NIH Public Access
Author Manuscript
Nat Genet. Author manuscript; available in PMC 2010 December 21.

Published in final edited form as:
Nat Genet. 2006 September ; 38(9): 1002–1004. doi:10.1038/ng1863.

In silico method for inferring genotypes in pedigrees

Joshua T Burdick[1], **Wei-Min Chen**[2], **Gonçalo R Abecasis**[2], and **Vivian G Cheung**[1,3]

[1]Department of Pediatrics, University of Pennsylvania, 3516 Civic Center Blvd., Philadelphia, Pennsylvania 19104, USA.

[2]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA.

[3]Department of Genetics, University of Pennsylvania, 3516 Civic Center Blvd., Philadelphia, Pennsylvania 19104, USA.

Abstract

Our genotype inference method combines sparse marker data from a linkage scan and high-resolution SNP genotypes for several individuals to infer genotypes for related individuals. We illustrate the method's utility by inferring over 53 million SNP genotypes for 78 children in the Centre d'Etude du Polymorphisme Humain families. The method can be used to obtain high-density genotypes in different family structures, including nuclear families commonly used in complex disease gene mapping studies.

Even though groups such as The SNP Consortium[1] and the International HapMap Consortium[2,3] have identified millions of polymorphic markers and stimulated the development of high-throughput genotyping techniques[4–6], genotyping of polymorphic markers remains a labor-intensive and costly step in genetic mapping studies. To decrease the cost of family-based genetic studies, we developed a computational approach that uses high-density genotype data for a subset of individuals in a pedigree to infer genotypes for the remaining relatives (see http://genomics.med.upenn.edu/genotypeinference and http://www.sph.umich.edu/csg/abecasis/Merlin/ for the software). This approach greatly reduces the amount of conventional 'wet-lab' experimentation required to carry out association analysis in pedigrees.

Many gene mapping projects use a tiered approach: first, genome-wide linkage analysis is carried out using widely spaced markers across the genome; then, genotypes are determined for many more markers near observed linkage peaks and are tested by association analysis. Our approach reduces work in the second stage because experimental genotyping is required for only a subset of individuals. Genotypes for the remaining individuals are obtained in two steps. First, low-resolution genotypes from linkage analysis are used to identify regions of shared identity-by-descent (IBD) between relatives. Then, with information on IBD sharing between individuals and high-density genotype data on some members of the family, we infer most of the unobserved high-density genotypes for the remaining individuals.

© 2006 Nature Publishing Group

Correspondence should be addressed to V.G.C. (vcheung@mail.med.upenn.edu).

Note: Supplementary information is available on the Nature Genetics website.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To illustrate this procedure, we used it to infer genotypes for the children in ten Centre d'Etude du Polymorphisme Humain (CEPH)–HapMap pedigrees. All the grandparents and parents of these pedigrees have been genotyped at about 1 million SNP markers in Phase I of the International HapMap Project[3]. First, we used genotypes of 6,564 genetic markers obtained previously on all individuals to determine the grandparental origin for every chromosomal segment in each child. Specifically, for each child and at every marker, we considered the allele from the mother and determined whether that allele was inherited from a transmitted chromosome that originated in the maternal grandfather or grandmother; we did the same for the paternal side. Results from adjacent markers allow us to confirm the grandparental origins of each genomic region (Fig. 1a). This step can be accomplished with existing pedigree analysis packages[7–11]. In the second step, we inferred unobserved genotypes in the children by combining information from the first step, which describes the genome of each child as a mosaic of the grandparental chromosomes, with high-density genotypes of the grandparents and parents (Fig. 1a). For example, at a particular SNP, suppose that the low-resolution genotypes show that the child inherited the chromosomal segment containing this SNP from the paternal and maternal grandfathers, and the high-resolution genotypes show the haplotypes transmitted from these grandparents carry alleles A and C, respectively; then the child's genotype must be AC.

