**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

**Keywords: GWAS, imputation, quality, family, siblings**

**Abstract**

Genome wide association studies (GWASs) conducted using sibling or parental genotypes as controls aim to remove confounding factors and to separate direct and indirect genetic effects. The underlying principle relies on relationships between relatives’ genotypes due to segregation of genetic material during meiosis. However, most GWAS data derive from genotyping arrays followed by imputation from a reference panel, 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. To assess the impact of imputation quality on family-based estimates, we are comparing imputed genotypes to genotypes from whole-genome sequencing (WGS) data in the UK Biobank. Preliminary results indicate much poorer correspondence between imputed and WGS sibling genotypes than expected based on imputation quality metrics. Our findings underscore the importance of stringent quality control in family-based analyses.

**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.