A comprehensive literature review of haplotyping software and methods for use with unrelated individuals

*Rany M. Salem,[1,2,3] Jennifer Wessel[1,2,3] and Nicholas J. Schork[1,2]**

[1] Polymorphism Research Laboratory, Department of Psychiatry, University of California, San Diego, CA, USA
[2] Department of Family and Preventive Medicine, University of California, San Diego, CA, USA
[3] Graduate School of Public Health, San Diego State University, San Diego, CA, USA
* *Correspondence to*: Tel: +1 858822 5571; Fax: +1 858822 2113; E-mail: nschork@ucsd.edu

Date received: 18th January 2005

Abstract

Interest in the assignment and frequency analysis of haplotypes in samples of unrelated individuals has increased immeasurably as a result of the emphasis placed on haplotype analyses by, for example, the International HapMap Project and related initiatives. Although there are many available computer programs for haplotype analysis applicable to samples of unrelated individuals, many of these programs have limitations and/or very specific uses. In this paper, the key features of available haplotype analysis software for use with unrelated individuals, as well as pooled DNA samples from unrelated individuals, are summarised. Programs for haplotype analysis were identified through keyword searches on PUBMED and various internet search engines, a review of citations from retrieved papers and personal communications, up to June 2004. Priority was given to functioning computer programs, rather than theoretical models and methods. The available software was considered in light of a number of factors: the algorithm(s) used, algorithm accuracy, assumptions, the accommodation of genotyping error, implementation of hypothesis testing, handling of missing data, software characteristics and web-based implementations. Review papers comparing specific methods and programs are also summarised. Forty-six haplotyping programs were identified and reviewed. The programs were divided into two groups: those designed for individual genotype data (a total of 43 programs) and those designed for use with pooled DNA samples (a total of three programs). The accuracy of programs using various criteria are assessed and the programs are categorised and discussed in light of: algorithm and method, accuracy, assumptions, genotyping error, hypothesis testing, missing data, software characteristics and web implementation. Many available programs have limitations (eg some cannot accommodate missing data) and/or are designed with specific tasks in mind (eg estimating haplotype frequencies rather than assigning most likely haplotypes to individuals). It is concluded that the selection of an appropriate haplotyping program for analysis purposes should be guided by what is known about the accuracy of estimation, as well as by the limitations and assumptions built into a program.

Keywords: *haplotype, haplotyping, genetic variation, phase, algorithm, software*

Introduction

The completion of the human genome project marks a significant milestone in genetic research, ushering in an era of research opportunities in the application of genomic technologies to medical and public health problems.[1–3] One area of application involves the identification and characterisation of DNA sequence variation and its relationship (or association) with, for example, disease susceptibility. Many initiatives have been put in place to facilitate relevant association studies, but the most important is the International HapMap Project (IHP).[4] The assignment and analysis of haplotype frequencies (ie the number of times alleles at different loci are observed together on the same chromosome in a sample of individuals) can not only lead to estimates of linkage disequilibrium (LD) strength, but can also be used as the basis for a number of additional phenomena and analyses — such as the comparison of population genetics structures (eg immigration rates, genetic distances, etc), the consideration of chromosome phylogeny and the estimation of the age of mutations.[5–15] Moreover, the use of haplotypes may result in considerable savings in terms of genotyping costs and power of an association study.[16–18]

