Genotype imputation accuracy with different reference panels

Guan-Hua Huang and Yi-Chi Tseng
National Chiao Tung University
TAIWAN

Genotype imputation

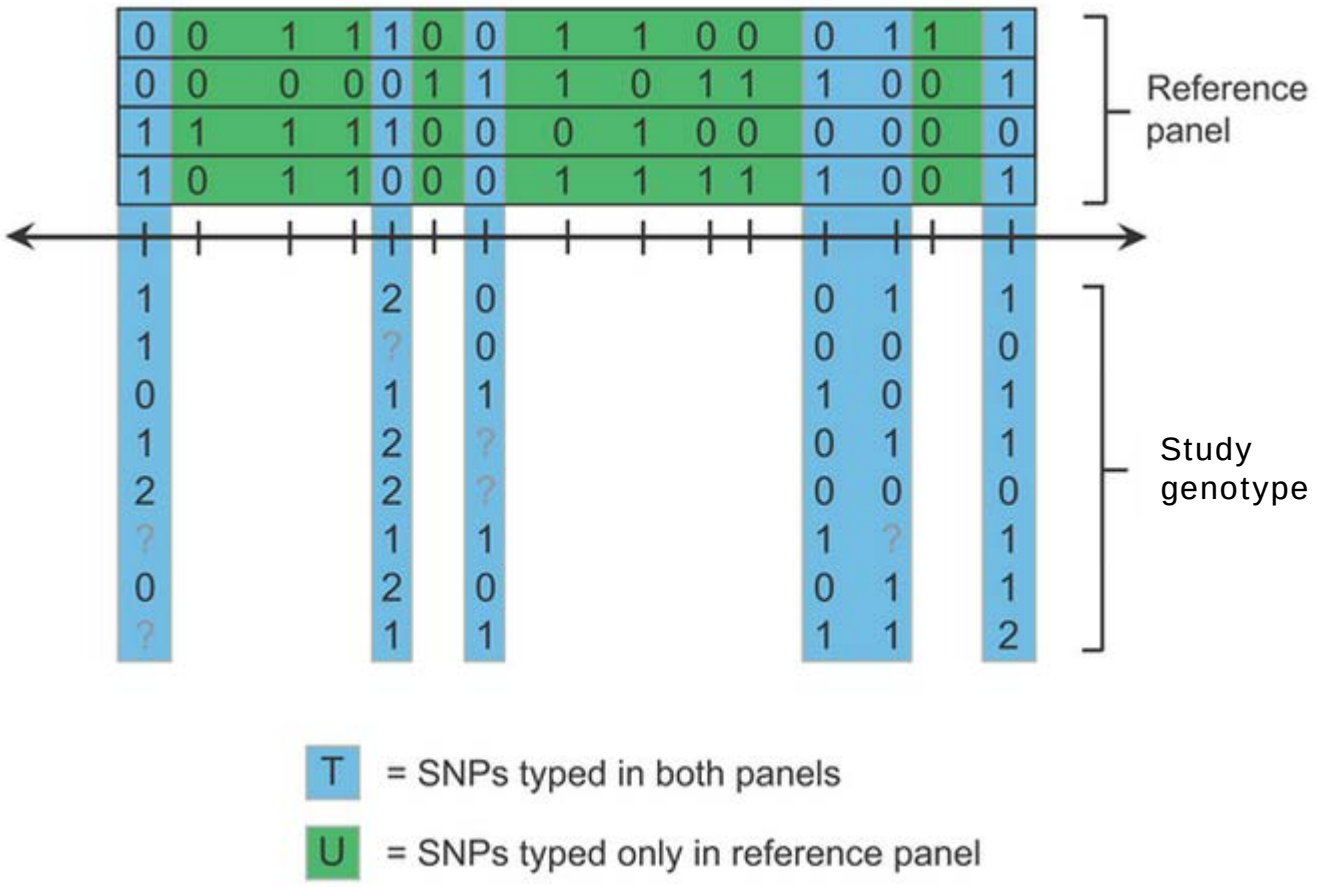
Background

- GWAS based on common SNPs have only identified a small fraction of the complex disease heritability.
  - Rare variants not included in the common genotyping platforms may contribute substantially to the genetic variation of these diseases.
- Using custom-made chips or next-generation sequencing to uncover rare SNPs' effects on the disease can be very expensive in current technology.

