**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 leverage random genetic variation within families to control for population stratification and distinguish direct from indirect genetic effects. Family-based GWAS (FGWAS) and sib-GWAS aim to address confounding issues affecting standard GWAS designs. However, real-world genotype data imperfections, particularly from imputed genotypes, pose challenges that have not been fully explored.

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.

Most GWAS data originate from genotyping arrays followed by imputation using a reference panel. Since imputation does not consider pedigree relationships, low-quality imputed genotypes in parents or siblings may fail to properly control for confounding in FGWAS or sib-GWAS. To examine this, we analyzed 19,290 sibling pairs and 5,324 parent-offspring pairs from the UK Biobank white British subsample.

We found that imputed genotype correlations between relatives decrease as the INFO score declines. For imputed dosages, the mean correlation between siblings across 1,000 SNPs with INFO scores of 0.30–0.31 and minor allele frequency ≥1% was 0.437 (S.E.=0.001), lower than the expected 0.5. Hard-call genotypes had an even lower correlation of 0.376 (S.E.=0.001). Even at INFO scores of 0.96–0.97, hard-call genotype correlations remained slightly below 0.5 (P=1.6x10⁻⁴).

These findings suggest that rigorous quality control is essential for family-based analyses using imputed genotypes. Additionally, developing imputation methods that incorporate pedigree information could improve accuracy in family-based genetic studies.