European Journal of Human Genetics (2005) 13, 898–901
© 2005 Nature Publishing Group All rights reserved 1018-4813/05 $30.00
www.nature.com/ejhg

LETTER

Typing without calling the allele: a strategy for inferring SNP haplotypes

European Journal of Human Genetics (2005) **13,** 898–901. doi:10.1038/sj.ejhg.5201440; published online 18 May 2005

There is an ongoing discussion in the literature regarding the value of phase data for detecting association to disease genes.[1] For some analyses,[2,3] haplotype information is the key for more powerful studies. Unfortunately, experimental methods for haplotype determination in unrelated individuals often require manipulation of separated single chromosomes or of large fragments thereof,[4] a technically complicated and cost prohibitive procedure. Meanwhile, current technologies offer an array of practical molecular techniques for determination of genotype data, without assignment of distinct alleles to the chromosomes that carry them.[5] As a consequence, the study design of choice is to infer haplotype architecture from genotype data, rather than measure it directly. Haplotype resolution, whether in genetic disease studies or in preceding efforts to map haplotypes, is currently approached by statistical or computational methods,[6–12] by a process called phasing or haplotyping. This strategy uses computation to make up for less informative, yet more practical and cost-effective experimentation (see Figure 1).

In this work, we show how additional reduction in the complexity of experimentation can be achieved with only minor effect on the informativeness of measurements: we explore the conditions under which a yet less detailed and so far unexploited type of data, may become attractive owing to its potentially reduced costs. This level of information, hereby called *xor-genotype*, differentiates only between homozygous and heterozygous states at biallelic loci such as single nucleotide polymorphisms (SNPs), and provides no data as to the allele identity of homozygotes. Several techniques are already available for xor-genotype determination. One such example is DHPLC,[13] which can determine whether an individual is homozygous or heterozygous for each SNP, but without additional information cannot readily distinguish between the two alternative homozygous states. (The name xor-genotypes originates from the logical 'exclusive-or', or xor operation, which like our typing can only distinguish if a pair of binary elements, the alleles, are identical or not). Moreover, since xor-genotypes contain less information than genotypes, development of cost-competitive methods for xor-genotyping may be possible.

We recently demonstrated[14] that computation can, under defined circumstances and with minimal genotype data, resolve xor-genotyping into individual genotypes or haplotypes (see Figure 1). Therein we described algorithms for reconstruction of the complete haplotype information from the partial, xor-genotype data, a computation we termed Xor Perfect Phylogeny Haplotyping (XPPH). We show here that computation can render xor-genotypes nearly as informative as full genotypes, at possibly a fraction of the cost, and with only a minor increase in the number of individuals that need to be examined. We base our theoretical analysis on the observation that large blocks of the genome have not undergone any significant recombination[4,15,16] and therefore have essentially evolved according to the no-recombination infinite-site coalescence model.[17] This is powered by the forces of selection and demography, which are likely to have reduced diversity within the human genome, leading to a lack of directly observable recombination events or branches of the genealogy tree. This allows use of the perfect phylogeny assumption[8] for haplotype structure, which guarantees that the haplotypes can fit into a tree that describes their genealogical ancestry. In such a tree, the path between two haplotypes is labeled by the SNPs that they are heterozygous for (See Figure 2b). This tree can be inferred from data while creating a catalog of variation (a haplotype map), and can be utilized in subsequent genetic studies.

Figure 2 illustrates the conceptual process of haplotype resolution from xor-genotype data, including alternating stages of data collection and computation in each 4-gamete block.

