In Silico Genotyping: Determining Genotypes in Pedigrees by Inference

Joshua T. Burdick, Weimin Chen, Gonçalo R. Abecasis, Vivian G. Cheung

Supplementary Note

Inferred genotypes for nuclear families

In nuclear families where low-resolution linkage data are available, most of the unobserved genotypes in offspring can be inferred by genotyping the parents and one of the offspring with high density markers. We applied our procedure to two generation CEPH families (we omitted information from the grandparents). In each family, we used sparse genotypes to determine IBD sharing between family members, and high-density genotypes on the parents and one child to infer genotypes for the remaining sibs. We obtained 93.7% of the missing genotypes. The proportion of inferred genotypes is higher (94% vs. 83%) than in the example with three-generation families where experimental genotypes are available for the grandparents and parents. This is because in the three generation families, an allele of the child's genotype could not be inferred if two grandparents and the corresponding parent were heterozygous; this can occur on the maternal and paternal sides or both. However, in a nuclear family, if the parents and one sib are genotyped with high-density markers and IBD sharing is known between the siblings, some of the sib genotypes can be inferred even when both parents and the genotyped child are heterozygous. These are the cases when the genotyped sib shares either zero or two of the alleles IBD with the "untyped" sib. For example, if both parents have AG genotypes at a marker, and one sib is AG and shares zero alleles identical-by-descent with his brother, then the brother's genotype must be AG.

Inferring additional genotypes using haplotype information

It is possible to infer some of the genotypes that cannot be inferred using family information alone by taking advantage of linkage disequilibrium at the population level. For example, when we used FastPHASE[1] to infer missing grandparental and parental genotypes, we were able to infer more genotypes than when we relied on family information alone.

Relative Efficiency Calculation

We estimated the "relative efficiency" of our method by comparing the median chi-squared statistics when data were analyzed using our proposed approach with two other designs that require the same number of experimental genotypes, but do not use inferred genotypes (Supplementary Tables 1 & 2). Alternative Design A uses the same number of families but no genotype inference; this allows us to compare efficiency with and without genotype inference in the same dataset. Alternative Design B uses fewer families but includes genotypes in all individuals for each family thus allowing us to evaluate the optimal design choice when transmission-disequilibrium tests[2] are used. For example, if 500 nuclear families with 3 offspring each are available, but funds are available for generating only 1500 high-density genotypes, analyzing two parents and one child in each family results in a median chi-squared of 13.41 (Alternative Design A), analyzing all individuals in 300 families results in a median chi-squared of 23.54 (Alternative Design B), and combining genotypes for two parents and one child in each family with inferred genotypes for the other children results in a median chi-squared of 32.19 (our approach). This corresponds to a relative efficiency (or increase in the effective sample size) of 2.40 versus Design A and 1.37 fold versus Design B (Supplementary Tables 1 & 2).

References

1. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet.* **78**, 629-44 (2006).
2. Abecasis, G.R., Cardon, L.R. & Cookson, W.O. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**, 279-92. (2000).