Imputation Around the World: Assessing Imputation Quality Across Diverse Global Populations

Jordan L. Cahoon, Xinyue Rui, Echo Tang, Christopher Simons, Jalen Langie, Ying-Chu Lo, Charleston W.K. Chiang

June 2022

Abstract

Genotype imputation is now an essential component in human genetic studies. By predicting unobserved genotypes based on sequenced individuals, imputation increases marker density and enables large-scale genome-wide association studies. The state-of-the-art imputation reference panel released by the Trans-Omics for Precision Medicine (TOPMed) contains a substantial number of admixed African and Hispanic/Latino samples. As a result, these populations are imputed with nearly the same efficacy as European cohorts. However, imputation for ethnic minorities primarily residing outside of North America still falls short in performance due to persisting underrepresentation. To illustrate this point, we curated genome- wide array data from 28 publications published between 2008 to 2021. In total, we imputed over 30k individuals across 145 populations around the world. We identified a number of populations where the imputation accuracy paled in comparison to that of European populations. For instance, the mean imputation r^2 for variants with minor allele frequency (MAF) between 1.0- 5.0 for Saudi Arabian (N=1061), Vietnamese (N=1264), Thai (N=2435), and Papua New Guineans (N=776) were 0.79, 0.78, 0.76, and 0.62, respectively. In contrast, the mean r^2 ranged from 0.90 to 0.93 for comparable European populations matched in sample size and SNP content. In addition to overall lower imputation quality, minority populations experienced a steeper drop in imputation accuracy for rarer variants. For example, the mean r^2 for Filipinos (N=1779) declined from 0.93 in alleles with 5-10 MAF to 0.59 in alleles with 0.5-1 MAF. In contrast, alleles in the same frequency classes imputed relatively well in Europeans with a decrease from 0.97 to 0.79. Despite tremendous effort in generating large imputation panels like TOPMed, our results suggest that global populations still incur at least a 10 drop in study power due to imputation accuracy alone and even greater power loss for rarer variants. While strategies leveraging smaller population-specific reference panels in conjunction with meta imputation may increase imputation quality, ultimately, reference panels must strive to increase diversity to promote equity within genetics research.