(i) Collect DNAs for a haplotype mapping cohort and determine the xor-genotype of each individual in this sample. Apply XPPH computation to resolve the structure of the perfect phylogeny tree.
(ii) Identify at most three sample individuals that need to be genotyped (see[14] for details). Genotype the selected individuals. Embedding these genotypes in the tree structure allows their homozygous alleles to resolve all haplotypes for the corresponding SNPs based on the

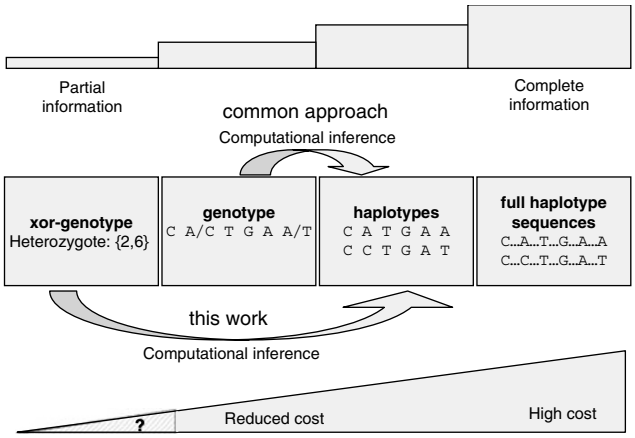

Figure 1 Levels of genetic information. The spectrum of levels of genetic information of an individual, depicted along a 6-SNP region: full haplotype sequences provide full information (including previously unknown SNPs); SNP haplotypes separate the alleles at known SNP sites into the two chromosomes; genotypes provide only the confluence of the allele identities at each SNP site, and xor-genotypes (introduced here) only tell which sites are heterozygous. Arrows indicate that haplotypes are computationally inferred from genotypes, and can also be computationally inferred from xor-genotypes by a process called xor-haplotyping (as shown in this work). Information decreases from full haplotype sequencing to xor-genotype determination (top). Costs diminish from full haplotypes to genotypes, and may potentially decrease further if only xor-genotypes are measured (bottom).

perfect phylogeny structure. This gives rise to a structured haplotype map of the block at hand.

(iii) Collect xor-genotype data for individuals in a disease study. Resolve the haplotypes of heterozygous individuals by the perfect phylogeny.

(iv) Genotype individuals who are completely homozygous in the current block.

Our implementation for XPPH computation (in step i), is available for public use at http://www.cs.tau.ac.il/~rshamir/greal/. We note that our method can detect 4-gamete blocks from xor-genotypes by sliding window analysis, and in principle can be modified to produce a chain of block-wise haplotypes along the genome, as done for genotyping data (eg Eskin *et al*[18]).

The economic advantage of our approach over regular genotyping will depend on the relative cost of regular *vs* xor-genotyping. Suppose the cost of regular genotyping one SNP in an individual is 1 unit, and the cost of its xor-genotyping is β. If we type t SNPs in N individuals, regular genotyping will cost Nt. If we xor-genotype them instead, we pay $\beta Nt + 3t$ (steps i and ii), but the overall cost also has to take into account the chances of XPPH failure (step iii) and the homozygosity rate (step iv). Chances of failure are $<10\%$ for $N>50$ (see Figure 3), and decrease as N grows; and a conservative estimate for homozygosity rate of a complete haplotype is $<30\%$ (as expected with 4–6 common haplotypes.)[16] In total, the expected cost of the whole strategy is at most $0.9(\beta Nt + 0.3Nt + 3t) + 0.1(\beta Nt + Nt)$. Hence, whenever $\beta < 0.6$ (and irrespective of t) the cost of the xor-genotyping strategy will be preferable to the cost of the standard approach. Note, that this estimate is conservative as it does not exploit allele correlation that crossblock boundaries. Such correlation may resolve homozygous haplotypes by information from adjacent blocks at no additional cost.

