The American Journal of Human Genetics

Supplemental Data

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure and Carlos D. Bustamante

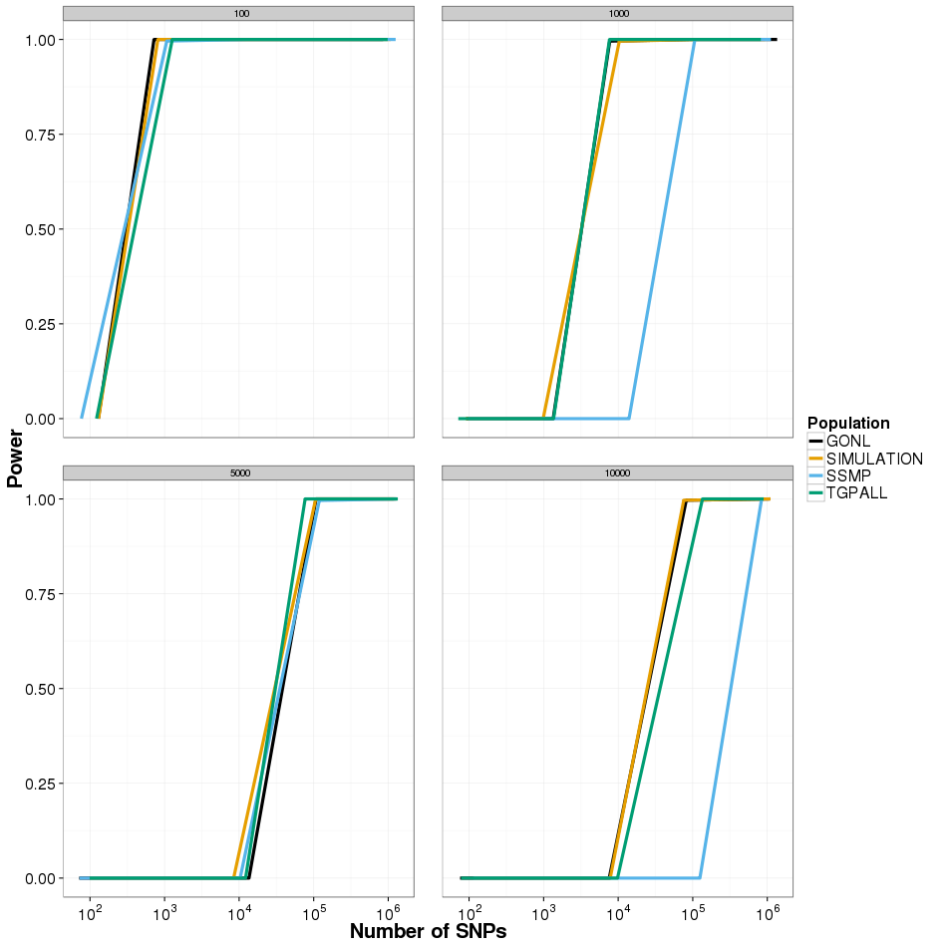

Figure S1: Theoretical power curves for SFS parameters estimated from different WGS datasets in Table S1 (TGPALL: 1000 Genomes Phase 1 WGS data). Panels show results for different beacon sizes. The curves are jittered to avoid overplotting.

