Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies

Julie A. Douglas[1,5], Michael Boehnke[2,5], Elizabeth Gillanders[6], Jeffrey M. Trent[6] & Stephen B. Gruber[3,4,5]

Published online: 9 July 2001, DOI: 10.1038/ng582

The study of complex genetic traits in humans is limited by the expense and difficulty of ascertaining populations of sufficient sample size to detect subtle genetic contributions to disease. Here we introduce an application of a somatic cell hybrid construction strategy called conversion[1–4] that maximizes the genotypic information from each sampled individual. The approach permits direct observation of individual haplotypes, thereby eliminating the need for collecting and genotyping DNA from family members for haplotype-based analyses. We describe experimental data that validate the use of conversion as a whole-genome haplotyping tool and evaluate the theoretical efficiency of using conversion-derived haplotypes instead of conventional genotypes in the context of haplotype-frequency estimation. We show that, particularly when phenotyping is expensive, conversion-based haplotyping can be more efficient and cost-effective than standard genotyping.

Haplotype data are extremely valuable in studies of linkage disequilibrium, particularly in tracking chromosomal segments that may harbor disease-susceptibility genes. Constructing haplotypes from conventional genotype data is complicated, however, because for any autosomal genotype, the maternal and paternal origins of the two constituent alleles are not directly observed. Hence, for unrelated individuals, the presence of two or more heterozygous genotypes guarantees haplotype uncertainty. Current haplotyping approaches include direct inference from family data, statistical estimation of haplotype frequencies generally via the expectation-maximization (EM) algorithm[5–8] or molecular haplotyping of very short segments of DNA by methods such as long-range polymerase chain reaction (PCR)[9]. Each of these approaches, however, is often an imperfect substitute for knowing an individual's full-length haplotypes. For example, for haplotypes consisting of just three single nucleotide polymorphisms (SNPs), genotypes of father-mother-offspring trios can fail to reveal offspring haplotypes up to 24% of the time[10]. In contrast, conversion has the potential for direct haplotype construction of individual haplotypes regardless of the availability or informativity of pedigree data.

The basis of the conversion strategy is the transformation of diploid to haploid cells by construction of somatic cell hybrids[11], capitalizing on the well-recognized observation that somatic cell hybrids often retain only a subset of human chromosomes. Although conversion was originally introduced to detect sequence mutations[1–4], it also provides an opportunity for direct haplotype construction through conventional genotyping of DNA from haploid cells. somatic cell hybrids have been used before to identify mutations[12] and even haplotypes[13] (M. Burmeister, pers. comm.), but their use has been restricted to a very limited number of subjects and chromosomal regions because of the inefficiencies and variations in fusion and selection conditions. Conversion, however, is a robust, high-throughput strategy to efficiently isolate any or all human chromosomes for analysis of haploid DNA.