When we applied this procedure to infer genotypes for children in ten CEPH-HapMap pedigrees, we obtained 53,666,501 genotypes, an average of 688,032 marker genotypes for each of 78 children (range: 629,731 to 698,165). The average of 688,032 inferred genotypes per child corresponds to ~83% of all the genotypes that can be obtained (the average number of genotypes available on each grandparent and parent in release 16 of the HapMap data is 832,703). Some genotypes were not inferred because the markers were located in regions where IBD sharing information was uncertain. In other cases, even though fully informative IBD information was available, the two grandparents in the maternal or paternal side (or both) and the corresponding parent were heterozygous at a SNP, so it was impossible to determine which alleles were transmitted. These results closely match analytical expectations: theoretically, we would expect to be able to infer ~97%, 83% and 77% of genotypes for SNPs with minor allele frequencies of 0.10, 0.30 and 0.50, respectively (Supplementary Methods online).

To determine the accuracy of the method, we compared the inferred genotypes with those generated experimentally by PCR-based SNP genotyping. Among the 3,210 genotypes in which both inferred and experimental genotypes were available, seven (0.2%) were discordant. Even if the inferred genotypes were incorrect in all seven discrepant cases, the error rate from inference would still be very low and comparable to the error rate obtained by experimental genotyping in the HapMap Project.

Next, we used the inferred genotypes to test for evidence of linkage and association of candidate transcriptional regulators with gene expression phenotypes. Previously, we had performed genome-wide linkage analyses to determine the chromosomal locations linked to the expression levels of genes[12].With the inferred genotypes, we performed family-based association analysis using the transmission disequilibrium test (QTDT)[13,14] with markers within the significant linkage peaks. As the linkage peaks are quite broad, we would have needed to perform millions of genotyping reactions. The inferred genotypes, however, allowed us to analyze a large number of parent-offspring transmissions without having to carry out any additional genotyping reactions. We illustrate this with ten expression phenotypes for which we have previously found highly significant linkage evidence for *cis*-acting regulators. We identified markers located under each significant linkage peak (pointwise $P < 4 \times 10^{-7}$) and carried out QTDT analysis with genotypes for (i) 30 genotyped CEPH-HapMap trios and (ii) the genotyped 30 CEPH-HapMap trios augmented with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

inferred genotypes of children in ten CEPH families. In each case, QTDT results confirmed the linkage findings and narrowed the candidate regions. However, results with the inferred genotypes included were more significant than the analysis with the 30 HapMap trios alone (Fig. 1b and Table 1). With just the 30 HapMap trios, for many phenotypes, there were not enough informative offspring to carry out the analysis. In the remaining cases, the findings of *cis* association were modest. With the inferred genotypes, we observed several-fold increases in χ^2 values (and therefore in effective sample size).

Simulations summarized in Table 2 show that the substantial increase in power is expected whether analyzing a variant that has a strong effect (such as the *cis*-acting variants for gene expression phenotypes examined above) or a weaker effect (as would be expected for most complex traits). The simulated data also show that genotyping one offspring per family with high-density markers further increases the power to very near what would be achieved if all the children in each family were genotyped (see rows 3 and 5 in Table 2).

Although the examples above focus on three-generation families, our method can be extended to other settings. For example, in nuclear families in which 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) and obtained 93.7% of the missing genotypes (Supplementary Note and Supplementary Tables 1 and 2 online). We confirmed these findings using simulated data (rows 6–8 in Table 2).

Gene mapping projects often begin with a linkage study with relatively sparse markers. When candidate regions are found, they are further investigated by association analysis. Because association studies require a dense set of markers, the cost of conventional genotyping can be very high. Here, we show that high-density genotypes can be inferred for the relatives of genotyped individuals with greatly reduced ‘wet lab’ experimentation. Of course, in some cases not all unobserved genotypes can be obtained, as haplotype phase may remain uncertain, or genotypes from a previous scan may not be available. In these cases, it is still possible to estimate a probability distribution for each of the unobserved genotypes conditional on the observed genotype data for the pedigree. It is then possible to carry out association tests that use these probability distributions in place of observed genotypes; these tests can extract information even from individuals whose genotype is uncertain (W.C. and G.R.A., unpublished data).

