In Silico Genotyping: Determining Genotypes in Pedigrees by Inference

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

METHODS

Families studied. We used the CEPH families where all 4 grandparents and 2 parents were genotyped by the International HapMap Project [1]. There are 10 such families and their CEPH IDs are: 1334, 1340, 1341, 1350, 1362, 1408, 1420, 1447, 1454, 1459.

IBD determination. To determine the grandparental origins of each chromosomal region of every child, we used MERLIN[2]. Every individual was genotyped at 6,564 SNP markers. For every child, at each marker, we determine if the maternal allele was inherited from the grandmother or grandfather, and similarly for the paternal allele. The median inter-marker distance is only 205 kb (mean = 408 kb), so we assume that there are no undetected double crossovers between markers.

Genotype Inference. For three-generation families, using the IBD information and high density genotypes of the grandparents and parents from the International HapMap Project (http://www.hapmap.org; release 16c1), we inferred genotypes of children in the families as described in the text. For a marker with allele frequency p, we expect to recover $(1-2p^2(1-p)^2)^2$ of missing offspring

genotypes when IBD is known and parental and grandparental genotypes are available.

To apply the method to two generation families, we pretended that the CEPH families only have two generations (we ignored information from the grandparents). For each of the families, we used the genotypes of the parents from the International HapMap project and the inferred genotypes of one child to infer genotypes for the remaining sibs in the family. For a marker with allele frequency p, we expect to recover $1-(p^2(1-p)^2)$ of missing offspring genotypes when IBD is known and parental and one sib's genotypes are available.

The software for inferring missing genotypes is available at http://genomics.med.upenn.edu/ibdgenotype/.

QTDT analysis. The analysis was performed with SNP markers that were under the significant linkage peaks ($P < 4.3 \times 10^{-7}$) and have minor allele frequency ≥ 3% among the CEPH-HapMap grandparents. QTDT program, (http://www.sph.umich.edu/csg/abecasis/QTDT/) with the –ao, -wega options was used.

Simulation & Relative Efficiency. We simulated genotypes for 8 SNPs in a 1 cM region and, in addition, for a trait-determining SNP with minor allele frequency of 0.3 that explained either (H^2) 50% or 5% of the total phenotypic variance. The associated SNP was placed between the 4^{th} and the 5^{th} SNPs. We varied the

family structures and the number of genotyped individuals in each family and for each scenario. For each scenario, we first obtained the unobserved genotypes by inference and recorded the percentage of genotypes that could be inferred and the error rate in the procedure. Second, we estimated power with either: only the observed (experimentally determined) genotypes; or with a combination of observed and inferred genotypes.

1. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
2. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).