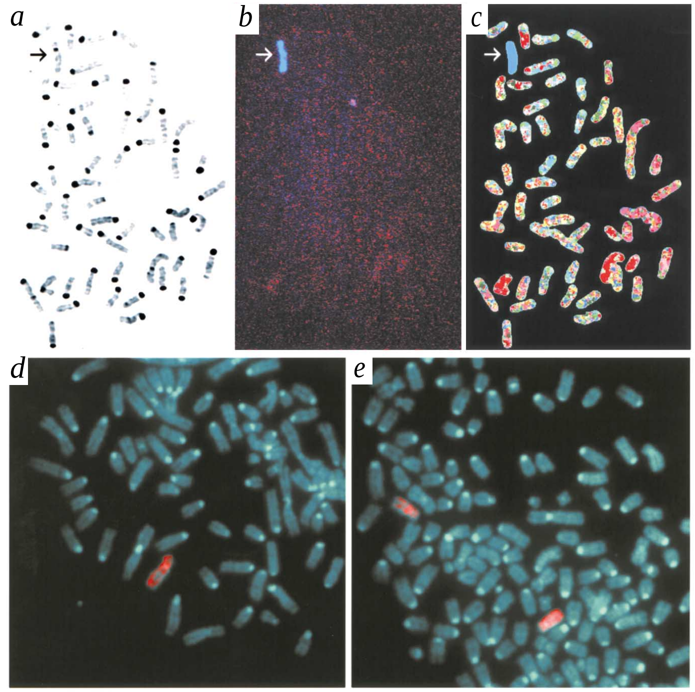

Fig. 1 Human chromosomal composition of mouse-human somatic cell hybrids. ***a*–*c***, Hybridization on metaphase chromosomes in a three-step fluorescence *in situ* hybridization (FISH) procedure (arrows indicate human chromosome 5). A 4',6-diamidino-2-phenylindole (DAPI) reverse staining pattern (to generate a G-banding pattern) is carried out (*a*), followed by analysis of the same chromosome by a whole-chromosome-painting probe (WCP) for human chromosome 5 (*b*). Finally, multicolor spectral karyotyping (SKY)[18] analysis of the same cell shows a single copy of human chromosome 5 (*c*). ***d*–*e***, Chromosome composition of a second monoallelic hybrid as defined by FISH with a WCP for human chromosome 5 detects a single copy in a near-diploid cell (*d*) and two copies of the same chromosome in a polyploid cell (*e*). Genotype analysis with polymorphic markers on chromosome 5 confirmed the derivation of these hybrids as monoallelic for that chromosome (data not shown). FISH images clearly show no detectable interspersed chromosome 5 material (recognized at the level of FISH); note that an additional 50 cells visually scored by FISH also showed no evidence of gross chromosomal rearrangements.

Departments of [1]Human Genetics, [2]Biostatistics, [3]Epidemiology and [4]Internal Medicine, and [5]Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. [6]National Human Genome Research Institute, Bethesda, Maryland, USA. Correspondence should be addressed to M.B. (e-mail: boehnke@umich.edu).

Table 1 • Fraction of hybrids null, monosomic and disomic for a given chromosome

Chromosome	Null hybrids	Monosomic hybrids	Disomic hybrids
1	0.32	0.22	0.19
2	0.26	0.23	0.40
3	0.29	0.24	0.28
4	0.26	0.29	0.29
5	0.27	0.22	0.29
6	0.23	0.29	0.24
7	0.22	0.18	0.40
8	0.28	0.23	0.21
9	0.32	0.19	0.27
10	0.27	0.31	0.26
11	0.24	0.29	0.34
12	0.26	0.26	0.29
13	0.37	0.28	0.29
14	0.21	0.23	0.22
15	0.31	0.29	0.29
16	0.23	0.30	0.26
17	0.28	0.23	0.29
18	0.53	0.12	0.21
19	0.33	0.17	0.10
20	0.23	0.33	0.24
21	0.24	0.28	0.42
22	0.41	0.18	0.39
All[a]	0.29±0.07	0.24±0.05	0.28±0.08

Data are based on 90 hybrids. In each row, any remaining fraction of hybrids could not be unequivocally characterized through single-pass genotyping. [a]Mean ± standard deviation for chromosomes 1–22.

Conversion entails fusion of viable human cells, typically lymphocytes or fibroblasts, with a rodent cell line to create hybrid cells that retain an apparently random subset of human chromosomes. Successful fusion events or hybrids are propagated under conditions that select for fused cells—for example, using the *HPRT1*/HAT (hypoxanthine, aminopterin, thymidine) system (see Methods). After 2–4 weeks of growth, fused clones are harvested, and DNA is prepared for analysis. Typically, 10–100 clones result from a single fusion experiment, with each clone being null, monosomic or disomic for each pair of human chromosomes. Clones that are monosomic for one or more human chromosomes (Fig. 1) can be identified by genotyping a few, highly polymorphic markers per chromosome, which minimally requires a single heterozygous genotype. Subsequent genotyping of these monosomic clones provides haplotypes for the chromosomes of interest.

