Hum Genet (2002) 110 : 148–156
DOI 10.1007/s00439-001-0656-4

ORIGINAL INVESTIGATION

**Chun-Fang Xu · Karen Lewis · Kathryn L. Cantone
Parveen Khan · Christine Donnelly · Nicola White
Nikki Crocker · Pete R. Boyd · Dmitri V. Zaykin
Ian J. Purvis**

Effectiveness of computational methods in haplotype prediction

Received: 30 July 2001 / Accepted: 9 November 2001 / Published online: 14 December 2001

Abstract Haplotype analysis has been used for narrowing down the location of disease-susceptibility genes and for investigating many population processes. Computational algorithms have been developed to estimate haplotype frequencies and to predict haplotype phases from genotype data for unrelated individuals. However, the accuracy of such computational methods needs to be evaluated before their applications can be advocated. We have experimentally determined the haplotypes at two loci, the *N-acetyltransferase* 2 gene (*NAT2*, 850 bp, *n*=81) and a 140-kb region on chromosome X (*n*=77), each consisting of five single nucleotide polymorphisms (SNPs). We empirically evaluated and compared the accuracy of the subtraction method, the expectation-maximisation (EM) method, and the PHASE method in haplotype frequency estimation and in haplotype phase prediction. Where there was near complete linkage disequilibrium (LD) between SNPs (the *NAT2* gene), all three methods provided effective and accurate estimates for haplotype frequencies and individual haplotype phases. For a genomic region in which marked LD was not maintained (the chromosome X locus), the computational methods were adequate in estimating overall haplotype frequencies. However, none of the methods was accurate in predicting individual haplotype phases. The EM and the PHASE methods provided better estimates for overall haplotype frequencies than the subtraction method for both genomic regions.

Introduction

Haplotype analysis has been widely employed in linkage studies for narrowing down the location of disease-susceptibility genes and in studies investigating population processes such as the origin and migration of ancestral alleles. It has also become an increasingly popular tool in assessing linkage disequilibrium (LD) and for mapping complex disease genes in association studies where phenotype-marker association may not be detectable as a first order association between single markers and phenotypes (Templeton 1999). Several studies indicate that extended marker haplotypes can provide additional power in detecting associations (Templeton et al. 1988; Kruglyak 1999; Judson et al. 2000; Martin et al. 2000; Zollner and von Haeseler 2000). Conventionally, haplotype phases have been resolved by tracing chromosomal transmission through extended families. Such extensive pedigree data are often not available in association studies where unrelated individuals or small nuclear families are used. Haplotype phases can also be determined by using molecular approaches, such as cloning, allele-specific polymerase chain reaction and single molecule dilution (Ruano and Kidd 1989; Ruano et al. 1990; Michalatos-Beloin et al. 1996; Clark et al. 1998). These molecular methods are labour-intensive and expensive to use in haplotype determination and, therefore, are not suitable for high-throughput applications.

A cheap and relatively straightforward alternative for haplotype estimation is the application of computational algorithms to predict haplotypes by using genotype data (Clark 1990; Excoffier and Slatkin 1995; Long et al. 1995; Hawley and Kidd 1995; Chiano and Clayton 1998; Stephens et al. 2001a). Most of these methods use the expectation-maximisation (EM) algorithm to predict haplotype frequencies in a population for which no assumption is made about LD between markers. Fallin and Schork (2000) have demonstrated high accuracy in haplotype frequency estimation for biallelic diploid samples by using the EM algorithm via extensive simulation studies. They have found that the estimation error is decreased by a number of factors: an increased sample size, a decreased ambiguity (unphased individuals), an increased dispersion of haplotype values, an increased LD between single nucleotide polymorphisms (SNPs) and an increased number of rare

C.-F. Xu (✉) · K. Lewis · K.L. Cantone · P. Khan · C. Donnelly
N. White · N. Crocker · P.R. Boyd · D.V. Zaykin · I.J. Purvis
Discovery Genetics, GlaxoSmithKline Research
and Development, Medicines Research Centre,
Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK
e-mail: cfx74267@gsk.com,
Tel.: +44-1438-768392, Fax: +44-1438-768097

SNPs. Much of the overall haplotype estimation error is attributable to sampling error, which is inherent in all studies even when the phases are experimentally determined. Templeton et al. (1988) first described the application of the EM algorithm in haplotype estimation in a study of association between phenotypic variation and genetic polymorphisms in the *apo AI-CIII-AIV* gene cluster. Haplotype phases, defined by three SNPs over a 10-kb region, were resolved for all individuals. The authors observed enhanced statistical power in detecting genetic effects by the application of haplotype analysis.

The subtraction algorithm described by Clark (1990) works in a stepwise manner. It starts by assigning haplotypes for unambiguous individuals who are either complete homozygotes or single-site heterozygotes. Subsequently, other individuals who carry a copy of the previously recognised haplotypes are identified. Each time a resolved haplotype is identified as one of the possible alleles in an ambiguous individual, the homologous allele is considered to be a recognisable haplotype. This exercise is repeated until the haplotype phases for all individuals are either resolved or identified as unresolved. In combination with AS-PCR, the subtraction algorithm has been used to resolve the haplotype phases of the lipoprotein lipase gene (Clark et al. 1998).

Stephens et al. (2001a) have recently introduced a Bayesian statistical method (called the “PHASE” method) for haplotype reconstruction from population data. PHASE incorporates the prior knowledge that unresolved haplotypes will be similar to known haplotypes. On the basis of their simulations, the authors have shown that PHASE improves on both the EM algorithm and the subtraction algorithm in haplotype reconstruction.

