Reliability of imputed genotype data for family-based analyses

Mahdi Mir[1], Tammy Tan[2], Patrick Turley[3,4], Daniel J. Benjamin[1,2,5], and Alexander Strudwick Young[1,5]

[1]UCLA Anderson School of Management, Los Angeles, CA, USA
[2]National Bureau of Economic Research, Cambridge, MA, USA
[3]Department of Economics, University of Southern California, Los Angeles, CA, USA
[4]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA
[5]UCLA Department of Human Genetics, David Geffen School of Medicine, Los Angeles, CA, USA

Introduction

Family-based analyses use random genetic variation within families to remove confounding from population stratification and to separate out direct genetic effects (effects of variants in an individual on that individual) from indirect genetic effects (effects of variants in an individual on another individual mediated through the environment). The underlying principle relies on relationships between relatives' genotypes due to random segregations of genetic material during meiosis[1–5]. Family-based genome-wide association studies (FGWASs) — which use parental genotypes as control variables — and sib-GWAS[6] — which use sibling genotypes as control variables — have been proposed as solutions to the confounding issues known to affect standard GWAS designs and downstream applications including estimation of heritability and genetic correlation, Mendelian randomization, and inferences of natural selection.

While family-based analyses have favorable theoretical properties, issues arising from imperfections in real-world genotype data have not been explored. Most GWAS studies use data derived from genotyping arrays — which measure genotypes at an incomplete set of variants — followed by imputation from a panel of reference haplotypes with more complete genome sequences. The imputation methods work by finding related sequences in the reference panel and using the reference haplotypes to fill in the genotypes at the positions not directly observed on the genotyping array, but they do not account for relationships between individuals in the target sample, including the sibling and parent-offspring relations used in family-based analyses. The output of imputation methods typically consists of a probability distribution over genotypes, from which the most likely genotype ('hard-call') or the expected genotype ('dosage') is used for downstream analyses. Imputation methods also provide a quality metric (INFO score or R^2) that estimates the fraction of genotype variation the imputation has recovered and is thus bounded in [0,1].

Results

Correlations between relatives' imputed genotypes

We sought to examine whether imputed data preserves the relationships between relatives' genotypes implies by Mendelian Laws and from which the theoretical properties of family-based analyses derive. We analyzed correlations between relatives' imputed genotypes (UKBv3 imputation) for 19,290 sibling pairs and 5,324 parent-offspring pairs from the UK Biobank White British subsample. We found that correlations between first-degree relatives' imputed genotypes decreased below the theoretical expectation with imputation quality (measured by INFO score with values between 0 and 1): for imputed SNPs (MAF>1%) with INFO scores between 0.30–0.31, the mean correlation between siblings was 0.437 (S.E.=0.001) for genotype dosages and 0.376 (S.E.=0.001) for hard-calls, lower than the expected correlation of 0.5. Even at INFO scores of 0.96–0.97, the correlation remained below 0.5 ($P=1.6x10^{-4}$) for

hard-call genotypes. For parent-offspring pairs, the mean correlation between parent-offspring pairs for SNPs with INFO scores between 0.30-0.31 was 0.434 (S.E.=0.001) for genotype dosages and 0.337 (S.E.=0.002) for hard-calls. Like for sibling pairs, the correlation remained below 0.5 even at INFO scores of 0.96-0.97 ($P=2.59x10^{-2}$) (Figure 1).