To assess the feasibility of conversion for whole-genome haplotype construction, we obtained DNA from 24, 19, 23, and 24 somatic cell hybrids derived from peripheral blood lymphocytes of one female and three males respectively. We genotyped DNA from these 90 hybrids and the 4 original donors for 90 markers distributed across the 22 autosomes and the X chromosome. On average, 29%, 24% and 28% of hybrids were null, monosomic and disomic, for a given autosome (Table 1); the remaining 19% of hybrids could not be unequivocally characterized with respect to a given autosome through single-pass genotyping. With the exception of chromosome 18, for which 53% of the 90 hybrids were null, there was no gross evidence for preferential loss or retention of particular chromosomes. Overall, 18±13 (mean ± standard deviation) human autosomes were retained in each hybrid, and 71 of the 90 hybrids retained at least one autosome (mean 7±4) in the monosomic state. Note that genotyping one or two markers per chromosome usually permitted identification of monosomic hybrids, and genotyping four markers was always sufficient.

Table 2 • Subset of chromosome 2 marker data

	D2S168	*D2S2333*	*D2S347*	*D2S125*
Donor	173, 175	101, 103	272, 294	92, 102
Hybrid 1[a]	175	103	272	92
Hybrid 2	173, 175	101, 103	272, 294	92, 102
Hybrid 3	–	–	–	–
Hybrid 4[b]	173	–	294	–
Hybrid 5[a]	175	103	272	92
Hybrid 6[a]	173	101	294	102

[a]Hybrid retains chromosome 2 in the monosomic state. [b]Upon repeat genotyping, hybrid is monosomic for chromosome 2 with haplotype 173-101-294-102.

Based on these data, we were able to infer haplotypes for all 22 autosomes and the X chromosome in all 4 subjects. In 78 of the 88 autosomal cases, we obtained each maternal and paternal chromosome in separate hybrids; in the remaining 10 of 88 autosomal cases, we obtained only one of the parentally derived chromosomes in the hybrid(s), but could still determine haplotypes by comparison with the donor genotypes. In general, we estimate that 23 hybrids are required to isolate a single set of monosomic chromosomes that cover the genome with 100% probability (see Methods); note that this estimate is based on only three subjects. Because the X chromosome contains *HPRT1*, which confers resistance to HAT, the *HPRT1*/HAT selection strategy automatically selects for retention of the X chromosome from the human donor. Thus, X chromosome haplotypes were easily determined for the three male subjects. For the female subject, 13 of 24 clones retained the X chromosome in the monosomic state, which permitted haplotype determination for her as well.

As an illustration of the haplotype construction strategy, consider the small subset of data for chromosome 2 (Table 2). Here, typing hybrids 1 and 6 or 5 and 6 discloses the subject's parentally derived haplotypes as 175-103-272-92 and 173-101-294-102. Alternatively, comparison of donor genotypes with hybrid 1, 5 or 6 also reveals this subject's haplotypes. Because the donor genotypes serve as an internal control, the latter approach may be the preferred strategy for haplotype construction. Note that, in this example, hybrids 2 and 3 are apparently disomic and null for chromosome 2, respectively, and therefore are uninformative for haplotype construction. Hybrid 4 was initially equivocal but upon repeat genotyping was determined to be monosomic for haplotype 173-101-294-102.