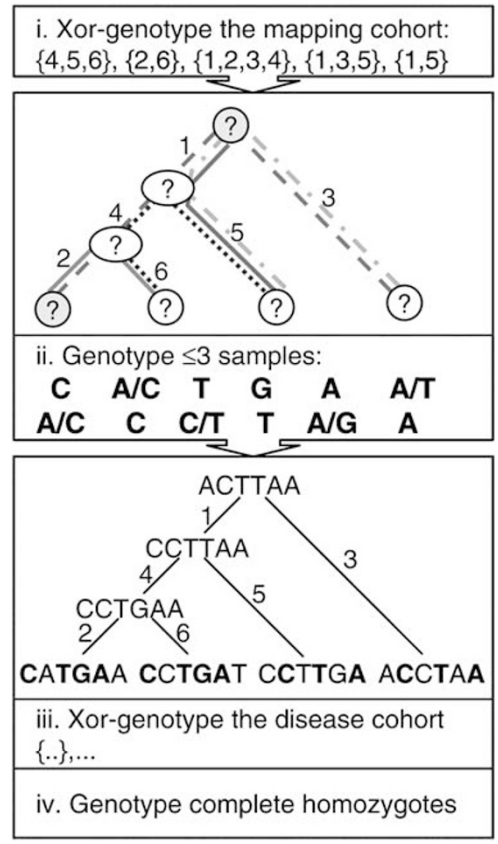

Figure 2 Haplotype resolution from xor-genotypes. Schematic example of the proposed paradigm for haplotype resolution from xor-genotype data, during the creation of a haplotype map (**a** and **b**) and its application in genetic disease studies (**c** and **d**). (**a**) The xor-genotypes (lists of heterozygous sites) of a five-individual sample are measured. XPPH computation reveals the perfect phylogeny structure that gave rise to the (yet unresolved) sample haplotypes (circles) and their evolutionary intermediates (ovals). Each xor-genotype corresponds to a path in the perfect phylogeny tree. (**b**) In order to resolve haplotype sequences the genotypes of two individuals are measured. Their homozygous alleles resolve all unknown haplotypes. (**c**) Xor-genotypes of the disease cohort are measured, and heterozygous samples are readily translated to haplotypes using the map (perfect phylogeny) information. (**d**) Completely homozygous individuals need to be genotyped.

Large-scale efforts like the HAPMAP projects,[19,20] are genotyping enough individuals to detect the regions where the perfect phylogeny description is most accurate and to

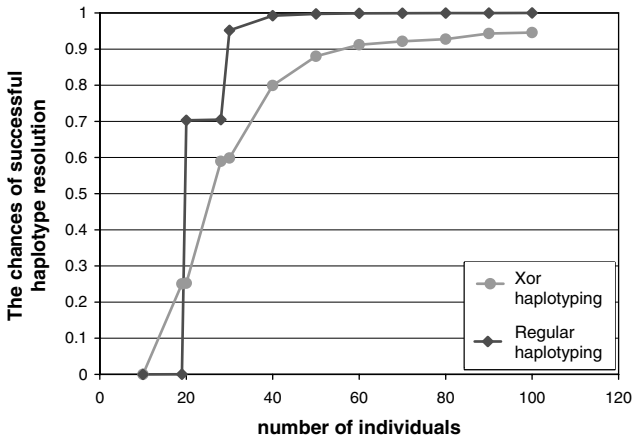

Figure 3 The power of xpph. The rate of successful haplotype resolution (*y*-axis) *vs* the number of individuals samples (*x*-axis). Each point (pink) represents an average success rate in XPPH computation on 5000 simulated samples of 50 common SNP haplotypes (minor allele frequency >0.05) for prescribed-size cohort, simulated by an infinite-site, no-recombination model[21] on a typical, 25 kb-long haplotype block of 50 SNPs. Statistics of haplotype recovery from regular genotypes were similarly obtained in Reference[22] are plotted for comparison (blue). For a cohort of size 60, XPPH guarantees $\sim$90% chance for a complete successful solution. Similar results were obtained also in simulations that assume exponential population expansion (data not shown).

describe the tree on these regions. If the full perfect phylogeny tree of a region is already in hand, the resolution of a single xor-genotype into haplotypes is unique (due to the 4-gamete rule) and easily obtained. Therefore the xor-haplotyping approach offers further potential advantages for future studies that exploit the high-resolution haplotype maps that will soon be available.