Unfortunately, many current genotyping technologies are unable to resolve the phase of maternal and paternal chromosomes in unrelated individuals, and hence the actual haplotypes an individual possesses may be in doubt. This ambiguity is referred to as the 'haplotype problem', and its

complexity increases exponentially with the number of loci being studied. Although there are technologies that can be used to unambiguously resolve phase at the chromosome or DNA level, they tend to be cost prohibitive.[19–24] Haplotype analysis involving related individuals (individuals collected from families and/or pedigrees) potentially offers more information and certain advantages compared with analysis involving unrelated individuals. Family based analysis imposes additional challenges and may not be suitable for all study designs or research objectives.[5,25–27] A companion review that focuses on computer programs and issues related to haplotype analyses involving related individuals will follow.[28] Statistical procedures are therefore required to both estimate haplotype frequencies and assign the most likely haplotypes to unrelated individuals from genotype data.[23,29,30] In this paper, available computer programs for haplotype frequency estimation will be considered as well as assignment of haplotypes involving unrelated individuals. The paper builds on an earlier review,[31] recent discussions of relevant algorithms[32,33] and articles comparing different procedures.[30,34–36] Some simple recommendations are made for addressing specific research questions using available software. Finally, web-based summaries of these evaluation are available and provide greater detail than that outlined here (URL: http://polymorphism.ucsd.edu/HapSoftwareReview/).

Materials and methods

Identification of software

Available software was identified through four means: 1) searching PUBMED through to June 2004; 2) reviewing cited references of retrieved papers and reviews of papers; 3) internet searches (eg via Google); and 4) communication with investigators working in the field. The PUBMED and/or internet searches included the following terms or combinations of terms: 'haplotyping', 'haplotype', 'analysis', 'methods', 'software', 'inference', 'assignment', 'problem', 'unrelated', 'population' and 'pooled'.

The methods, features and limitations of the identified programs were evaluated using the original published articles describing the methods, the manuals associated with the software and articles comparing programs and methodologies. The assessments provided here, build on an earlier review,[31] published discussions of algorithms for haplotype analysis[32,33] and articles contrasting different methodologies.[30,34–36] Accuracy of the methods used for estimating haplotype frequencies and assigning haplotypes to individuals was considered to be of particular importance. Ideally, validation of an indirect (ie statistically-based) haplotyping method should be compared with direct, DNA sequence-derived haplotype information. Although studies with simulated data are also informative, allowing discrimination of program performance under a variety of situations, without a 'gold standard' for comparison purposes it is hard to assess the true reliability of a method. The large number of reviewed programs precludes systematic testing of the identified programs' accuracy, performance and claims. The evaluation of this large group of programs is complicated by the diversity of methods used, measures of reliability algorithms used, varying datasets and assumptions and program characteristics which limit or prevent a program from working in all instances. The authors have endeavoured to provide a thorough review of the literature of haplotyping software in unrelated individuals, but it is acknowledged that not all original authors' claims have been validated (Supplemental Table S-A provides a brief summary of reviewed articles in which programs were actually compared). Thus, there is a reliance on some authors' claims that have not been independently verified. The majority of identified programs are freely available to academic and non-profit users. Finally, recommendations are provided for specific research objectives.

Evaluation criteria

The identified computer programs were evaluated on the basis of a number of criteria and/or software features. Many of these features and criteria were considered because they reflect items that should guide the use of particular haplotyping software.

1. *Algorithms and methods:* the analytical methods and algorithms implemented in the available programs are considered. Essentially, algorithms can be divided into two broad classes: parsimony and likelihood methods.
2. *Accuracy:* the accuracy of haplotyping algorithms is considered in terms of the algorithms ability to assess haplotype frequencies from a sample of unrelated individuals, as well as to assign haplotypes to particular individuals. Measures of accuracy are discussed briefly in the accuracy section and are detailed on the above-mentioned website (see supplementary Table S-B).
3. *Assumptions:* haplotyping programs often make assumptions about, for example, Hardy-Weinberg equilibrium (HWE), LD, population history and recombination. These assumptions can have an impact on the accuracy of haplotype frequency estimates and assignments.
4. *Genotyping error:* the accommodation of genotyping error in haplotype inference is considered. Programs that identify and accommodate genotyping errors are noted.
5. *Hypothesis testing:* not all programs have the ability to conduct statistical tests of hypotheses, so this feature is considered as well.
6. *Missing data:* the accommodation of missing data in haplotype analysis is considered.
7. *Software characteristics:* issues related to the usability of programs are considered, including computer system requirements, input data formats, interfaces, output, run time and sample size.
8. *Web implementation:* web-based implementations of available computer programs are considered.