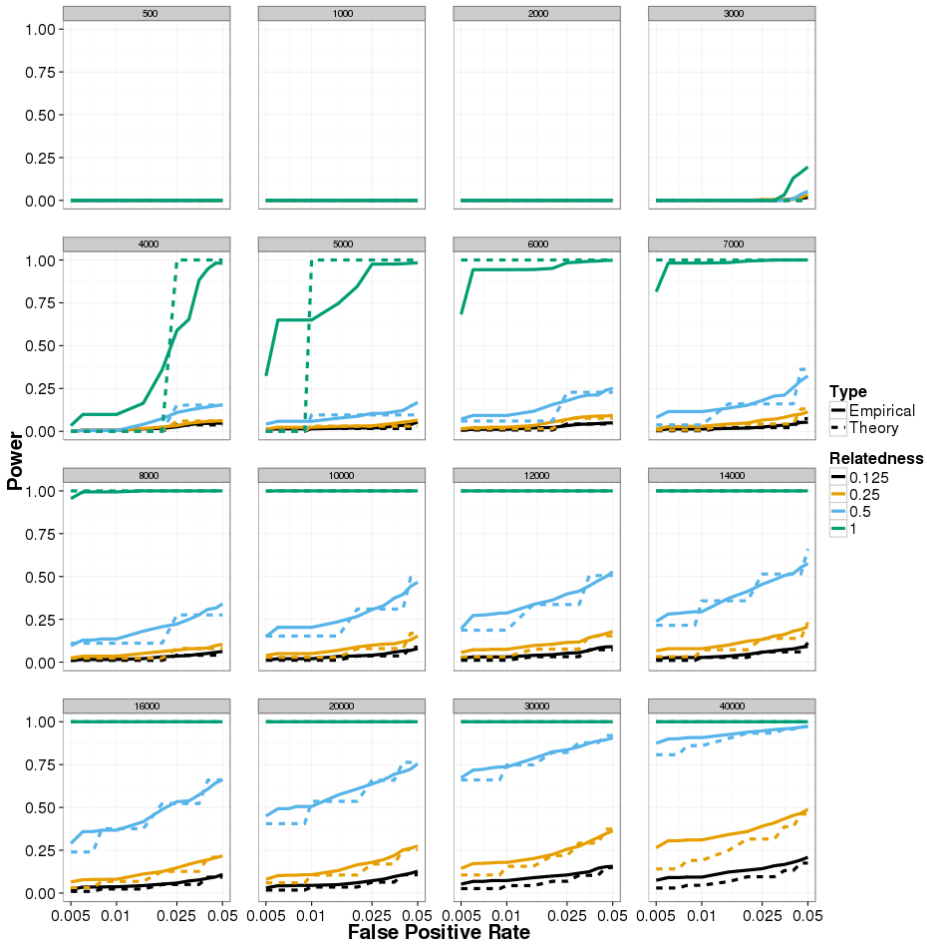

Figure S2: Receiver Operating Characteristic (ROC) curve of the LRT test for detecting relatives. Power at detecting relatives is diminished for detecting parents and siblings. Different panels show different number of SNPs queried. The X-axis is logarithmic in scale.