Our experience and data indicate that conversion-based haplotyping presents some technical challenges. For example, 20 of the 90 hybrids were equivocal with respect to the retention of chromosome 2 on the basis of single-pass genotyping. Of these 20 equivocal cases, 10 were the result of low DNA concentrations and were later resolved by reconcentrating and genotyping. For each chromosome, genotypes from distal and proximal markers

Table 3 • Relative haplotype-frequency information

Number of SNPs[a]	Mean Fisher information ratio[b] (range)	Mean sample size reduction[c]
2	0.79 (0.55–0.95)	1.3×
3	0.59 (0.39–0.90)	1.7×
4	0.43 (0.15–0.85)	2.3×
5	0.30 (0.08–0.80)	3.3×
6	0.21 (0.04–0.75)	4.8×

Under the assumptions of Hardy-Weinberg and linkage equilibrium, the numbers given represent the means and ranges of the Fisher information ratio for all combinations of allele frequencies from 0.1 to 0.9 (intervals of 0.2) resulting in distinct haplotype frequencies. [a]Number of SNPs forming the haplotype. [b]Ratio of Fisher information when estimating the frequency of a specific haplotype from standard genotypes compared with conversion-derived haplotypes. [c]Mean sample size reduction (based on equivalent information) resulting from the use of haplotype instead of genotype data.

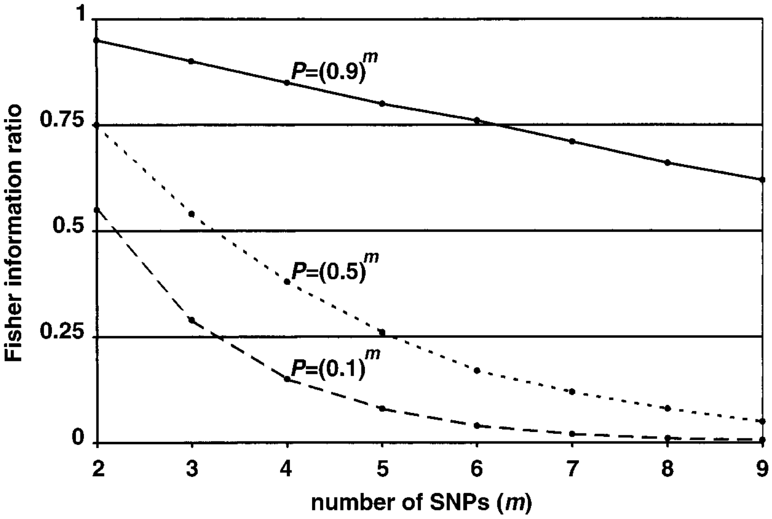

Fig. 2 Select haplotype-frequency information ratios. *p*, haplotype frequency; *m*, number of SNPs forming the haplotype. Results assume linkage equilibrium and the same allele frequency at each locus.

indicate that whole chromosomes rather than chromosomal fragments were generally retained in the hybrids, which is consistent with multicolor fluorescence *in situ* hybridization (FISH) analysis results on several of these hybrids (Fig. 1; B. Vogelstein, pers. comm.). Still, repeat genotyping indicates that 4 of the remaining 10 equivocal cases may be the result of insertions or deletions of chromosomal segments instead of low DNA concentrations or preferential amplification, although the presence of the former could not be verified because we had DNA only from these hybrids. There was also evidence for the absence of X chromosome retention (distal and proximal to *HPRT1*) in 6 of the 90 hybrids, possibly resulting from incomplete selection for *HPRT1* or chromosomal rearrangements.

Even in the presence of the technical challenges above, the integrity of haplotype construction was not compromised for any chromosome-subject combination in our sample. In particular, there was evidence of only one somatic recombination event (out of 483 clones monosomic for a given autosome), an essential requirement for accurate haplotype construction. Moreover, we detected only five apparent genotyping errors or mutations, which corresponds to an error rate of 0.06% of 8,460 genotypes.

To address the utility of conversion for gene-mapping studies, we considered the context of linkage disequilibrium mapping and addressed the question: what reduction in sample size can be realized by using conversion-derived haplotypes instead of standard genotypes? We calculated the Fisher (or average) information matrix for haplotype frequency estimation, assuming haplotypes consisting of two or more SNPs. In large samples, Fisher information is equal to the inverse of the variance of a maximum-likelihood estimator. Consequently, for a single parameter, such as haplotype frequency, information is inversely proportional to the required sample size and therefore provides a practical means for comparing the relative efficiency of two different experimental designs or estimation schemes.