However, it remains to be seen how effective the algorithms are in haplotype prediction when applied to actual data. Tishkoff et al. (2000) have empirically determined the haplotypes at the CD4 locus over a 9.8-kb region consisting of a short tandem repeat polymorphism and an ins/del polymorphism and have assessed the accuracy of the EM algorithm in haplotype frequency estimation. They have shown that EM-algorithm-estimated frequencies of common haplotypes do not differ significantly with that empirically determined, whereas rare haplotypes are occasionally miscalled. Recently, Zhang et al. (2001) compared PHASE and EM by using both simulated data and phase-known data derived from a subset of the CD4 genotype data of Tishkoff et al. (2000). They found that the performances of PHASE and EM were similar in both haplotype construction and haplotype phase prediction. Stephens et al. (2001b) compared EM and PHASE for haplotypes determined from pedigree data at three loci, each being approximately 4–5 kb. They argued that PHASE should outperform EM when there was “clustering” in the true haplotype configuration and showed PHASE either outperformed EM or did about the same. However, none of these studies empirically evaluated the accuracy of computational methods with real data when such “clustering” of haplotypes is not present. The performance of these computational methods is likely to be influenced by LD between markers. None of the previous studies have presented LD data (Tishkoff et al. 2000; Stephens et al. 2001a, 2001b; Zhang et al. 2001). There probably was reasonable LD between markers in these studies as the regions studied were relatively small (ranged between 4 kb and 10 kb). It remains to be seen how these algorithms work when applied to relatively large genetic loci. This is important as most of the studies involving haplotype construction, such as studies in population genetics and in mapping disease genes, include genomic regions of hundreds of kilobases or more. In addition, none of the previous studies has empirically evaluated the performance of the subtraction method over the EM and PHASE method. In this study, we have used experimental data from two loci, one with pronounced LD over an 850-bp region, the other with less LD over a 140-kb region. We have comprehensively evaluated and compared the accuracy of the EM algorithm, the subtraction algorithm, and the PHASE method in both haplotype frequency estimation and in individual haplotype prediction based on experimental data.

Materials and methods

DNA samples

Blood was collected from 81 GlaxoSmithKline employees (Caucasians from North Carolina, USA) and 154 males (Caucasians from San Francisco, USA) under informed consent. DNA was extracted by PPGX (Research Triangle Park, North Carolina) by using the Puregene DNA isolation kit (Gentra Systems, supplied by Flowgen Instruments, Lichfield, UK) or by Whatman Bioscience (Cambridge, UK). All DNA samples were anonymised.

Genotyping

Five SNPs from the *N-acetyltransferase* 2 (*NAT2)* gene on chromosome 8 (Fig. 1; Agundez et al. 1996) were genotyped for each of the 81 GlaxoSmithKline employees by PCR and direct sequencing. An 850-bp fragment of the *NAT2* gene was amplified by using primers F1 and R1 and subsequently sequenced with the initial PCR primers and two additional nested primers (F2 and R2) on an ABI 377 Sequencer (PE Applied Biosystems, Foster City, USA). The sequences of the primers were as follows: F1 (forward PCR primer): 5’-CTATAAGAACTCTAGGAACAAATTGGAC-3’;

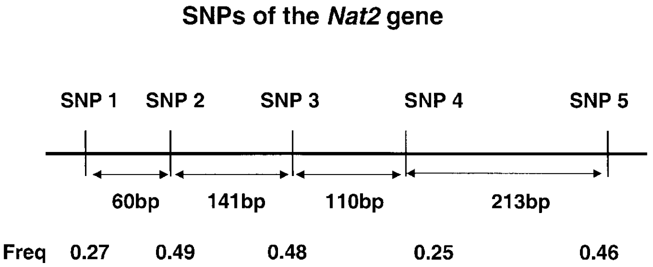

Fig. 1 Schematic representation of the five SNPs of the *NAT2* gene. SNP1(C/T), SNP2(T/C), SNP3 (C/T), SNP4 (G/A) and SNP5(A/G) correspond to nucleotide positions 282, 341, 481, 590 and 803 (U53473), respectively. Frequency (*Freq*) refers to the minor allele frequencies. The distances between neighbouring SNPs are shown

Table 1 OLA primer and probe sequences for the five SNPs on the X chromosome

Primer or probe	Sequences
SNP6 (T/G)	
F PCR primer	TTTTGGTGTTGCAGTATTGACAG
R PCR primer	TCTTGGGAAGCATAGGTCTCTTG
Allele 1 probe	GCTGTCAGAACAGGAATT (FAM)
Allele 2 probe	ACAGCTGTCAGAACAGGAATG (FAM)
Common probe	TCCAAACTGCTCTAGCTGAAGACAG
SNP7 (G/T)	
F PCR primer	CCACAAATCTTTGCTGTGATGAG
R PCR primer	ACCCCATGCTAGACATGCTATTC
Allele 1 probe	AATGAGTGGTCCGGGAAG (HEX)
Allele 2 probe	CAAAATGAGTGGTCCGGGAAT (HEX)
Common probe	CCCTTGCTATAGACGGGAGAATGCTA-CAGTCTC
SNP8 (A/G)	
F PCR primer	GAGCTGGAAAGCACCAGAACATG
R PCR primer	GAGGCGATCTCCAGCCTCC
Allele 1 probe	TCCTTTTCCCAAACCAGA (FAM)
Allele 2 probe	TCATCCTTTTCCCAAACCAGG (FAM)
Common probe	GCTCTATATGTTCAAGGAAATGCAGC-GGTATGTGTGCCT
SNP9 (C/G)	
F PCR primer	AGACACGAAGGAGTGCATTCTG
R PCR primer	TCTAGCCCAAACCTCTTTTGAAG
Allele 1 probe	TTACAAAGTCAACTCACC (HEX)
Allele 2 probe	TTTTTACAAAGTCAACTCACG (HEX)
Common probe	CGTTAGCCACCACTGAGATCAAGAGC
SNP10 (T/C)	
F PCR primer	CCACATAGATGCTTCCAGCAGC
R PCR primer	GTTCAGTTTTGCCTGACGATC
Allele 1 probe	AATGCTACAGAGAAGCTT (FAM)
Allele 2 probe	AAGAATGCTACAGAGAAGCTC (FAM)
Common probe	AAGTAGTGAACATAGTGGGGAGCTT-GAGTCAC