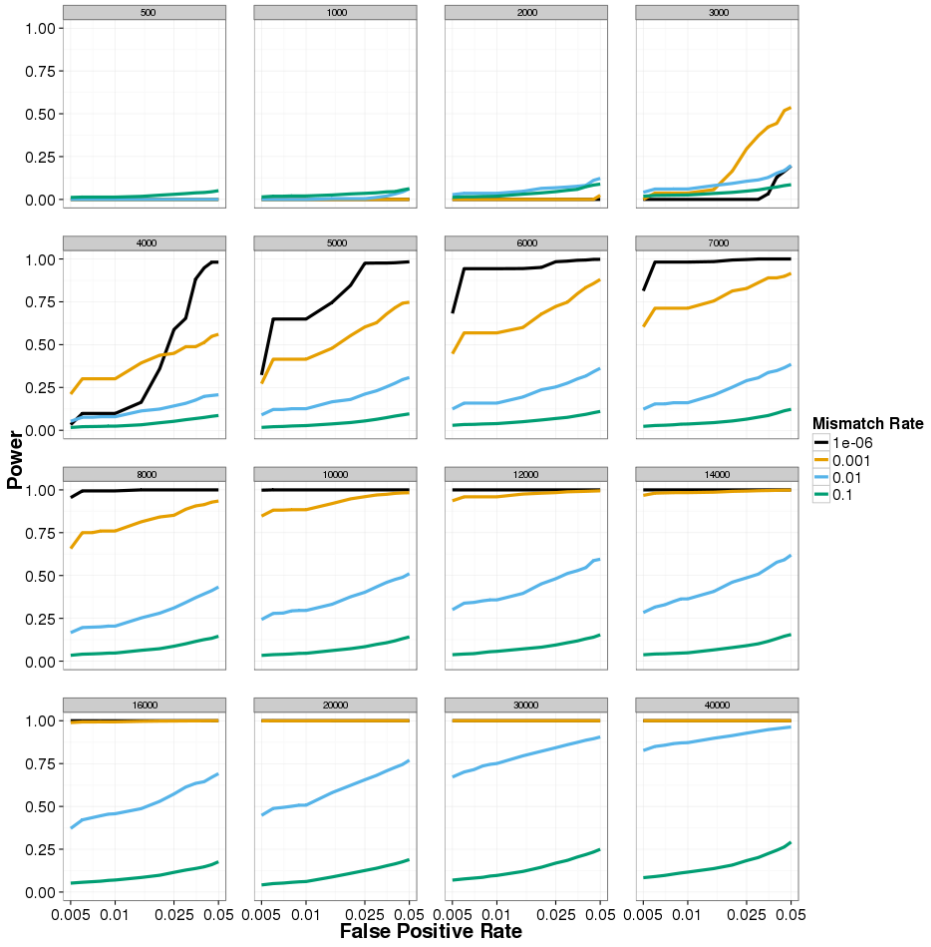

Figure S3: ROC curve of the LRT test for different error rates. Power is diminished with increasing errors. Different panels show different number of SNPs queried. The X-axis is logarithmic in scale.

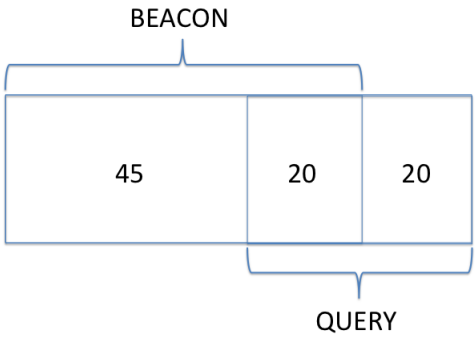

Figure S4: 1000 Genomes Phase 1 CEU beacon setup. Of the 85 CEU samples, 65 are used in the beacon. 20 samples from the beacon and the remaining 20 samples are used as query genomes.

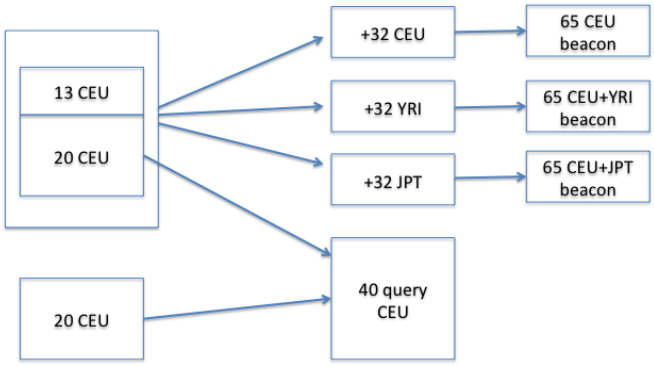

Figure S5: Multi-population beacon setup. 33 CEU individuals are common to each beacon and 32 are unique to each beacon. The same 40 CEU individuals are used to query each beacon.

Dataset	Number of samples	Estimated a'	Estimated b'
Simulation	1000	0.1300	1.1300
SSMP	100	0.1848	0.8500
GoNL	498	0.1131	0.8574
1000 Genomes Phase 1	1092	0.0735	1.0096
1000 Genomes Phase 1 Affymetrix array	1074	0.6483	1.2876

Table S1: Beta distribution parameters estimated from site frequency spectra for simulation data and some public whole-genome datasets. Only the SNP array data has considerably different parameters from the rest.

Mismatch rate δ	Error type modeled
10^{-6}	Ideal setting, almost zero noise
0.001	Genotype discordance between the same SNP in two replicates
0.01	Typical fraction of unique SNPs in two replicates
0.1	Upper bound on fraction of unique SNPs in two replicates

Table S2: Error types modeled by the mismatch rate δ

Beacon	Public data	Phenotype
1000 Genomes Project	Yes	N.A.
1000 Genomes Project Phase 3	Yes	N.A
AMPLab	Yes	N.A.
Broad Institute	No	Mixture of phenotypes
Cafe CardioKit	No	Cardiac diseases
ICGC	No	Cancer
Kaviar	No	Mixture of phenotypes
NCBI	No	Mixture of phenotypes
PGP	Yes	N.A.
IBD	No	Inflammatory bowel disease
Native American+Egyptian	No	None
UK10K	No	Mixture of phenotypes
SFARI	No	Autism spectrum disorder

Table S3: Phenotypes linked with the genomic data indexed by beacons. A dataset is denoted as "public" if individual-level genotype data can be downloaded without requesting access. Beacons which index non-public data are shown in bold. For beacons indexing public data, the phenotype is listed as "N.A." (not applicable).