Background

- Many researchers thus turn to use the “genotype imputation” approach to predict the genotypes at these rare SNPs that are not directly genotyped in the study sample.

Genotype imputation



Genotype imputation

- A reference panel of individuals genotyped at a dense set of SNPs
- A study sample genotyped at a subset of these sites
- Phase genotypes in the study sample
- Look for matches between the resulting haplotypes and the corresponding partial haplotypes in the reference panel
- Matched haplotype patterns in the reference panel are used to predict unobserved genotypes in the study sample.

How to choose a reference panel?

- Use reference panels from public databases, like HapMap 3 and 1,000 Genomes Project
- A two-stage approach for genotype imputation:
  - the reference panel—a subset of individuals for whole genome sequence (WGS)
  - the study sample—the remaining samples genotyped on commercial genome-wide SNP arrays

Public database reference panels

- Collected from a variety of ethnic populations
- Include the individuals that most closely match the ancestry of the study population as the reference panel
  - Pros: reduce the computational burden of imputation
  - Cons: yield suboptimal accuracy with using partial information, or in studies with no clear reference matches
- Howie *et al*. (2011):
  - Larger and more diverse reference collections could actually make it easier to identify haplotype sharing with simple models, thereby making imputation faster and more accurate.

Two-stage genotype imputation

- Create a reference panel that is genetically similar to the study sample
  - greatly increase the imputation accuracy
- Come at the extra cost of next-generation sequencing

Objectives

- Analyze 464 individuals with both WGS and GWAS data from the GAW18 data set
- Compare genotype imputation accuracy when adopting different reference panels

Data: GAW18 real data set

- 464 individuals with both WGS and GWAS data
- Only impute SNPs on chromosome 3
- Randomly selected 345 individuals (~2/3 of 464) as the study sample

Reference panels compared

1) 1,000 Genomes Phase 1 for 1,094 individuals from Africa, Asia, Europe, and the Americas
2) 120 randomly selected individuals from 1,000 Genomes Phase 1
3) 246 Africans, 286 Asians, 381 Europeans, and 181 Americas from 1,000 Genomes Phase 1
4) GAW18 WGS data for 119 individuals that were not selected as the study sample