Results

Forty-six haplotyping programs were identified and reviewed. The programs were divided into two groups: those designed for analyses involving individual genotype data from unrelated individuals (a total of 43 programs) and those designed for analysis of DNA pools (three total programs). An overview of reviewed programs is presented in Tables 1–4 and in Supplemental Tables S1-S4, S-A and S-B: (http://polymorphism.ucsd.edu/HapSoftwareReview/). Additional information on the software programs discussed in this paper, links and contact information for programs, all supplemental tables, updates to existing software and newly released software are available at the following website: http://polymorphism.ucsd.edu/HapSoftwareReview/.

The majority of identified programs for estimating haplotype frequencies and assigning them to individuals use methods rooted in likelihood theory (eg for estimation purposes — primarily the maximum likelihood approach). From a survey of the literature, it appears that most of the programs give similar results, although performance is not always consistent. No group or individual program appears to work well in all situations, or have all the features one might like to see implemented in a haplotype analysis program. It appears that accuracy and performance are affected by the characteristics of the data to be analysed and the characteristics of the population from which the individuals are sampled.

Haplotyping in unrelated populations

Algorithms and methods. A number of different analytical methods have been proposed for haplotype analysis involving unrelated individuals (see Table 1 and Supplemental Table S1). Ultimate classification of haplotyping algorithms is difficult, since implemented algorithms are often modified and combined in programs. A broad classification can be made, however, between algorithms based on parsimony and algorithms based on likelihood theory. An overview of each of these classes is provided below.

Methods based on parsimony: in 1990, Clark[37] proposed an innovative method of constructing haplotypes using a rule-based algorithm. This simple method uses the frequencies of individuals whose haplotypes are known with certainty (eg individuals homozygous at every loci) to draw inferences about the most likely haplotypes for individuals whose haplotypes are ambiguous, given their genotype data. HAPINFREX, which employs Clark's method, is computationally fast and efficient and has been used in a great deal of research[14,38] Limitations of the method include the requirement of unambiguous individuals in the study population, sensitivity to the order in which data are analysed, the inability to assign haplotypes to all individuals and potentially erroneous haplotype assignments.[37,39] To overcome these limitations, a pure parsimony extension, using integer linear programming, has been proposed[40,41] and implemented in the program HAPAR.[39] Extensions of parsimony methods take advantage of the 'perfect phylogeny framework'.[40] These programs apply the results of recent research that indicates that recombination is uncommon within LD blocks[16–18] for efficient and effective haplotype analysis. Perfect phylogeny haplotyping (PPH) reduces the haplotype analysis problem to a phylogeny problem[40] by making the assumptions of no recombination and infinite site mutations. Along this framework, unphased genotype data are reduced to a 'graph realisation problem' and solved using metroid theory and graph analysis in GPPH, although a unique solution is not guaranteed.[40,42] A simpler alternative method based on graph analysis is employed by DPPH.[43] Since empirical data may violate the prefect phylogeny assumption,[44] the assumption is relaxed in the 'perfect phylogeny' model implemented in HAP^{H}[44] and BPPH.[45] HAP^{H} constructs haplotypes within LD blocks using a maximum likelihood method.

Methods based on likelihood theory: the majority of programs that could be located are rooted in likelihood theory. Methods that exploit likelihood theory can be further broken down into maximum likelihood and Bayesian methods. The expectation maximisation (EM) algorithm is the most widely used haplotyping algorithm based on likelihood theory. In 1995, three research groups separately implemented and published EM-based haplotyping programs, 3locus.PAS,[46] HAPLO^{H}[47] and MLHAPFRE.[48] Excoffier and Slatkin[48] present a discussion of the challenges and limitations of applying the EM algorithm to haplotype analysis. In brief, the EM method has two parts, a likelihood function using initial parameter inputs and estimating sets of haplotypes that maximise the posterior probabilities of given genotypes. The estimates are iteratively updated to maximise the likelihood function.