R1 (reverse PCR primer): 5’-AAGGGTTTATTTTGTTCCTTAT-TCTAAAT-3’; F2 (nested forward primer): 5’-CACCTTCTCCT-GCAGGTGACCA-3’; R2 (nested reverse primer): 5’-TGTCAAG-CAGAAAATGCAAGGC-3’. Sequencher (Genecodes, Ann Arbor, USA) was used to analyse the sequences in order to generate genotype results for each of the five polymorphic sites.

Five SNPs over a region of 140-kb on the X chromosome were identified by PCR and direct sequencing of DNA samples from 11 female individuals (Coriell Cell Repositories, New Jersey, USA). Oligo ligation assays (OLA; Landegren et al. 1988; Grossman et al. 1994) were used to generate genotype calls for the 154 males. Table 1 shows the sequences of the primers and probes used in the PCR and OLA assays. The data was analysed by using Genotyper NT (PE Applied Biosystems) to generate genotype calls.

Molecular determination of the haplotypes

The 850-bp PCR fragment of the *NAT2* gene was cloned into a TA cloning vector (Invitrogen, Groningen, The Netherlands). Between six and twelve subclones from each of the 81 individuals were sequenced. The sequence data were analysed by using Lasergene (DNASTAR, Madison, USA) to resolve the haplotypes for both chromosomes of each individual. The haplotypes from the five SNPs on chromosome X were assigned directly according to the genotype data, as each individual male has only one X chromosome.

Computational estimation of the haplotypes

For the chromosome X region, artificial diploid genotypes were constructed by combining random pairs of males. The haplotypes for each diploid for both genetic regions were assigned by using the subtraction algorithm (Clark 1990). Briefly, haplotypes for individuals who were either complete homozygotes or single-site heterozygotes were assigned initially and a preliminary list of haplotypes present in the samples was recorded. Other individuals who carried a copy of the previously recognised haplotypes were then identified. Each time a resolved haplotype was identified as one of the possible alleles in an ambiguous individual, the homologous allele was considered to be a recognisable haplotype and added to the haplotype list. This exercise was repeated until the phase information for all individuals was either resolved or identified as unresolved. Depending on the order in which the genotypes are entered, the algorithm may produce a different set of haplotypes. The haplotype frequencies were calculated by gene counting in individuals with resolved haplotype phases.

The sample haplotype frequencies and individual conditional haplotype probabilities for both genomic regions were also estimated by using the EM algorithm with multiple restarts (computer program is available from D. Zaykin upon request). All haplotype pairs that can yield an unphased genotype pattern were enumerated. The probability for each of the haplotype configurations was calculated by using the estimated population haplotype frequencies. For example, suppose one haplotype pair that generates the unphased pattern is i/j, where i and j represent two of the haplotypes with *p(i)* and *p(j)* frequencies as estimated by the EM algorithm. From Bayes’ rule, the conditional probability that the unphased genotype G_{ij} has the haplotype pair i/j is

$$\Pr(p(i),\, p(j)\,\middle|\,G_{if}) = \frac{p(i)\,p(j)}{\Sigma_{x,y}\,p(x)\,p(y)}$$

where *x* and *y* indicate a haplotype pair that can yield the same unphased genotype and the sum is taken over all such pairs including *i* and *j*. The haplotype pair with the greatest probability was considered to be the haplotype phase for each diploid. We also evaluated the accuracy of EM in haplotype construction by increasing the probability threshold to 99%, i.e. the haplotype phase was considered to be resolved only if the probability of a haplotype pair was greater than 99%. In this case, a diploid was classified as unphased if neiher of the haplotype pair had a probability greater than 99%.

The default parameter values (10,000 iterations, a thinning interval of 100, and a burn-in value of 10,000) specified in PHASE were used to evaluate the performance of this method (Stephens et al. 2001a). Haplotype phase was specified by the most probable haplotype pair that is compatible with the individual multi-site genotypes. Similarly, we examined the accuracy of PHASE in haplotype construction by increasing the threshold probability to 99%, i.e. the haplotype phase was considered to be resolved only if the probability of each haplotype call at ambiguous positions was greater than 99%.

Measures of estimation accuracy

The two measures, I_F and I_H, introduced by Excoffier and Slatkin (1995) were used to estimate the effectiveness of computational algorithms when predicting haplotype frequencies. I_F (the similarity index) describes how close the estimated haplotype frequencies are to the actual frequencies and is defined as the proportion of haplotype frequency in common between estimated and true frequencies:

$$I_F = \sum_{k=l}^{h} \min(p_{ek},\, p_{tk}) = 1 - \frac{1}{2}\sum_{k=l}^{h} |\,p_{ek} - p_{tk}|$$

where h is the number of haplotypes in the data set, p_{ek} and p_{tk} are the estimated and true (experimentally determined in this case) haplotype frequencies for the k haplotype, k=1....h. I_F varies between 0 and 1 (a value of 1 is achieved when the actual and estimated frequencies are identical). I_H compares the number of different haplotypes seen experimentally with the number of different haplotypes identified by the computer programs. A haplotype is defined as being detected if it has an estimated frequency of at least 1/(2n) in a population of n individuals (Excoffier and Slatkin 1995). I_H is given by:

$$I_H = \frac{2(m_{true} - m_{missed})}{m_{true} + m_{est}}$$

where m_{true} is the number of haplotypes determined experimentally, m_{est} is the number of estimated haplotypes with frequency above the threshold, and m_{missed} is the number of haplotypes identified experimentally but not computationally. The value of I_H can vary between 1 (when the computational identified haplotypes are exactly the same as those determined experimentally) to 0 (when none of the true haplotypes are identified computationally).