In silico genotype inference provides a cost-effective way to scan many existing family collections for association, either genome-wide or within candidate genes or regions. All that is required is to genotype several well-chosen individuals in each family at very high density. This approach will facilitate genome-scale family-based association studies and, thus, the identification of susceptibility genes for complex diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Devlin and M. Morley for technical assistance, M. Boehnke and R. Spielman for discussion and critical reading of the manuscript. This work was supported by grants from the US National Institutes of Health to G.R.A. and V.G.C.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

1. Matise TC, et al. Am. J. Hum. Genet 2003;73:271–284. [PubMed: 12844283]
2. International HapMap Consortium. Nature 2003;426:789–796. [PubMed: 14685227]
3. The International HapMap Consortium. Nature 2005;437:1299–1320. [PubMed: 16255080]
4. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. Nat. Genet 2005;37:549–554. [PubMed: 15838508]
5. Syvanen AC. Nat. Genet 2005;37:S5–S10. [PubMed: 15920530]
6. Matsuzaki H, et al. Nat. Methods 2004;1:109–111. [PubMed: 15782172]
7. Sobel E, Lange K. Am. J. Hum. Genet 1996;58:1323–1337. [PubMed: 8651310]
8. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Nat. Genet 2002;30:97–101. [PubMed: 11731797]
9. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Am. J. Hum. Genet 1996;58:1347–1363. [PubMed: 8651312]
10. S.A.G.E. Statistical Analysis for Genetic Epidemiology. Cork, Ireland: Statistical Solutions; 2003.
11. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Nat. Genet 2000;25:12–13. [PubMed: 10802644]
12. Morley M, et al. Nature 2004;430:743–747. [PubMed: 15269782]
13. Abecasis GR, Cardon LR, Cookson WO. Am. J. Hum. Genet 2000;66:279–292. [PubMed: 10631157]
14. Abecasis GR, Cookson WO, Cardon LR. Eur. J. Hum. Genet 2000;8:545–551. [PubMed: 10909856]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

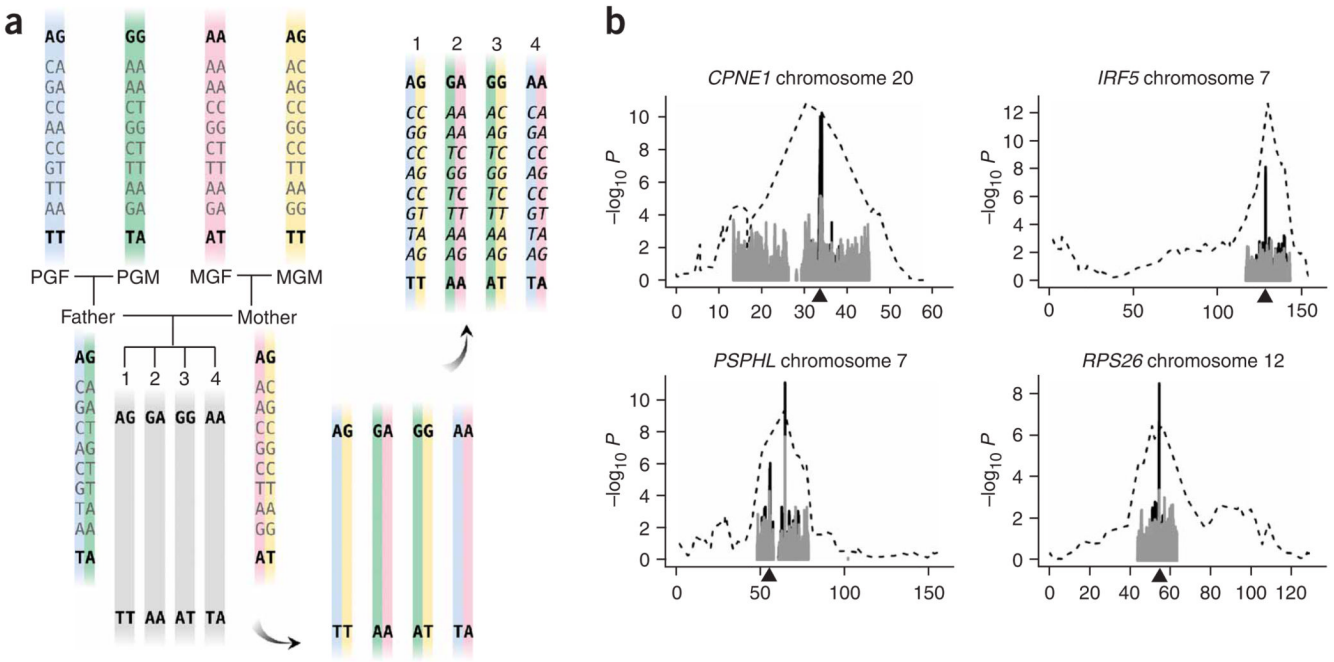

Figure 1.