The EM algorithm has been shown to be accurate via simulations,[49] and produces haplotype frequency estimates comparable to molecular haplotype frequencies.[23,29,30] Moreover, much of the error in haplotype frequency estimation associated with the EM algorithm has been found to be due to sampling error.[29,40] The EM algorithm may occasionally miscall rare or low frequency haplotypes.[29,30,49,50] Accuracy of the EM algorithm improves with increasing sample size.[49] The EM algorithm does have some limitations: it may converge to a non-global maximum, requiring restarts to ensure that a global maximum is reached[48,49] and it can make demands on memory requirements that may limit its utility with large numbers of subjects and datasets.[48,51]

Variants of the EM algorithm have been developed that allow the EM algorithm to overcome some of these constraints. The SNPHAP program handles the limitations by progressively expanding the subsets of markers and eliminating low frequency haplotypes from consideration at each step (refereed to as posterior and prior 'trimming').[52] The THESIAS program uses a stochastic variant of the EM algorithm to overcome many of its limitations.[53] Alternatively,

Table 1. Description of unrelated haplotyping programs, divided into four classes based on method.

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
Parsimony methods									
1. Simple parsimony									
HAPAR	Parsimony	HA	No	None	Overcomes limitations of HAPINFREX	May be susceptible to HWE departures	Practical limit, biallelic	PC/UNIX	39
					Increasing sample size improves accuracy				
HAPINFERX	Clark's	HA	No	None	Intuitive method, fast	May fail to start	Practical limit, biallelic/ multiallelic	UNIX	37
					Reduced number of haplotypes	Sensitive to data order			
					No limit on number of loci	Unstable and erroneous estimates			
2. Phylogeny									
BPPH	IP	HA	No	IP	Similar to HAP[H]	User interface	Practical limit, biallelic	MAC	45
					Speed				
DPPH	PP	HA	No	PP	Handles large datasets	Theoretical	Practical limit, biallelic	MAC	40,43
					Speed	Strict population assumptions			
GPPH	PP	HA	No	PP	Handles large datasets	Theoretical	Practical limit, biallelic	MAC/ PC/UNIX	40,42
					Speed	Strict population assumptions			

HAP[H]	IP	HA/HF	Yes	HWE, IP	Predicts haplotype blocks	No probability for haplotype assignments	Max 500 loci, Practical limit biallelic	Web-based	44
					Constructs haplotypes within blocks				
					Identifies block structure				
					Web-based				
Likelihood methods									
1. Maximum likelihood									
Arlequin v2.0	EM	HA/HF	No	HWE	Includes numerous population genetic analysis tools	EM issues	EM Practical Limits, biallelic/ multiallelic	JRE on MAC/PC/ UNIX	89
CHAPLIN	ECM	HF	Yes	HWE	Graphical interface	ECM algorithm needs to be compared with standard EM methods	Practical limits, biallelic/ multiallelic	PC	91
					Association tests				
					HWE assumption relaxed in case sample				
EH	EM	HF	No	HWE	Estimates haplotype frequency	EM issues	No Max, 3–4 practical max, biallelic/ multiallelic	PC	85,104
					Compares case-control HF under different assumptions	Must specify mode of inheritance and penetrance of disease			

(*continued*)