The mean squared error (MSE) described by Fallin and Schork (2000) was also used to measure the accuracy of computational algorithms in haplotype frequency estimation. The MSE measure incorporates all the k haplotype frequencies and thus reflects the overall difference in haplotype frequencies between estimated and true values for a particular data set:

$$MSE = \sum_{k=l}^{h} (p_{ek} - p_{tk})^2 / h$$

where h, p_{ek} and p_{tk} are defined as above.

Pair-wise LD between SNP markers

LD was measured by using the standardised D' first proposed by Lewontin (1964). D' is the LD relative to its maximum value for a given set of allelic frequencies for the pair of sites. It is calculated by dividing the raw D value by the absolute maximal value possible. In this sense, D' is a normalised value of LD.

Results

Molecular determination of genotypes and haplotypes for the *NAT2* locus

Figure 1 shows the distribution of the five SNPs utilised in this study over the *NAT2* locus. The minor allele frequencies of the five polymorphisms determined in the 81 individuals ranged from 0.25 to 0.49, which were similar to those reported for Caucasians (Agundez et al. 1996). The genotype distribution for each SNP did not deviate significantly from the Hardy-Weinberg equilibrium.

To determine the haplotypes molecularly, the 850-bp PCR fragment of *NAT2* was cloned into a TA cloning vector and between six and twelve subclones from each individual were analysed by PCR and sequencing. In the absence of recombination, recurrent mutation and back mutation, the maximum number of haplotypes for a locus with five biallelic variable sites is 6 (i.e. n+1), with n being the number of SNP sites. On the other hand, if there is random association between polymorphic sites, the maximum number of potential haplotypes for a locus with five SNPs is 32 (2^5). Analysis of the 162 alleles in the *NAT2* locus revealed seven haplotypes suggesting strong LD in this small chromosomal region (Table 2). Indeed, there were maximal or nearly maximal D' values among all SNP pairs indicating that there was complete or near complete linkage disequilibrium and that recombination was rare over such a short physical distance in the *NAT2* gene locus (Table 3). The five SNPs generated ten SNP pairs (Table 3). Each of the eight SNP pairs created only three haplotypes. The remaining two SNP pairs created all four possible haplotypes with three haplotypes accounting for 98%–99% of the alleles.

Table 2 Haplotype frequencies determined by molecular and computational methods for the *NAT2* locus (n=81 individuals)

Haplotype[a]	Experimentally determined	Subtraction[b]	EM[c]	PHASE[d]
12212 (H1)	0.444	0.430	0.444	0.444
11111 (H2)	0.235	0.164	0.234	0.235
21121 (H3)	0.247	0.313	0.247	0.247
21111 (H4)	0.025	0.031	0.025	0.025
12211 (H5)	0.031	0.039	0.031	0.031
12112 (H6)	0.012	0.016	0.012	0.012
11112 (H7)	0.006	0.008	0.005	0.006
I_F		0.914	0.999	1
I_H		1	0.923	1
MSE		1.4E-03	2.9E-07	9.3E-08

[a] Allele 1 refers to the major allele and allele 2 refers to the minor allele for all of the five SNPs in this locus

[b] The haplotype phases were resolved for 64/81 individuals according to the subtraction method (Clark 1990). The haplotype frequencies were calculated from the 64 phase-resolved individuals. The remaining 17 phase unresolved individuals were triple heterozygotes with the genotype distribution being 11,12,12,11,12

[c] EM algorithm with 100 restarts

[d] PHASE method (Stephens et al. 2001a)

Table 3 Linkage disequilibrium (absolute D' value) between SNP markers in the *NAT2* locus and chromosome X locus

Locus and markers	Markers			
Nat 2 locus	SNP2	SNP3	SNP4	SNP5
SNP1	1	1	1	1
SNP2		1	1	0.97
SNP3			1	0.92
SNP4				1
Chr X locus	SNP7	SNP8	SNP9	SNP10
SNP6	0.27	0.40	0.59	0.17
SNP7		0.23	0.1	0.40
SNP8			0.23	0.21
SNP9				0.71

Computational estimation of haplotypes for the *NAT2* locus

We inferred haplotypes from the genotyping results for the 81 individuals by using the subtraction algorithm (Clark 1990). Thirty-one individuals were either homozygous for all SNP sites or heterozygous at only one SNP

Table 4 Comparison of computational methods in predicting haplotype phases

Method	*NAT2* (*n*=81)		Chromosome X (*n*=77)	
	Phase-resolved individuals	Accuracy[a]	Phase-resolved diploids	Accuracy[a]
Subtraction	64	100%	43	95%
EM	81	100%	77	78%
Phase	81	100%	77	77%

[a]We calculated overall accuracy from phase-resolved individuals

site; thus, their haplotypes could be assigned directly. Eight, eighteen, three and twenty-one individuals were heterozygous at two, three, four and all five SNP sites, respectively. Using the subtraction method, we resolved the haplotype phases for 64 individuals (79%, Table 4). There was 100% concordance between experimentally determined haplotype phases and those predicted computationally. The remaining 17 individuals were heterozygous at the same three SNP sites and each had two possible haplotype configurations. The haplotype frequencies were calculated from the 64 phase-resolved individuals (Table 2). The similarity index (I_F) value was 0.91, which was close to its maximal value, suggesting that the subtraction method was effective in estimating haplotype frequencies for this region. The overall estimation error (MSE) was 1–2 orders of magnitude greater than that reported by Fallin and Schork (2000) using the EM algorithm, probably because the subtraction method used a reduced number of individuals in the haplotype frequency estimation (Table 2).