Genotype inference. (**a**) Inferred genotypes for eight SNPs. The inferred genotypes for each child are shown in italics. To determine the inferred genotypes, we identified regions of shared IBD (color-coded) between the child and her parents and grandparents using low-density genotypes (boldface) and used-high density genotypes (light gray) of the grandparents and parents. For illustration purposes, the grandparental genotypes are shown as phased. In practice, this is resolved by an IBD estimation program. (**b**) Comparison of data from QTDT analysis for four representative expression phenotypes. For each phenotype, results from linkage analysis based on 14 CEPH families (dotted line), QTDT analysis with genotypes for 30 trios from the HapMap Project (gray bars) and with the combined HapMap and inferred genotypes (black bars) are shown. For the linkage analysis, the null hypothesis is no linkage, and for the QTDT, the null hypothesis is no linkage or no association.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Comparison of QTDT results without and with inferred genotypes

		30 CEPH-HapMap trios 'without inferred genotypes'			30 CEPH-HapMap trios and ten CEPH families 'with inferred genotypes'			
Phenotype	**SNP**[a]	χ^2	***P***	**Number of informative offspring**[c]	χ^2	***P***	**Number of informative offspring**[c]	**Increase in** χ^2[d]
CSTB	rs2838393	NT	NT	< 20	19.3	1×10^{-5}	49	–
CTSH	rs1036937	4.89	0.027	20	15.52	8×10^{-5}	82	3.2
PPAT	rs2139512	NT	NT	< 20	15.49	8×10^{-5}	57	–
PSPHL	rs2419485	NT	NT	< 20	46.7	8×10^{-12}	31	–
AA827892	rs3752279	NT	NT	< 20	40.09	2×10^{-10}	30	–
HLA-DQB1	rs9275141	10.5	0.0012	22	33.08	9×10^{-9}	77	3.2
RPS26	rs2271194[b]	12.43	4×10^{-4}	20	35.43	3×10^{-9}	71	2.9
CPNE1	rs11167280	NT	NT	< 20	42.73	6×10^{-11}	40	–
IRF5	rs7789423	1.28	0.2586	24	33.43	7×10^{-9}	81	26.1
LRAP	rs2762	NT	NT	< 20	88.51	5×10^{-21}	70	–

NT: tests were not carried out when the number of informative offspring is < 20.

[a]Markers listed are those with the most significant evidence of association when inferred genotypes were used.

[b]For *RPS26*, the same marker was identified as most significant with and without the inferred genotypes.

[c]Informative offspring are individuals with at least one heterozygous parent.

[d]Ratio of χ^2 with inferred genotypes to χ^2 without inferred genotypes.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Comparison of simulation results

Family structures, with selected individuals genotyped for candidate SNPs	*In silico* genotyping			Power (%)			
	Missing genotypes (% of total)	Percentage of missing genotypes inferred	Error rate (%)	Simulated[a] H^2	Using observed genotypes	Using inferred genotypes	After genotyping all individuals
CEPH pedigrees in Table 1	46.4	82.4	0.4	0.5	0.0	83.9	97.1
Ten three-generation pedigrees with eight offspring each							
Grandparents and parents genotyped	57.1	82.4	0.4	0.5	0.0	74.1	92.2
Grandparents, parents and one child genotyped	50.0	99.2	0.7	0.5	0.0	91.2	92.2
100 three-generation pedigrees with eight offspring each							
Grandparents and parents genotyped	57.1	82.4	0.4	0.05	0.3	44.3	82.2
Grandparents, parents and one child genotyped	50.0	99.2	0.7	0.05	1.3	70.8	82.2
500 nuclear families with three offspring each							
Parents and one child genotyped	40.0	94.6	1.1	0.05	4.7	62.1	82.7
Parents and two children genotyped	20.0	96.9	0.6	0.05	40.8	77.4	82.7
500 nuclear families with two offspring each							
Parents and one child genotyped	25.0	94.7	1.0	0.05	4.5	27.3	41.2

[a]We simulated genotypes for eight SNPs within an 8-cM region and a trait-determining SNP with a minor allele frequency of 0.3 that explained either 50% or 5% (H^2) of the total phenotypic variance. The associated SNP was placed between the 4th and the 5th SNP (for details, see Supplementary Methods).