Table 1. *Continued.*

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
EHPLUS	EM	HF	No	HWE	Improves EH, more loci and polymorphic markers	Long run times for permutation calculations	Max 5 loci, 15 alleles in analysis	PC/UNIX	84
					Incorporates model-free analysis				
EM-DeCODER	EM	HA/HF	No	HWE	Program with standard EM algorithm	EM issues	Max 15 loci, biallelic	UNIX	57
FASTEHPLUS	EM	HF	No	HWE	Similar to EHPLUS, with speed improvements	EM issues	Max 5 loci, 15 alleles in analysis	PC/UNIX	105
GENECOUNTING	EM	HA/HF	Yes	HWE	Provides posterior probabilities for assigned haplotypes	Missing data limited to biallelic loci	10–15 loci practical limit, biallelic/ multiallelic	PC/UNIX	106
					Compares global and specific haplotypes between groups	EM issues			
GCHAP	EM	HA/HF	YES	HWE	Haplotypes with zero likelihood dropped to improve speed and accuracy	EM issues	20 loci practical limit, biallelic	JRE on PC/UNIX	107,108
					Similar to SNPHAP				
GS-EM	EM	HA/HF	Yes	HWE	Includes algorithm for assigning probability to genotype calls from several genotyping methods	EM issues	Practical limit, biallelic	Web-based	73

					Haplotypes constructed using assigned genotypes probability	Limited to biallelic SNPs			
					Web-based				
HAP^{Z}	EH	HA/HF	Yes	HWE	Modified version of SNPHAP that accommodates multiallelic loci	EM issues	Practical limit, biallelic/ multiallelic	PC/UNIX	106
HAPMAX	MLE	HF	No	HWE	Ease of use	Accommodates a limited number of SNPs	8 loci, biallelic	PC	109
						Interface			
HAPLO^{H}	EM	HF	Yes	HWE	Handles some missing data	EM issues	10 loci, 40 alleles max, biallelic/ multiallelic	UNIX	47
					Utilises pedigree data, if available				
					Calculates standard error				
HAPLOSCOPE	EM/MCMC	†	†	†	Platform program, incorporates SNPHAP and PHASE v1.0	See individual programs for limitations/ features	†	UNIX/ Windows	110
					Facilitates comparison/testing				
					Graphical interface, identifies tagging SNPs and LD blocks				
HAPLOVIEW	EM+PL	HA/HF	Yes	HWE	Calculates pairwise LD	EM issues	100s, practical limit, biallelic	JRE on MAC/PC/ UNIX	56

(*continued*)

Table 1. *Continued.*

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
					Checks for recombination				
					Identifies tagging SNPs				
					Accepts pedigree and unrelated genotype data				
HAPLO.STATS	EM	HA/HF	Yes	HWE	Incorporates method similar to SNPHAP, with user inputs	Requires knowledge of S-Plus 6.0 or R	Practical limit, biallelic/ multiallelic	S-PLUS 6.0 on UNIX/R on UNIX & PC	86
					Separate programs that:	EM issues			
					(1) assign haplotypes with posterior probability of assignments				
					(2) allow linear regression for trait to haplotype analysis				
					(3) calculates score statistic for haplo-type phenotype association				
HIT	EM/MCMC/ MC+PL	†	†	†	Platform program, incorporates SNPHAP and PHASE v1.0	See individual programs for limitations/ features	†	*	111
					Facilitates comparison				

					Graphical interface, identifies tagging SNPs and LD blocks				
HPLUS	EM+EE+PL	HA/HF	Yes	HWE	Provides posterior probabilities for assigned haplotypes	Requires Matlab	100 loci, biallelic	MATLAB on PC/ UNIX	55,83
					Compares haplotype frequencies between groups, adjusts for covariates	EM issues			
					Utilises pedigree data, if available				
LDSUPPORT	EM	HA/HF	Yes	HWE	Provides posterior probabilities for assigned haplotypes	EM issues	*	UNIX	29,112
					Identifies LD blocks for haplotype reconstruction				
					Examines association with disease, automation speeds process				
LOGINSERM ESTIHAPLO	EM	HA/HF	Yes	HWE	Program uses ML method to infer haplotypes for individuals with missing data	EM issues	Practical limit, biallelic/ multiallelic	PC/ UNIX	80
					Offers option to exclude individuals with missing data				