We estimated the haplotype frequencies by using the EM algorithm with 100 restarts to minimize chances of local convergence. A comparison of the haplotype frequencies determined molecularly with those that were estimated showed very high concordance (Table 2). The I_F value was 0.999 and the MSE value was four orders of magnitude smaller than that obtained by using the subtraction method. The estimated haplotypes were in 100% agreement with that experimentally determined for all of the 81 individuals, indicating pronounced accuracy for haplotype assignment (Table 4).

There was no difference between PHASE estimated haplotype frequencies and those experimentally determined. The I_F value reached the maximal value of 1 and the MSE value was the smallest among the three methods (Table 2). PHASE-constructed haplotypes showed 100% agreement with those experimentally determined indicating that PHASE was effective and accurate in haplotype construction for this region.

The EM method and PHASE method outperformed the subtraction method both in estimating haplotype frequencies and in predicting individual haplotype phases for the *NAT2* region, where there was pronounced LD. There were no significant differences in the effectiveness and accuracy between PHASE and EM, though PHASE performed marginally better than the EM method in haplotype frequency prediction.

Molecular determination of genotypes and haplotypes for the X chromosome locus

Figure 2 shows the distribution of the five SNPs over a region of 140 kb on chromosome X. The minor allele frequencies of the five SNPs ranged from 0.06 to 0.40. Table 3 presents the pair-wise linkage disequilibrium between markers.

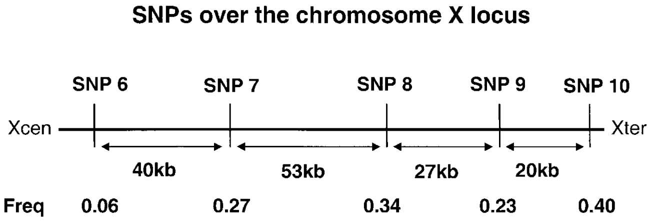

Fig. 2 Schematic representation of the five SNPs spanning 140 kb on chromosome X. The distances between neighbouring SNPs are shown. Frequency (*Freq*) refers to the minor allele frequencies of the five SNPs on chromosome X

Table 5 Haplotype frequencies determined by molecular and computational methods for the chromosome X region

Haplotype[a]	Experimentally determined	Subtraction[b]	EM[c]	PHASE[d]
12212 (h1)	0.169	0.186	0.176	0.149
12211 (h2)	0.162	0.221	0.172	0.156
11211 (h3)	0.123	0.128	0.122	0.136
12112 (h4)	0.117	0.081	0.095	0.136
12221 (h5)	0.097	0.081	0.117	0.136
12111 (h6)	0.084	0.116	0.080	0.071
11221 (h7)	0.045	0.035	0.039	0.039
11112 (h8)	0.039	0.047	0.053	0.039
12121 (h9)	0.032	0.012	0.033	0.019
11121 (h10)	0.019	0.023	0.023	0.019
22112 (h11)	0.019	0.035	0.045	0.045
11212 (h12)	0.013	0.012	0.010	0.013
11222 (h13)	0.013	0.000	0.007	0.000
12222 (h14)	0.013	0.012	0.009	0.013
22212 (h15)	0.013	0.000	2.76E-6	0.000
11111 (h16)	0.006	0.000	4.23E-10	0.006
21111 (h17)	0.006	0.012	0.007	0.000
21211 (h18)	0.006	0.000	0.011	0.019
22111 (h19)	0.006	0.000	4.14E-6	0.000
22121 (h20)	0.006	0.000	7.19E-10	0.000
22211 (h21)	0.006	0.000	1.75E-15	0.000
I_F		0.856	0.914	0.89
I_H		0.800	0.865	0.83
MSE		3.68E-04	1.13E-04	2.02E-04

[a]For SNP6, SNP9 and SNP10, allele 1 refers to the major allele and allele 2 refers to the minor allele. For SNP7 and SNP8, allele 1 refers to the minor allele and allele 2 refers to the major allele
[b]According to the subtraction method (Clark 1990). The haplotype frequencies in this column were calculated from 43/77 diploids. The phases of the remaining 34 diploids (19 double heterozygotes, 12 triple heterozygotes, 3 quadruple heterozygotes) remained ambiguous, i.e. there was more than one possible haplotype configuration
[c]EM algorithm with 100 restarts
[d]PHASE method (Stephens et al. 2001a)

The haplotypes for the 154 males were assigned directly according to the genotype data, as each individual male has only one X chromosome. The five polymorphisms established 21 out of the 32 (2^5) potential haplotypes (Table 5). Six of the haplotypes (h16–h21) were observed only once and four haplotypes (h12–h15) were seen only twice. These 10 rare haplotypes (h12–21) represented 9% of all the 154 alleles. Six haplotypes (h1–h6) had allele frequencies above 5%, representing 75% of the 154 alleles.

Computational prediction of haplotypes
for the chromosome X locus

To evaluate the effectiveness and accuracy of the computational methods in predicting haplotype phases and estimating haplotype frequencies over a relatively large genetic region (140 kb), we artificially created genotype data for 77 diploids by combining random pairs of males. According to the subtraction algorithm (Clark 1990), we resolved the haplotype phases for a total of 43 diploids (56%), including 38 diploids that were heterozygous at 0 or 1 site and five diploids that were heterozygous at multiple sites. There was more than one possible haplotype configuration for each of the remaining 34 diploids (44%) and the haplotype phases of these diploids remained unresolved. For diploids that were either complete homozygotes or single-site heterozygotes, there was a 100% match between the estimated and real haplotypes. However, the computationally assigned haplotype configurations were in agreement with those experimentally determined for only three out of the five multi-site heterozygotes (Table 4). Our data suggested that the subtraction method was neither effective nor accurate in predicting haplotype phases for diploids that were heterozygous at multiple SNP sites in genomic regions where pronounced LD was not maintained.