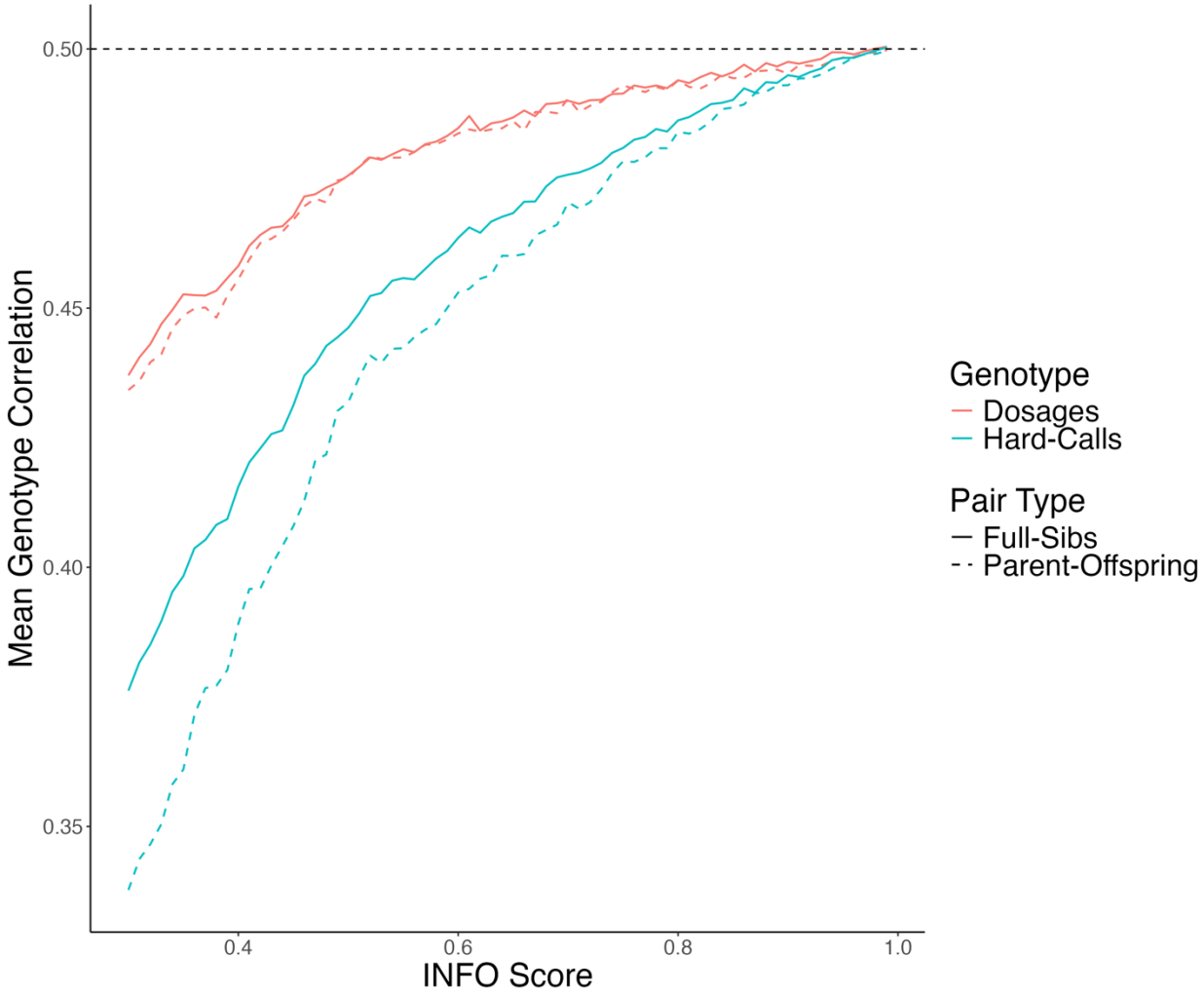

Figure 1. Correlations between relatives' genotypes as a function of imputation quality (INFO score). *We show the mean correlation from 1000 SNPs in each INFO score bin for both imputed dosages (expected genotype given genotype probabilities) and imputed hard-call genotypes (most likely genotype). Results from 5,324 parent-offspring pairs and 19,290 sibling pairs from the white British subsample of the UK Biobank.*

Assessing bias due to imputation in sib-GWAS

We now show how the bias in sib-GWAS induced by imputation can be assessed by comparing sibling imputed genotypes and genotypes from whole genome sequencing (WGS) data. For a sibling pair from family i, the FGWAS model[7] is:

$$Y_{i1} = \delta g_{i1} + \alpha g_{\mathrm{par}(i)} + \epsilon_{i1};$$
$$Y_{i2} = \delta g_{i2} + \alpha g_{\mathrm{par}(i)} + \epsilon_{i2};$$

where Y_{i1} and g_{ij} are, respectively, the phenotype and SNP genotype of sibling j in family j; $g_{\mathrm{par}(i)}$ is the sum of paternal and maternal genotypes in family i; δ is the direct genetic effect (DGE), the target of sib-GWAS and FGWAS; α is the average non-transmitted coefficient (NTC); and ϵ_{i1} are the residuals.