- The degrees of genetic similarity to the study sample from farthest to closest

Order of genetic similarity to the study sample

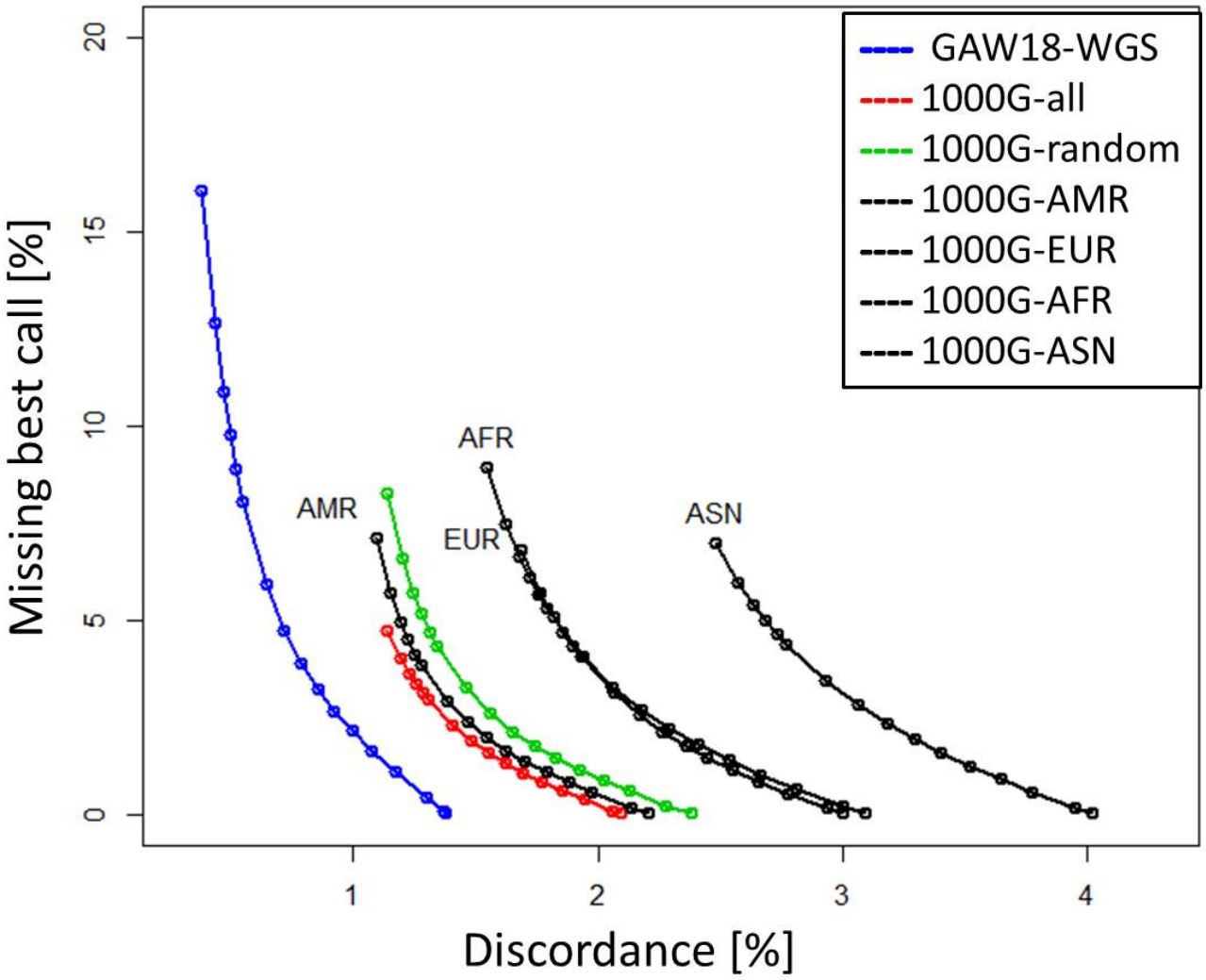
- Selected a set of uncorrelated SNPs ($r^2 < 0.2$)
- Computed genome-wide identity by state (IBS)
- Genome-wide IBSs to the study sample:

	AFR	All	Random	ASN	EUR	AMR	WGS
IBS	0.655	0.677	0.678	0.682	0.683	0.688	0.692

Imputation method used

- Software package IMPUTE2 (version 2.2.2) was used to impute SNPs
- IMPUTE2 provides probabilities for each probable genotype
- Under a given threshold, calculate the percentage of all imputed genotypes for which no probability exceeds the threshold (i.e., no call)
- Among calls, calculate the percentage of the best-guess imputed genotypes disagree with the observed WGS genotypes (i.e., discordance)

Results



For a reference panel, plot no call vs. discordance rates for calling thresholds ranging from 0.33 to 0.99

Discussion

- Reference panels can be obtained from publicly available databases, or from a two-stage approach where a subset of individuals in the study population are selected for whole genome sequencing.
- A reference panel that closely matches the ancestry of the study population can increase imputation accuracy, but it can also result more missing genotype calls.

- For the admixed study sample, the simple selection of a single best reference panel among HapMap African, European or Asian population is not appropriate. The composite reference panel combining all available reference data should be used.