The haplotype frequencies were estimated by using the 43 “phase-resolved” diploids by the subtraction method (Table 5). The combined haplotype frequency for the six common haplotypes (h1–h6) that each had true allele frequencies greater than 5% was 81%, which was higher than that determined molecularly (75%). The I_F value of the subtraction method was lower for this region than that for the *NAT2* region (Table 2, 5).

We estimated the haplotype frequencies for the 77 artificially generated diploids by using the EM algorithm with 100 restarts. For the 10 haplotypes (h12–21) observed only once or twice molecularly, 0 to 1.5 alleles were predicted computationally, accounting for 4% of all the alleles, which was lower than that observed empirically (9%). The estimated (76%) and true (75%) combined frequencies were similar for the common haplotypes (h1–h6). The reduced I_F value and increased MSE value for this locus in comparison with that observed for the *NAT 2* locus suggested that the estimation error for overall haplotype frequencies was increased with decreasing LD when using the EM algorithm (Tables 2, 5).

As expected, EM predicted haplotypes for the 38 diploids that were either complete homozygotes or single-site heterozygotes were in agreement with those experimentally determined. For the 39 diploids that were multi-site heterozygotes, the haplotype phases assigned by the EM algorithm were in agreement with those experimentally determined for 22 diploids (56%). The overall accuracy for predicting haplotype phases was 78% for all of the diploids (Table 4). One haplotype pair for each of 11 multi-site heterozygotes had the conditional probability of being greater than 99%; the predicted haplotypes were in agreement with the true haplotypes for only five of these diploids (45%). Thus, the accuracy was not improved by increasing the probability threshold.

PHASE slightly over-estimated the combined haplotype frequencies (78%) for the six common haplotypes (h1–h6), whereas it under-estimated the combined frequency (5%) for the 10 rare haplotypes (h12–h21; Table 5). The overall estimation error for the chromsome X locus was greater than that for the *NAT 2* locus (Tables 2, 5), indicating decreased accuracy with decreased LD between markers. PHASE performed marginally better than the subtraction method and marginally worse that the EM method in estimating haplotype frequencies for this locus.

PHASE accurately assigned the haplotypes for 21 of the 39 (54%) multi-site heterozygotes, in addition to the 38 diploids that were heterozygous at 0 or 1 site. The overall accuracy in haplotype construction with PHASE was 77% (Table 4). The probability of each haplotype call at an unknown position being greater than 99% was found for only one multi-site heterozygote, the predicted haplotypes for which was in disagreement with the true haplotypes. For this locus, there was no considerable difference in the effectiveness and accuracy in haplotype construction between the PHASE method and the EM method.

To make a more comprehensive comparison between the three computational methods in haplotype frequency estimation, we performed simple computer simulation to assess the accuracy of these algorithms when there was uniformity of haplotype frequencies. The amount of heterozygosity, and therefore the number of ambiguous haplotypes, was increased by equalising haplotype frequencies, thereby presenting more of a challenge to the computational algorithms. We took the empirical pool of haplotypes in Table 5 and assumed equal haplotype frequencies for all haplotypes (1/21). Random samples of 77 individuals were taken from populations with these frequencies, assuming random union of haplotypes. In comparison with the results presented in Table 5, there was a decrease in the I_F value for the simulated data set, indicating that there was increased estimation error for haplotype frequencies by using the EM and PHASE algorithms when the haplotype frequencies reached uniformity (Table 6). The subtraction method allowed the haplotype phases to be resolved for only 21/77 (27%) diploids and the haplotype frequencies were calculated by using these phase-resolved diploids (Table 6). This exercise suggested that there was increased estimation error for haplotype frequencies with increased ambiguity by using all three computational methods.

Table 6 Haplotype frequency estimation by using computational algorithms for the chromosome X region assuming uniform population haplotype frequencies from simulated data (*n*=77)

Haplotype	Sample frequency	Subtraction estimation[a]	EM estimation[b]	PHASE estimation[c]
11111	0.039	0.024	0.012	0.033
11112	0.045	0.024	0.042	0.006
11121	0.039	0.024	0.016	0.033
11211	0.097	0.071	0.120	0.013
11212	0.045	0.071	0.036	0.006
11221	0.026	0.024	0.030	0.013
11222	0.039	0.048	0.040	0.006
12111	0.026	0.024	0.009	0.020
12112	0.058	0.095	0.078	0.020
12121	0.084	0.071	0.116	0.059
12211	0.026	0.071	0.063	0.045
12212	0.045	0.071	0.038	0.013
12221	0.058	0.048	0.036	0.032
12222	0.032	0.024	0.015	0.019
21111	0.032	0.000	0.092	0.065
21211	0.045	0.000	0.0001	0.019
22111	0.064	0.048	0.021	0.051
22112	0.064	0.095	0.049	0.012
22121	0.039	0.048	0.048	0.006
22211	0.039	0.048	0.041	0.013
22212	0.058	0.071	0.072	0.026
I_F		0.795	0.785	0.74
I_H		0.950	0.976	1
MSE		5.40E-04	6.56E-04	8.91E-04

[a] The haplotypes were assigned according the subtraction method (Clark 1990). The haplotype frequencies in this column were calculated from 21/77 phase-resolved diploids
[b] EM algorithm with 100 restarts
[c] PHASE method (Stephens et al. 2001a)

Table 7 Comparison of the run-times for the three computational methods on a SUN UltraSparc-II

Locus	Subtraction	EM	PHASE
NAT 2 (Table 2)	0 min 0.01 s	0 min 13.48 s	128 min 51.51 s
Chr X (Table 5)	0 min 0.01 s	0 min 11.78 s	42 min 29.79 s
Chr X (Table 6)	0 min 0.03 s	0 min 40.21 s	61 min 1.66 s

We also compared the run-time for each of the three computational methods in predicting haplotypes from diploid data (Table 7). The subtraction method had the least computational burden; the calculation was completed in a fraction of a second for data presented in Tables 2, 5, and 6. PHASE took the longest time to complete the computational process among all three methods (Table 7). It is anticipated that the difference in computational burden between the three methods would even more pronounced if an increase in the number of markers is included in the analysis.

