**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

**Abstract**

Family-based analyses have become an important part of human genetic analyses, especially for socio-economic and behavioural 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). 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.

A graph of different colored lines

Description automatically generated

**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.

While FGWAS has favorable theoretical properties, issues arising from imperfections in real-world genotype data have not been thoroughly explored. Most GWAS 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.

The fact genotypes are imputed without reference to pedigree relations suggests that low quality imputed genotypes of parents or siblings may inadequately control for confounding factors in FGWAS or sib-GWAS. We examined this issue using 19,290 sibling pairs and 5,324 parent-offspring pairs from the white British subsample of the UK Biobank. We found that the correlations between sibling and parent-offspring pairs’ imputed genotypes fall below the expected value of 0.5 as the INFO score of the SNP declines below 1 (Figure 1). For imputed genotype dosages, the mean correlation between siblings’ genotypes across 1000 SNPs with INFO scores between 0.30 and 0.31 and minor allele frequency (MAF) at least 1% was 0.437 (S.E.=0.001), much lower than the expected correlation of 0.5 based on Mendelian laws; for imputed hard calls, the correlation was even lower at 0.376 (S.E.=0.001). The correlation between siblings’ genotypes only approaches the expected correlation of 0.5 as the INFO score gets very close to 1, with detectable deviations from 0.5 for hard-call genotypes for SNPs with INFO scores between 0.96-0.97: mean correlation 0.4983 (S.E.=4.7x10-4, P=1.6x10-4 for correlation below 0.5). Results for parent-offspring pairs were similar for imputed dosages but worse for hard-call genotypes (Figure 1).

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 performing analyses comparing the imputed genotypes in UK Biobank to the whole-genome sequencing data. 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.

Family-based analyses have also become important in the analysis of polygenic predictors (called polygenic indices, or PGIs). Use of low-quality imputed genotypes in polygenic prediction analyses is standard practice since it often increases out-of-sample prediction accuracy. However, our results indicate that use of low-quality imputed genotypes could introduce biases and confounding. Furthermore, low correlation between relatives’ imputed genotypes may bias estimates of assortative mating from PGIs.

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.