We assume that sibling genotypes are mean-normalized. The sib-difference regression can proceed by least-squares regression of Y_{i1} onto $(g_{i1} - g_{i2})$. This will give an unbiased estimate of δ. We can rewrite the equation in terms of the regressions onto $(g_{i1} - g_{i2})$ and $(g_{i1} + g_{i2})$ (which are uncorrelated with each other):

$$Y_{i1} = \delta(g_{i1} - g_{i2}) + (\delta/2 + 2\alpha/3)(g_{i1} + g_{i2}) + \epsilon'_{i1},$$

for some ϵ'_{i1} uncorrelated with $(g_{i1} - g_{i2})$ and $(g_{i1} + g_{i2})$. We can then ask what we would obtain if we instead performed regression onto the difference in imputed sibling genotypes:

$$\frac{\mathrm{Cov}(Y_{i1}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} = \delta \frac{\mathrm{Cov}(g_{i1} - g_{i2}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} + (\delta/2 + 2\alpha/3) \frac{\mathrm{Cov}(g_{i1} + g_{i2}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} + \frac{\mathrm{Cov}(\epsilon'_{i1}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})}$$

We can thus assess the bias that comes from using imputed sibling genotypes by performing two regressions:

1. Slope of regression of $g_{i1} - g_{i2}$ on $\hat{g}_{i1} - \hat{g}_{i2}$, which gives bias coming from the sib-difference component;
2. Slope of regression of $g_{i1} + g_{i2}$ on $\hat{g}_{i1} - \hat{g}_{i2}$, which gives bias coming from the sib-sum component, i.e. the component that the sib-difference should be orthogonal with as it captures parental indirect genetic effects and confounding factors.

This can be done by obtaining the sibling genotypes from WGS data (which we assume to be true) and forming $g_{i1} - g_{i2}$ and $g_{i1} + g_{i2}$ and regressing these values onto the difference in sib imputed genotypes, $\hat{g}_{i1} - \hat{g}_{i2}$.

From these slopes, we could express the expected result of sib-difference regression using imputed genotypes as a linear combinations of δ and α, enabling us to quantify the bias and confounding in terms of the DGE and NTC from the FGWAS model.

Comparison of sibling imputed and WGS genotypes.

We regressed the difference between siblings' genotypes, derived from WGS data in the UKB, onto the difference from imputed data:

$$(g_{i1} - g_{i2}) \sim (\widehat{g_{i1}} - \widehat{g_{i2}}).$$

Theoretically, we would expect an intercept of zero and a slope of one if the imputed genotypes are the same as the WGS genotypes.

We analyzed 68 high-quality imputed SNPs (mean info score of 0.966) and 46 low-quality imputed SNPs (mean info score of 0.312) — see Figure 2. For high-quality SNPs, the average slope estimate was 0.957 (S.E. = 0.0015), whereas for low-quality SNPs the average slope was 0.062 (S.E. = 0.0142), suggesting a much weaker relationship between imputed and WGS genotypes than would expected given the INFO score.