(*continued*)

Table 1. *Continued.*

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
MLHAPFRE	EM	HF	Yes	HWE	Performance improves with presence of LD	Incorporated into Arlequin	16 loci, biallelic	JRE on Mac/PC/ UNIX	48
					Performs well with large sample size	EM issues			
MLOCUS	EM	HA/HF	Yes	HWE	Provides posterior probabilities for assigned haplotypes	EM issues	11 loci, biallelic/ multiallelic	PC	46,113
					Notes observed vs. inferred haplotypes				
					Calculates pairwise LD				
OSLEM	EM	Yes	No	HWE	Modified EM algorithm that runs 2× faster	EM issues	Practical limit, biallelic	Web-based	114
PL-EM	EM+PL	HA/HF	Yes	HWE	Combines PL with EM	EM issues	100s, practical limit, biallelic	PC/UNIX	54
					EM-based version of HAPLOTYPER				
					Calculates variance of haplotype frequency estimates				
SAS Genetics	EM	HA/HF	Yes	HWE	Provides posterior probabilities for assigned haplotypes	Requires SAS	Practical limit, biallelic/ multiallelic	SAS on PC/UNIX	115
					Incorporates statistical tests and procedures	EM issues			

SNPEM	EM	HF	No	HWE	Estimates haplotype frequency by population	EM issues	10 loci, biallelic	UNIX	10
					Compares global and specific haplotype between 2 groups				
SNPHAP	EM	HA/HF	Yes	HWE	Uses posterior and prior trimming to handle large number loci	EM issues	Practical limit, biallelic	UNIX	52
					Provides posterior probabilities for assigned haplotypes				
THESIAS	S-EM	HF	Yes	HWE	Stochastic EM avoids issues of standard EM programs	S-EM algorithm needs to be compared with standard EM methods	Practical limit, 20 loci, biallelic	PC/UNIX	53,88
					Includes tests for haplotype-phenotype association				
					Accommodates large sample sizes				
WHAP	EM	†	†	†	Uses haplotype output from SNPHAP for association testing	EM issues	†	PC/UNIX	116
					Allows weighted association analysis	Requires separate haplotyping program			

(*continued*)

Table 1. *Continued.*

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
Zaykin *et al.*	EM	HF	No	HWE	Program on analysis of haplotype–phenotype association	EM issues	Practical limit, biallelic/ multiallelic	PC/UNIX	82
						Subjects with missing data ignored			
Zou and Zhao	MLE/EM	HF	Yes	HWE	Adjust haplotype frequency estimates for genotyping error	Assumes genotyping errors are random	Practical limits, biallelic/ multiallelic	*	68
					Program also works for nuclear families	Assumes error rates are known			
3locus.PAS	EM	HF	Yes	HWE	Handles some missing data	EM issues	3 loci, biallelic/ multiallelic	PC/UNIX	46
					Various tests available				
					Improves with increasing sample size				
2. Simple Bayesian									
HAPLOTYPER	MC+PL	HA/HF	Yes	HWE	Uses PL algorithm to construct haplotypes with many loci	Long run times	256 max, biallelic	UNIX	57
					Provides posterior probabilities for assigned haplotypes	Posterior probabilities may be difficult to interpret			

HAPLOREC	MC-VL	HA/HF	Yes	HWE	Uses variable length chain based on maximising LD	Restarts avoid non-global optimum	Practical limit, biallelic	Java virtual machine, v1.4 or newer	62
					Handles large number loci				
3. Coalescent-based Bayesian[e]									
Arlequin v3.0	ELB	HA/HF	No	Adaptive window	Includes numerous population genetics analyses	Long run times	1,000s, biallelic/ multiallelic	JRE on LINUX/ PC/Mac	60,89
					Handles recombination				
PHASE v2.0	MCMC+PL	HA/HF	Yes	Coalescent/ HWE	Improved run time	Departure for coalescent model may impact performance	Practical limit, biallelic/ multiallelic	PC/MAC/ UNIX	59
					Compares haplotype frequency between groups	Posterior probabilities may be difficult to interpret			
					Handles recombination				
					Provides posterior probabilities for assigned haplotypes				
PHASE v1.0	MCMC	HA/HF	No	Coalescent/ HWE	Incorporates pop-genetics and coalescence ideas	Departures for coalescent model may impact performance	Practical limit, biallelic/ multiallelic	UNIX	51
					Incorporates known phase and trios pedigrees into analysis	Slow run times			