Discussion

We evaluated the effectiveness and accuracy of three computational algorithms in estimating haplotype frequencies and in predicting haplotype phases by using molecular data. We found that all three methods performed well in overall haplotype frequency estimation for genetic regions with high LD (Table 2). The accuracy of computational methods was decreased with decreasing LD and increasing ambiguity. The EM and PHASE algorithms gave better overall estimates of haplotype frequencies than the subtraction method for both genomic regions. This may reflect that the EM and PHASE algorithms include all individuals from the samples in the haplotype frequency estimation, whereas the subtraction method only uses the phase-resolved individuals. PHASE out-performed the EM algorithm for the *NAT2* region, whereas the opposite was true for the chromosome X region. All three algorithms gave better estimates of haplotype frequency for the genomic region with pronounced LD (the *NAT2* locus) than that for the region where substantial LD was not maintained (chromosome X). This observation is in agreement with that reported recently in a simulation study to assess the accuracy of the EM algorithm in haplotype frequency estimation (Fallin and Schork 2000). Fallin and Schork (2000) demonstrated that the EM algorithm performed very well under a wide range of population and data set scenarios. We have shown that haplotype frequencies can be estimated from genotype data computationally without additional laboratory cost and that the estimation error increases with decreasing LD.

We also evaluated and compared the effectiveness and accuracy of the computational algorithms in predicting haplotype phases for individuals. All algorithms predicted individual haplotypes effectively and accurately for the *NAT2* region, where there was near complete LD between SNP sites. The EM and PHASE algorithms gave better estimates than the subtraction method. Such effectiveness and accuracy in haplotype prediction were reduced when marked LD was not maintained, as demonstrated at the chromosome X locus. The subtraction method resolved haplotype phases for only 56% of the diploids with the overall accuracy being 95%. The EM and PHASE methods assigned the haplotype pair with the highest probability to each of the diploids, with the overall accuracy being 78% and 77%, respectively. Increasing the probability threshold of the haplotype phase being resolved to 99% resulted in a decrease in the number of phase-resolved diploids with no improvement in the accuracy of both methods. In contrast to the observation by Stephens et al. (2001a, 2001b), our comparisons showed that there were no significant differences between EM and PHASE in haplotype construction for both genomic regions. Each of our two datasets involved five SNPs; the relative performances of these methods with much larger number of SNPs remained to be seen. Our results indicated that the computational algorithms could provide an effective and accurate prediction for haplotype phases in genetic regions with pronounced LD but not in regions where marked LD is not maintained. It has to be stressed that the degree of inherent phase ambiguity for multiple-site heterozygotes is increased with decreased LD between markers (Hoh and Hodge 2000). The performance of computational algorithms in predicting haplotype phases should be interpreted in the light of this inherent ambiguity.

Our observations may have potential implications for genome-wide association studies. Although experimental evidence supporting the proposal for genome-wide association studies is emerging (Cambien et al. 1999; Cargill et al. 1999; Martin et al. 2000; Moffatt et al. 2000), sequential examination of individual SNPs in an attempt to identify disease-susceptibility genes is fraught with problems of interpretation**.** First, the number of analyses performed will be enormous for the 100,000–500,000 SNPs that will be used in such studies, making it necessary to correct the statistical result and to ensure the authenticity of the signal from each SNP. Second, in contrast with monogenic diseases where the causative single nucleotide changes may have unambiguous phenotypes, the contributions of genetic variations in the underlying network of interactions that are responsible for the phenotypes of complex diseases are much more complicated. The pattern of genotype-phenotype association might be more complex than initially envisaged. Indeed, the very fact that a large number of SNPs have been identified within coding and regulatory regions of a specific candidate gene raises the possibility that several of them within a gene might be functional (Keavney et al. 1998; Nickerson et al. 1998; Cambien et al. 1999; Cargill et al. 1999; Moffatt et al. 2000). Third, our current knowledge about the actual distribution of LD across the human genome is limited (Goddard et al. 2000; Kidd et al. 2000). If complete or near complete association between several polymorphisms within a gene is present, haplotypic combinations of the polymorphic sites may play a significant role in the functionality of the gene. Therefore, it is necessary to explore multiple alternative analytical approaches to identify disease genes in association studies. The incorporation of haplotype analyses will ensure additional use of valuable information in association studies and provide additional evidence about the strength and nature of the associations. Once a collection of SNPs has been discovered and genotyped over a gene locus, a chromosomal region or even the entire genome, they can be organised sequentially into haplotypes. This should allow sequential haplotype scans for two, three, four, five or more SNPs on fragments of DNA ranging from 10 kb to 150 kb in association studies. Sequential haplotype scanning may be able to provide a richly detailed view of specific genomic fragments and reveal the inter-relationships between SNPs surrounding the regions, thus offering an additional method for identifying genomic fragments that harbour the variants causing the phenotype. If the haplotype information is derived from genotype data by using computational methods, it needs to be noted that the accuracy of such haplotype information is decreased with decreasingd LD and increasingd ambiguity between markers.