In summary, we introduced here a novel type of genetic data called xor-genotypes, and described computational tools to resolve haplotypes based primarily on xor-genotype data. We argued for the potential economical advantage of xor-genotypes over the full genotypes common today. We showed that simulated genetic data in blocks indicates that xor-genotypes are nearly as informative as full genotypes, potentially at a fraction of the cost, both for haplotype mapping and for genetic disease studies, and determined the boundaries under which these conditions hold. Additionally, our computational work[14] revealed that xor-genotypes and their perfect phylogeny provide important insights even if full genotypes were obtained, and, for instance, allow selection of tag SNPs without phased data. Hence, genotyping methods that distinguish only between heterozygotes and homozygotes, together with the appropriate computational solutions, may offer a cost-effective alternative to genotyping in genetic studies, and may play a role in whole genome mapping approaches.

Acknowledgements

We thank Orna Man for helpful discussions. Jacques S Beckmann is the Hermann Mayer Professor of Molecular Genetics. Ron Shamir was supported in part by the Israel Science Foundation (Grant 309/02). Itsik Pe'er was supported by a grant from the Israeli Ministry of Science.

Tamar Barzuza*[,1], Jacques S Beckmann[2,3], Ron Shamir[4] and Itsik Pe'er[5]

[1]*Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel;*
[2]*Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel;*
[3]*Département de Génétique Médicale, CHUV-UNIL, Ch-1011 Lausanne, Switzerland;*
[4]*School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel;*
[5]*Medical and Population Genetics Group, Broad Institute, Cambridge MA 02142, USA*

**Correspondence: T Barzuza, c/o Professor Ron Shamir, School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. Tel: +972 3 640 5383; Fax: +972 3 640 5384; E-mail: tamar.barzuza@weizmann.ac.il*

References

1 Chapman JM, Cooper JD, Todd JA *et al*: Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Heredity* 2003; **56**: 18–31.
2 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
3 Lin S, Chakravarti A, Cutler DG: Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* 2004; **36**: 1181–1188.
4 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
5 Kwok PY: Genetic association by whole-genome analysis. *Science* 2001; **294**: 1669–1670.
6 Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990; **7**: 111–122.
7 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12**: 921–927.
8 Gusfield D: Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. *Proceedings of the Sixth Annual International Conference on Computational Biology (RECOMB'02)* 2002, pp 166–175.
9 Bafna V, Gusfield D, Lancia G, Yooseph S: Haplotyping as perfect phylogenty: a direct approach. *Technical Report U.C. Davis*, CSE-2002-21, 2002.
10 Eskin E, Halperin E, Karp RM: Efficient reconstruction of haplotype structure via perfect phylogeny. *J Bioinform Comput Biol (JBCB)* 2003; **1**: 1–20.
11 Pe'er I, Beckmann JS: Resolution of haplotypes and haplotype frequencies from SNP genotypes of pooled samples. *Proceedings of the Seventh Annual International Conference on Computational Biology (RECOMB '03)* 2003; 237–246.
12 Kimmel G, Shamir R: Maximum likelihood resolution of multi-block genotypes. *Proceedings of the Eighth Annual International*

Conference on Computational Biology (RECOMB '04) 2004; 282–289.
13 Xiao W, Oefner PJ: Denaturing high-performance liquid chromatography: A review. *Hum Mutat* 2001; **17**: 439–474.
14 Barzuza T, Beckmann SJ, Shamir R, Pe'er I: Computational Problems in Perfect Phylogeny Haplotyping: Xor-Genotypes and Tag SNPs. *The Fifteenth Annual Symposium on Combinatorial Pattern Matching (CPM)* 2004; 15–27.
15 Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001; **29**: 217–222.
16 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in human genome. *Science* 2002; **296**: 2225–2229.
17 Hudson RR: Properties of a neutral allele model with intragenic recombination. *Theoret Population Biol* 1983; **23**: 183–201.
18 Eskin E, Halperin E, Karp R: Large Scale Reconstruction of Haplotypes from Genotype Data. *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB)* 2003; 104–113.
19 www.hapmap.org.
20 Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
21 Hudson RR: Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002; **18**: 337–338.
22 Chung RH, Gusfield D: Empirical Exploration of Perfect Phylogeny Haplotyping and Haplotypers. *Proceedings of the Cocoon Conference* 2003; 5–19.