Under the assumptions of Hardy-Weinberg and linkage equilibrium, we present ratios of Fisher information when using unrelated individuals to estimate the frequency of a specific haplotype from standard genotypes compared with conversion-derived haplotypes (Table 3). As a consequence of the relationship between Fisher information and large-sample variance, these numbers provide the relative sample sizes required by haplotype versus genotype data to obtain equivalent information. For example, for a 6-SNP haplotype, standard genotype data provide on average only 21% as much Fisher information per subject as do experimentally-derived haplotypes. As a result, to attain the same level of precision of haplotype-frequency estimation, sample size requirements are on average about 1/0.21, or 4.8 times greater for standard genotypes than for conversion-derived haplotypes. Even for a 3-SNP haplotype, genotypes provide an average of 59% as much information as do haplotypes. These results are consistent with research on two-locus haplotype frequency estimation[14–16]. For example, based on empiric data, Tishkoff *et al.*[14] reported large decreases in the standard errors of two-locus haplotype frequency estimates with haplotype (compared with genotype) data, and, based on the disequilibrium coefficient *D*, Hill[15] and McKeigue[16] demonstrated analytically that two-locus genotype data were 50% efficient compared with haplotype data.

The advantage of experimentally-derived haplotypes increases with increasing numbers of loci (Table 3), reaching a factor of nearly 5 for a 6-SNP haplotype. This advantage also increases with decreasing haplotype frequency. For example, consider the Fisher information ratio for several distinct haplotype frequencies under the assumption of linkage equilibrium and the same allele frequency at each locus (Fig. 2). For a 4-SNP haplotype of frequency ~66% or (0.9^4)—that is, each allele of frequency 90%—genotype data provide at least 75% as much Fisher information as do haplotype data. In contrast, when this same haplotype occurs with frequency ~6% or (0.5^4)—that is, each allele of frequency 50%—relative information is less than 50%. For markers not in linkage equilibrium, gains will be correspondingly less.

Conversion-based haplotyping is more costly than standard genotyping because of the requirement for hybrid construction and characterization and duplicate genotyping. This extra cost, however, must be balanced against the additional information provided by the direct knowledge of haplotypes. The information gain on a per subject basis reduces the required sample size, and consequently the overall recruitment and phenotyping costs, which tend to dominate those for genotyping. Thus, depending on the relative costs of recruiting, phenotyping and genotyping, conversion-based haplotyping may be more cost-effective in the long run than standard genotyping.

To assess cost-effectiveness, we calculated the ratio R that compares information per subject for standard genotype data with that for conversion-derived haplotype data. On the basis of R, if a sample of n subjects is conventionally genotyped, then haplotyping a subset of $m{=}R{\cdot}n$ of these subjects provides the same amount of haplotype-frequency information as does genotyping. If C, P and G represent the per subject costs of conversion and initial hybrid testing, recruitment and phenotyping, and genotyping, respectively, then the total cost for genotyping is $n(P{+}G)$, whereas that for haplotyping is $m(C{+}P{+}2{\cdot}G)$. Under this model, conversion-based haplotyping is more cost-effective than standard genotyping when $m(C{+}P{+}2{\cdot}G){<}n(P{+}G)$ or, equivalently, $R{<}(P{+}G)/(C{+}P{+}2{\cdot}G)$. For example, in the context of fine mapping, if C=\$350 (the current commercial cost for a single chromosome), G=$200 (as it might be to type a dense set of markers in a region of interest), and P=$1,500, $1,000 or $500, then conversion-based haplotyping is more cost-effective than genotyping when $R < 0.75$, 0.69, or 0.56, respectively. These inequalities may be achieved when estimating the frequency of a haplotype consisting of as few as three or four SNPs, where the average information ratio is 0.59 or 0.43, respectively (Table 3).