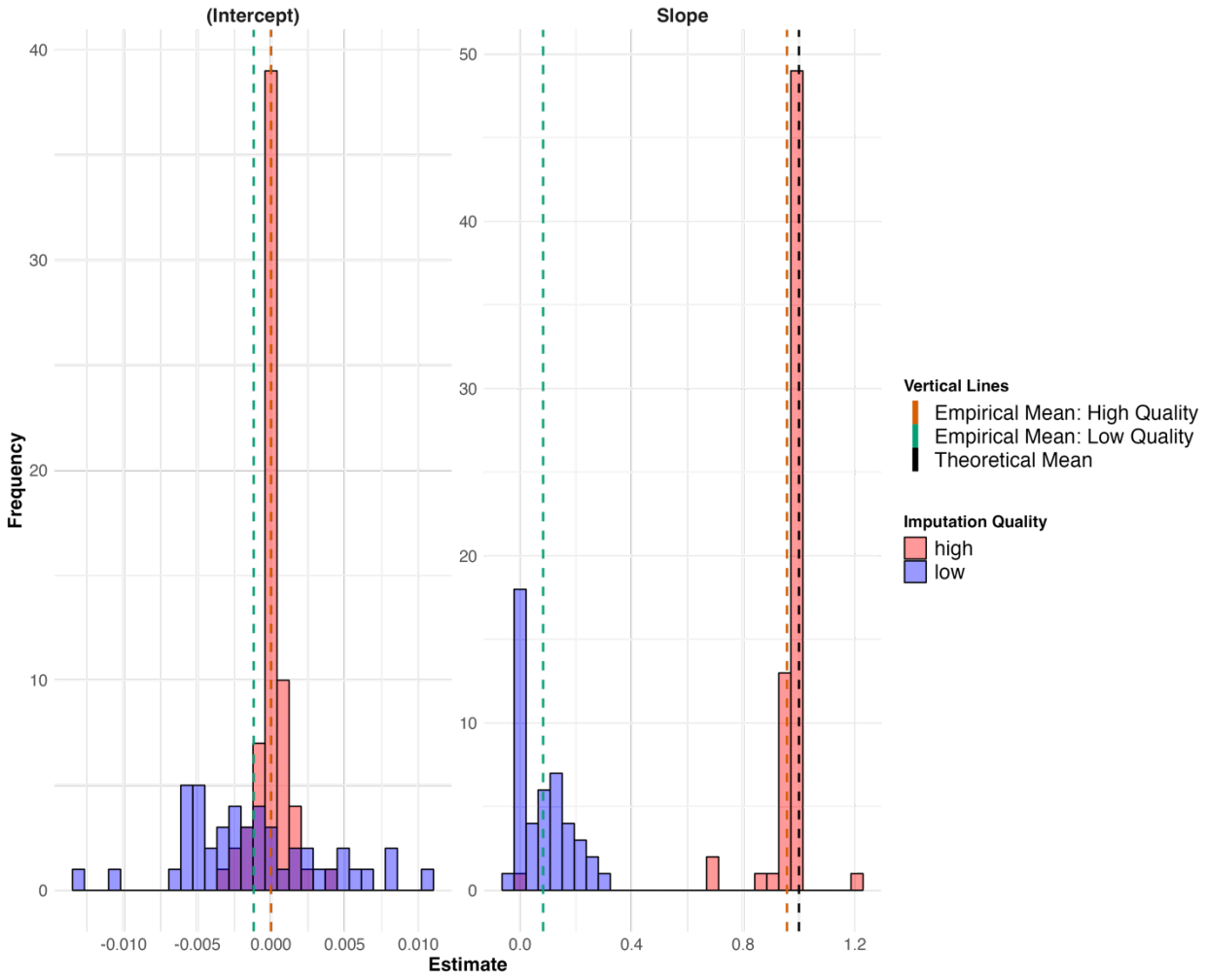

Figure 2. The distribution of intercepts and slopes from the regression of $g_{i1} - g_{i2}$ (from WGS) onto $\hat{g}_{i1} - \hat{g}_{i2}$ (imputed). *We analyzed two groups of SNPs: 68 high-quality SNPs (mean INFO score of 0.966) and 46 low-quality SNPs (mean INFO score 0.312). The regression was performed for 19,052 White British sibling pairs from the UK Biobank. If imputed genotypes are the same as WGS genotypes, we expect an intercept of zero and a slope of 1.*

INFO score is an unreliable metric of imputation quality in UK Biobank

The above results showed much poorer concordance between imputed and WGS genotypes than we expected based on INFO score. Thus, we performed a simpler analysis that looked at the regression R^2 between imputed and WGS genotypes as a function of SNP INFO score using 19,052 White British individuals, one from each sibling pair.

Figure 3 presents the relationship between the INFO score and the R^2 between imputed and WGS data. The results indicate that the INFO score is highly unreliable — rather than matching the R^2, as expected, the actual R^2 values are often much lower. For high-quality SNPs, some have R^2 values as low as 0.5,

while for low-quality SNPs (with INFO scores around 0.3), the imputed and WGS genotypes are mostly uncorrelated.