Our observations also have potential implications for linkage studies involving SNPs. Currently, approximately 400 microsatellite markers are used for genome-wide linkage scans to localise regions harbouring disease genes, with an average genetic distance being approximately 10 cM. Because of the remarkable advances in the technologies for SNP identification and genotyping, it is proposed that it may be more efficient to use 1500–2000 SNP markers to replace microsatellite markers in a typical genome scan. The combined polymorphic information content from several highly informative SNPs within a region may be equivalent to the polymorphic information content from a single multi-allelic microsatellite marker. For a locus with n biallelic variable sites, the maximum number of haplotypes is n +1 in the absence of recombination, repeated or back mutations, whereas the potential number of haplotypes could reach 2^n if there is linkage equilibrium between polymorphic sites. We observed 7 and 21 haplotypes for the five SNPs over the *NAT2* locus and the X chromosome locus, respectively, in the populations studied, reflecting the different magnitudes of LD operating over the 850-bp and the 140-kb regions. The haplotype heterozygosity (H) in the X chromosome region (H=0.82, based on the artificial diploid data) is higher than that at the *NAT2* locus (0.69), although the opposite is true for average H of individual SNPs over the two regions (0.31 and 0.48, respectively). This is consistent with the strong negative correlation between the mean pair-wise LD and haplotype heterozygosity, supporting the concept that the stronger the non-random association between SNPs, the lower the information added by each SNP to a set of other SNPs (Cambien et al. 1999). On the other hand, some information will be lost because of the ambiguous haplotyping of multiple SNPs, if linkage equilibrium is reached between SNP sites (Hodge et al. 1999; Hoh and Hodge 2000). A fine balance of LD between multiple SNP sites is therefore required to obtain the maximum information within a genomic region and the knowledge of the extent and magnitude of LD between SNPs should be invaluable in the selection of the SNP set for linkage analysis.

This study indicates that computational methods can provide an effective prediction of haplotype frequencies by using genotype data from unrelated individuals for genomic regions with LD. Computational algorithms could give effective and accurate prediction for haplotype phases in regions with high values of LD between markers and a small probability of recombination events. The EM and PHASE algorithms are better computational methods than the subtraction method, both in estimating haplotype frequencies and in predicting haplotype phases. Our observation may shed some light on alternative statistical approaches in association studies and linkage studies with SNPs.

Acknowledgements Our thanks go to the Core-Sequencing Group in UK Discovery Genetics, GlaxoSmithKline, for their effort and assistance. We should also like to thank S. Lumsden for the PCR assay development and the Physical Mapping Group for the identification of the SNPs over the chromosome X locus. We are grateful to Drs. Linda McCarthy, Michael Stubbins and John Riley for the discussions and critical readings of the manuscript.

References

Agundez JA, Olivera M, Martinez C, Ladero JM, Benitez J (1996) Identification and prevalence study of 17 allelic variants of the human NAT2 gene in a white population. Pharmacogenetics 6:423–428

Cambien F, Poirier O, Nicaud V, Herrmann SM, Mallet C, Ricard S, Behague I, Hallet V, Blanc H, Loukaci V, Thillet J, Evans A, Ruidavets JB, Arveiler D, Luc G, Tiret L (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. Am J Hum Genet 65:183–191

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238

Chiano MN, Clayton DG (1998) Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 62:55–60

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Grossman PD, Bloch W, Brinson E, Chang CC, Eggerding FA, Fung S, Iovannisci DM, Woo S, Winn-Deen ES, Iovannisci DM (1994) High-density multiplex detection of nucleic acid sequences: oligonucleotide ligation assay and sequence-coded separation. Nucleic Acids Res 22:4527–4534

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hodge SE, Boehnke M, Spence MA (1999) Loss of information due to ambiguous haplotyping of SNPs. Nat Genet 21:360–361

Hoh J, Hodge S (2000) A measure of phase ambiguity in pairs of SNPs in the presence of linkage disequilibrium. Hum Hered 50:359–364

Judson R, Stephens JC, Windemuth A (2000) The predictive power of haplotypes in clinical response. Pharmacogenomics 1:5–16

Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. Hum Mol Genet 7:1745–1751

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua RE, Odunsi A, Grigorenko E, Bonne-Tamir B, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations . Am J Hum Genet 66:1882–1899

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. Science 241:1077–1080

Lewontin RC (1964) The interaction of selection and linkage. Genetics 49:49–67

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

Martin ER, Gilbert JR, Lai EH, Riley J, Rogala AR, Slotterbeck BD, Sipe CA, Grubber JM, Warren LL, Conneally PM, Saunders AM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM (2000) Analysis of association at single nucleotide polymorphisms in the APOE region. Genomics 63:7–12

Michalatos-Beloin S, Tishkoff SA, Bentley, KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841–4843

Moffatt MF, Traherne JA, Abecasis GR, Cookson WOCM (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. Hum Mol Genet 9:1011–1019

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Ruano G, Kidd KK (1989) Direct haplotyping of chromosomal segments from multiple heterozygotes via allele-specific PCR amplification. Nucleic Acids Res 17:8392

Ruano G, Kidd KK, Stephens JC (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Pro Nat Acad Sci USA 87:6296–6300

Stephens M, Smith NJ, Donnelly P (2001a) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stephens M, Smith NJ, Donnelly P (2001b) Reply to Zhang et al. Am J Hum Genet 69:912–914

Templeton AR (1999) Uses of evolutionary theory in the human genome project. Annu Rev Ecol Syst 30:23–49

Templeton AR, Sing CF, Kessling A, Humphries S (1988) A cladistic analysis of phenotype association with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural population. Genetics 120:1145–1154

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–522

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparison of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906–912

Zollner S, Haeseler A von (2000) A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am J Hum Genet 66:615–628