In the future, conversion-based haplotyping may also be cost-effective in the context of whole-genome association studies, especially as automated methods of genotyping continue to improve and costs continue to decrease, and phenotyping costs remain high. For example, if C=$1,500 (the current estimated cost to construct and characterize 23 hybrids for whole-genome coverage), G=$300 (the cost to genotype 30,000 SNPs at $0.01 per genotype) and P=$1,500, then conversion-based haplotyping is more cost-effective than genotyping when $R<0.50$. Again, this inequality may be realized when estimating the frequency of a haplotype consisting of four SNPs, where the average information ratio is 0.43 (see Table 3). Note that, regardless of haplotype construction strategy, significant reductions in genotyping costs must be achieved for whole-genome association screening to be feasible.

Conversion-based haplotyping has the potential to be very useful in a variety of contexts, including linkage analysis and population genetic studies, as well as linkage disequilibrium mapping. The advantages will probably be greatest for studies in which subjects or families are difficult or impossible to ascertain or are expensive to phenotype. Studies of late-onset diseases and case-control designs in particular may benefit from conversion-based haplotyping because the approach eliminates the need to collect and genotype DNA from family members who may be deceased or simply unavailable. Moreover, for rare diseases or studies that rely on critical individuals, valuable information can be recovered with experimentally-derived haplotypes.

The comparisons and results presented here address efficiency on the basis of haplotype-frequency estimation and therefore are relevant to the extent that haplotypes are useful for linkage disequilibrium mapping. Comparison of the efficiency of haplotype versus genotype data for detecting linkage disequilibrium with a disease-predisposing variant requires independent consideration.

Methods

Somatic cell hybrids. The Vogelstein laboratory generated[3] and generously donated anonymous somatic cell hybrids. GMP Genetics generated additional hybrids for cytogenetic analysis. In brief, immortalized mouse recipient cells (E2) were derived from mouse embryonic fibroblasts that carry a mutation in *HPRT1* and are sensitive to HAT. We obtained human lymphocytes from four individuals as part of an IRB-approved study at Johns Hopkins University. Lymphocytes were electrofused with the E2 recipient cells and selected in medium containing HAT and geneticin to permit the selective growth of somatic cell hybrids only. Colonies appearing after two weeks of growth were then expanded and grown for another two weeks. DNA was harvested from 24, 24, 23 and 19 colonies for genotyping.

DNA analysis. We genotyped hybrids with 90 dinucleotide microsatellite markers distributed across the 22 autosomes and the X chromosome (2 markers per chromosome arm and an additional 8 markers over a 7-megabase region on 1q). We set up PCRs with a TECAN Genesis200 robot. PCR amplification was done with the GeneAmp 9600 (Applied Biosystems). We separated PCR products using the ABI 3700 DNA sequencer, which allows multiple fluorescently labeled markers to be run in a single lane. We ran the ROX 400 size standard as an internal size standard and calculated allele sizing with the local southern algorithm available in the Genescan software program (Applied Biosystems). Allele calling and binning were done with the Genotyper software (Applied Biosystems).

Fisher information calculations. For m=2 SNPs, Fisher information calculations are reasonably straightforward. For phase-unknown (genotype) data, there are nine possible two-locus genotypes, so the data follow a 9-nomial distribution. For phase-known (haplotype) data, the double-heterozygote class is no longer ambiguous, and therefore the data follow a 10-nomial distribution. Under the assumption of Hardy-Weinberg equilibrium, the multinomial class probabilities are simple functions of the four haplotype frequencies or, equivalently, three functionally independent haplotype frequencies. Thus, for two SNPs, Fisher information is a 3×3 matrix, which is easily calculated and inverted. Analytic calculation is substantially more difficult for m>2 SNPs. In this case, there are $2^{m-1}(2^m+1)$ phase-known and 3^m phase-unknown multilocus genotypes, and the information matrices are of dimension 2^{m-1}. For m>2 SNPs, we developed an algorithm that permits numerical computation and inversion of the information matrices. For the phase-unknown case, our algorithm is of order 6^m, which permits nearly instantaneous efficiency comparisons for m<7 and acceptably fast comparisons for m<10.