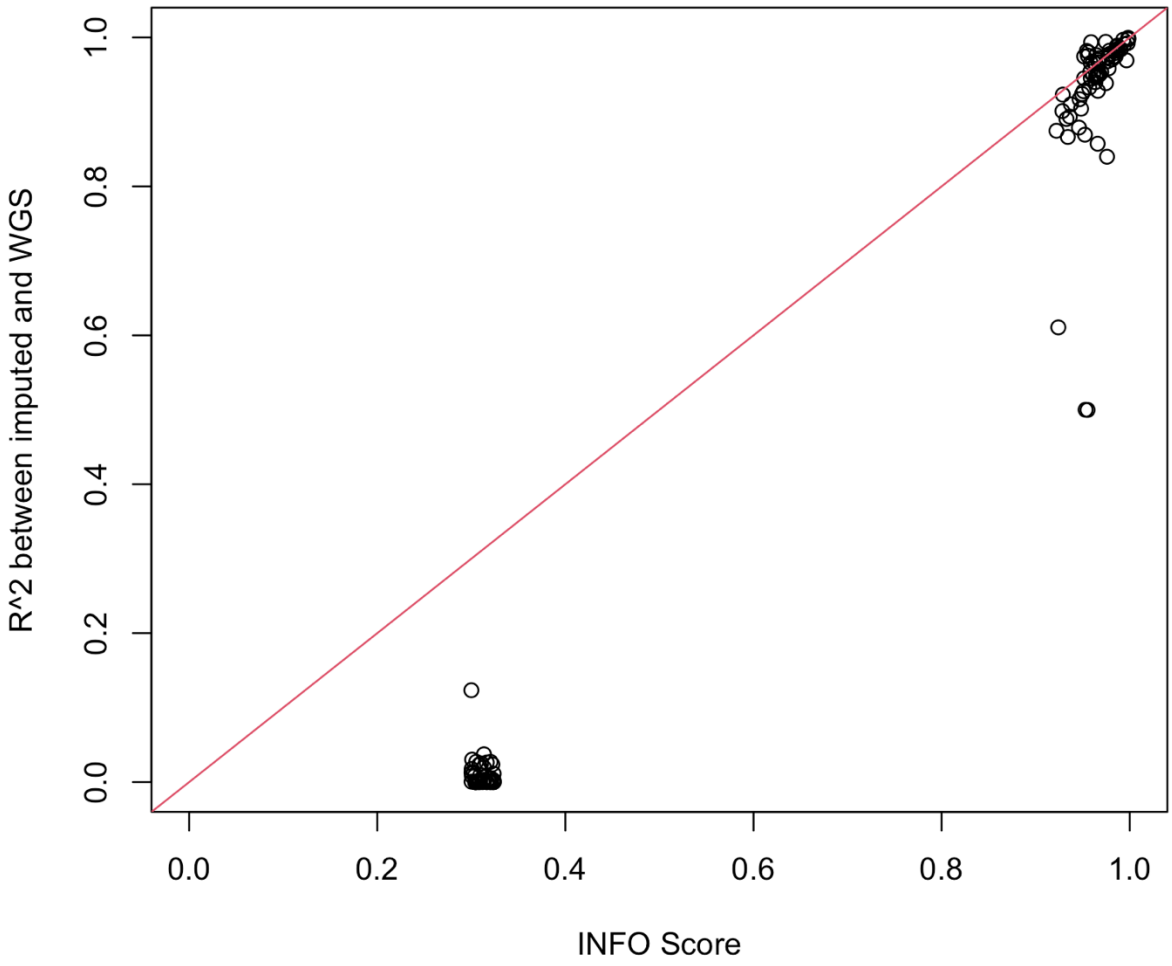

Figure 2. INFO score vs R^2 between imputed and WGS data in UK Biobank.

Discussion

These results raise concerns about sib-GWAS results that use low-quality imputed genotypes. For example, a sib-GWAS meta-analysis by Howe et al.[6] used hard-call genotypes for all SNPs with MAF>1% and INFO score greater than 0.3, implying many low-quality imputed genotypes were used. To investigate the impact that low-quality imputed genotypes have on sib-GWAS and FGWAS, we are going to perform more analyses comparing the imputed genotypes in UK Biobank to the whole-genome sequencing data. We will extend the analysis to a greater number of SNPs across different INFO score values. We will then perform sib-GWAS and FGWAS using WGS data and compare the results from downstream analyses to those from imputed data.

Our results argue that stringent quality control should be applied for family-based analyses using imputed genotype data to ensure that such analyses have the theoretical properties that they promise. Further investigations should be performed to evaluate the impact of low quality imputed genotypes in polygenic index (PGI) analysis. Furthermore, imputation methods that account for pedigree relations between close

relatives should be developed to enable better family-based analyses of data derived from genotyping arrays.

Beyond implications for family-based analyses, our results raise more general questions about the reliability of imputed data, currently the vast majority of data that has been analyzed in GWAS. An important question is whether the unreliability of the imputed data in the UK Biobank applies to other imputed datasets. Even if it does not, since the UK Biobank imputed genotype data[8] is probably the most used genetic data in the world, the issues with this particular imputed dataset deserve attention.

1. Young, A. S. Genome-wide association studies have problems due to confounding: Are family-based designs the answer? *PLOS Biol.* **22**, e3002568 (2024).
2. Benjamin, D. J., Cesarini, D., Turley, P. & Young, A. S. Social-Science Genomics: Progress, Challenges, and Future Directions. (2024).
3. Tan, T. *et al.* Family-GWAS reveals effects of environment and mating on genetic associations. *medRxiv* 2024–10 (2024).
4. Junming Guan, Seyed Moeen Nehzati, Daniel J. Benjamin, & Alexander I. Young. Novel estimators for family-based genome-wide association studies increase power and robustness. *bioRxiv* 2022.10.24.513611 (2022) doi:10.1101/2022.10.24.513611.
5. Veller, C. & Coop, G. Interpreting population- and family-based genome-wide association studies in the presence of confounding. *PLoS Biol.* (2024).
6. Howe, L. J. *et al.* Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).
7. Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).
8. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

Grant Support: The study was supported by Open Philanthropy and the National Institute on Aging/National Institutes of Health through grants R24-AG065184 and R01-AG042568 to D.J.B., and R01-AG083379 to A.S.Y.