**Low quality genotype data is not appropriate for family-based analyses**

Mahdi Mir1, Tammy Tan2,­Patrick Turley3,4, Daniel J. Benjamin1,2,5, and Alexander Strudwick Young1,5

UCLA Anderson School of Management, Los Angeles, CA, USA

National Bureau of Economic Research, Cambridge, MA, USA

Department of Economics, University of Southern California, Los Angeles, CA, USA

Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA

UCLA Department of Human Genetics, David Geffen School of Medicine, Los Angeles, CA, USA

**Introduction**

Family-based analyses have become an important part of human genetic analyses, especially for socio-economic and behavioral phenotypes. 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 segregation of genetic material during meiosis. Family-based genome-wide association studies (FGWASs) — which use parental genotypes as control variables — and sib-GWAS — 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 FGWAS has favorable theoretical properties, issues arising from imperfections in real-world genotype data have not been thoroughly explored. Most GWAS studies use imputed data. The data is derived from genotyping array data — which measures genotypes at an incomplete set of variants — followed by imputation from a panel of reference haplotypes with more complete genome sequencing data. The imputation methods work by finding closely 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. 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 R2) that estimates the fraction of genotype variation the imputation has recovered.

**Results**

Imputation is a process that does not consider family relations in the GWAS sample. Thus, imputed genotypes — especially those of low quality — may not preserve the expected properties of family data, potentially leading to misleading results.

To examine this, we analyzed 19,290 sibling pairs and 5,324 parent-offspring pairs from the UK Biobank. 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. Similar results were obtained for parent-offspring genotype correlations.

A graph of different colored lines

AI-generated content may be incorrect.

**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 along with its 95% confidence interval 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.

In a separate analysis, we regressed the difference between siblings’ genotypes, derived from whole-genome sequencing (WGS) data in the UKB, onto the difference from imputed data . Theoretically, we would expect an intercept of zero and a slope of one.

Figure 2 illustrates the distribution of the intercepts and slopes from these regressions for individual SNPs. The results indicate significant deviations from theoretical expectations for both high- and low-quality imputed SNPs, with the deviations being more pronounced for low-quality imputed genotypes.

For high-quality imputed genotypes (68 SNPs with a mean info score of 96.6%), the average slope estimate is 95.7% (mean S.E. = 0.0015, mean P = 0.0146). In contrast, for low-quality imputed SNPs (46 SNPs with a mean info score of 31.2%), the mean slope estimate is only 6.2% (mean S.E. = 0.0142, mean P = 0.0090).

A screenshot of a graph

AI-generated content may be incorrect.

Figure 2. The distribution of intercept and the slope for the regression of onto .

We also conducted a straightforward analysis by regressing the WGS-derived genotype of 19,052 White British individuals onto their imputed genotype at the same SNP. To avoid dependence between observations, we included only one sibling from each sib pair in the UKB data. We then plotted the INFO scores against the R2 values from the fitted models.

Figure 3 presents the relationship between the INFO score and the R2 between imputed and WGS data. The results indicate that the INFO score is highly unreliable—rather than matching the expected R2, the actual R2 values are often much lower. For high-quality SNPs, some exhibit R2 values as low as 0.5, while for low-quality SNPs (with INFO scores around 0.3), the imputed and WGS genotypes are nearly uncorrelated.

A diagram of a red line

AI-generated content may be incorrect.

Figure 3. INFO score vs R2 between imputed and WGS data.

**Discussion**

These results raise concerns about results from a recent sib-GWAS meta-analysis by Howe et al. that 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. For the next step we are going to extend the analysis to a greater number of SNPs across different INFO score values.

This will enable us to measure the bias and confounding introduced by low quality SNPs. We will compare sib-GWAS and FGWAS results from both imputed genotypes and whole-genome sequencing data.

Our results, although preliminary, indicate that stringent quality control is required for family-based analyses using imputed genotype data. 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.

**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.