Estimated hybrid construction requirements. We estimated the number of hybrids required for haploid genome coverage by enumerating all combinations of the observed data for the current set of 90 hybrids (24, 24, 23 and 19 from each of 4 individuals) and determining the fraction for which genome coverage was achieved. Empiric results from this exercise give the estimated number of hybrids needed to isolate a single set of monosomic chromosomes that cover the genome with a desired level of certainty, say >95%. Calculations presuppose genotyping diploid DNA from the original sample and haploid DNA from a single set of monosomic chromosomes covering the genome in order to obtain both maternally and paternally derived haplotypes for each individual.

Acknowledgments

We thank B. Vogelstein and H. Yan of Johns Hopkins University for generously donating the somatic cell hybrids. We would also like to acknowledge the technical assistance of T. Dennis and A. Dutra of the National Human Genome Research Institute Cytogenetic and Confocal Microscopy Core. This work was supported in part by a University of Michigan Predoctoral Fellowship to J.A.D. and by National Institutes of Health grants to M.B. (R01 HG00376) and S.B.G. (R01 CA81488).

Received 22 February; accepted 31 May 2001.

1. Papadopolous, N., Leach, F.S., Kinzler, K.W. & Vogelstein, B. Monoallelic mutation analysis (MAMA) for identifying germline mutations. *Nature Genet.* **11**, 99–102 (1995).
2. Laken, S.J. *et al.* Analysis of masked mutations in familial adenomatous polyposis. *Proc. Natl. Acad. Sci.* USA **96**, 2322–2326 (1999).
3. Yan, H. *et al.* Conversion of diploidy to haploidy. *Nature* **403**, 723–724 (2000).
4. Yan, H., Kinzler, K.W. & Vogelstein, B. Genetic testing—present and future. *Science* **289**, 1890–1892 (2000).
5. MacLean, C.J. & Morton, N.E. Estimation of myriad haplotype frequencies. *Genet. Epidemiol.* **2**, 263–272 (1985).
6. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
7. Hawley, M.E. & Kidd, K.K. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* **86**, 409–411 (1995).
8. Long, J.C., Williams, R.C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810 (1995).
9. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L., Kidd, K.K. & Ruano, G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res.* **24**, 4841–4843 (1996).
10. Hodge, S.E., Boehnke, M. & Spence, M.A. Loss of information due to ambiguous haplotyping of SNPs. *Nature Genet.* **21**, 360–361 (1999).
11. Weiss, M.C. & Green, H. Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes. *Proc. Natl. Acad. Sci.* USA **58**, 1104–1111 (1967).
12. Nishimura, D.Y. *et al.* The forkhead transcription factor gene FKHL7 is responsible for glaucoma phenotypes which map to 6p25. *Nature Genet.* **19**, 140–147 (1998).
13. Glaser, T. *et al.* The beta-subunit of follicle-stimulating hormone is deleted in patients with aniridia and Wilms' tumour, allowing a further definition of the WAGR locus. *Nature* **321**, 882–887 (1986).
14. Tishkoff, S.A., Pakstis, A.J., Ruano, G. & Kidd, K.K. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *Am. J. Hum. Genet.* **67**, 518–522 (2000).
15. Hill, W.G. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–239 (1974).
16. McKeigue, P.M. Efficiency of estimation of haplotype frequencies: use of marker phenotypes of unrelated individuals versus gene counting of phase known gametes. *Am. J. Hum. Genet.* **67**, 1626–1627 (2000).
17. Veldman, T., Vignon, C., Schrock, E., Rowley, J.D. & Ried, T. Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping. *Nature Genet.* **15**, 406–410 (1997).