Jian Yang et al. "GCTA: A Tool for Genome-wide Complex Trait Analysis". In: *The American Journal of Human Genetics* 88.1 (Jan. 2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011. URL: http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

[18] Siddharth Krishna Kumar et al. "Limitations of GCTA as a solution to the missing heritability problem". In: *Proceedings of the National Academy of Sciences* 113.1 (Dec. 2015), E61–E70. DOI: 10.1073/pnas.1520109113. URL: http://dx.doi.org/10.1073/pnas.1520109113.

[19] 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.

[20] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (July 2006), pp. 904–909. DOI: 10.1038/ng1847. URL: http://dx.doi.org/10.1038/ng1847.

[21] Richard M. Durbin et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (Oct. 2010), pp. 1061–1073. DOI: 10.1038/nature09534. URL: http://dx.doi.org/10.1038/nature09534.

[22] Yan-Hui Fan and You-Qiang Song and. "IPGWAS: An integrated pipeline for rational quality control and association analysis of genome-wide genetic studies". In: *Biochemical and Biophysical Research Communications* 422.3 (June 2012), pp. 363–368. DOI: 10.1016/j.bbrc.2012.04.117. URL: http://dx.doi.org/10.1016/j.bbrc.2012.04.117.

[23] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. DOI: 10.1086/519795. URL: http://dx.doi.org/10.1086/519795.

[24] Jonas Halfvarson et al. "Inflammatory bowel disease in a Swedish twin cohort: a long-term follow-up of concordance and clinical characteristics". In: *Gastroenterology* 124.7 (June 2003), pp. 1767–1773. DOI: 10.1016/s0016-5085(03)00385-8. URL: http://dx.doi.org/10.1016/S0016-5085(03)00385-8.

[25] Alexander J. MacGregor et al. "Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins". In: *Arthritis and Rheumatism* 43.1 (Jan. 2000), pp. 30–37. DOI: 10.1002/1529-0131(200001)43:1<30::aid-anr5>3.0.co;2-b. URL: http://dx.doi.org/10.1002/1529-0131(200001)43:1%3C30::AID-ANR5%3E3.0.CO;2-B.

[26] Paul Lichtenstein et al. "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study". In: *The Lancet* 373.9659 (Jan. 2009), pp. 234–239. DOI: 10.1016/s0140-6736(09)60072-6. URL: http://dx.doi.org/10.1016/S0140-6736(09)60072-6.

[27] V. Hyttinen et al. “Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs: A Nationwide Follow-Up Study”. In: *Diabetes* 52.4 (Apr. 2003), pp. 1052–1055. DOI: 10.2337/diabetes.52.4.1052. URL: http://dx.doi.org/10.2337/diabetes.52.4.1052.

[28] Marjorie E. Marenberg et al. “Genetic Susceptibility to Death from Coronary Heart Disease in a Study of Twins”. In: *The New England Journal of Medicine* 330.15 (Apr. 1994), pp. 1041–1046. DOI: 10.1056/nejm199404143301503. URL: http://dx.doi.org/10.1056/NEJM199404143301503.

[29] J Sofaer and. “Crohn’s disease: the genetic contribution.” In: *Gut* 34.7 (July 1993), pp. 869–871. DOI: 10.1136/gut.34.7.869. URL: http://dx.doi.org/10.1136/gut.34.7.869.

[30] S Harney and B.P. Wordsworth and. “Genetic epidemiology of rheumatoid arthritis”. In: *Tissue Antigens* 60.6 (Dec. 2002), pp. 465–473. DOI: 10.1034/j.1399-0039.2002.600601.x. URL: http://dx.doi.org/10.1034/j.1399-0039.2002.600601.x.

[31] N Craddock et al. “Mathematical limits of multilocus models: the genetic transmission of bipolar disorder.” In: *The American Journal of Human Genetics* 57 (1995), pp. 690–702.

[32] Swapan Kumar Das and Steven C Elbein. “The Genetic Basis of Type 2 Diabetes”. In: *Cellscience* 2 (2006), pp. 100–131.