(*continued*)

Table 1. *Continued*.

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	MAX subjects, loci and type	Platform	Ref.[d]
					Provides posterior probabilities for assigned haplotypes	Posterior probabilities may be difficult to interpret			
SLHAP v1.0	MCMC	HA/HF	Yes	Neutral coalescent/ HWE	Similar to PHASE v1.0	Departures for coalescent model may impact performance	Practical limit, biallelic/multiallelic	UNIX	58
					Missing data				
					Improved run time				

[a] Program haplotype output, individual assignment, frequency estimates or both.
[b] Ability of program to accept missing data.
[c] Program assumptions.
[d] List of references.
[e] Programs in this section make assumptions based on or draw inference from coalescent model.
[*] Could not determine from available data.
[†] See incorporated programs for features and limitations.
EE: Estimating equation; ECM: Expectation conditional maximisation algorithm; ELB: Excoffier-Laval-Balding algorithm, Bayesian; EM: Expectation maximisation algorithm; EM issues: May be sensitive to HWE departures, long run times, and non-global max (requiring multiple restarts); HF: Haplotype frequency estimate; HA: Individual haplotype assignment; HWE: Hardy–Weinberg equilibrium; IP: Imperfect phylogeny-based method; JRE: Java runtime environment; LD: Linkage disequilibrium; MAC: Program runs on Apple computer; MC: Monte Carlo algorithm, Bayesian algorithm; MCMC: Markov Chain Monte Carlo algorithm, Bayesian algorithm; MC-VL: Monte Carlo-variable length chain algorithm, Bayesian Algorithm; MLE: Maximum likelihood estimation algorithm; PC: IBM compatible personal computer; PL: Partition ligation algorithm; PP: Perfect phylogeny-based method; Practical Limit: program has no upper limit on number of markers and /or subjects, however computational and practical considerations limit this value; S-EM: Stochastic EM algorithm; UNIX: Runs on Unix operating system, including Linux, FORTRAN, Solaris and others.

Table 2. Web-based haplotyping programs and related websites.

Haplotyping program	Website	Comments
GS-EM	http://episun7.med.utah.edu/~alun/software.html	Assigns probability to genotype calls
		Constructs haplotypes using probabilities
HAP[H]	http://www.calit2.net/compbio/hap/	Imperfect phylogeny method
		Handles missing data
OSLEM	http://genome3.cpmc.columbia.edu/~genome/HDL/	Modified EM algorithm that runs faster
		No missing data
PHASE v2.02	http://archimedes.well.ox.ac.uk/pise/	MCMC method
		Comparison HF between groups
		Handles missing data and recombination
SNPEM	http://polymorphism.ucsd.edu/snpem/	EM method
		Comparison HF between groups
SNPHAP	http://archimedes.well.ox.ac.uk/pise/	EM method
		Handles missing data
Haplotyping-related websites		
Boas Center for Genomics and Human Genetics: North Shore LIJ Research Institute	http://www.nslij-genetics.org/soft/	Comprehensive list of statistical genetics software
		Mirrored by Rockefeller site
International HapMap Project	http://www.hapmap.org	HapMap project news, data and information
Laboratory of Statistical Genetics at Rockefeller University	http://linkage.rockefeller.edu/soft/	Comprehensive list of statistical genetics software
		Mirror of North Shore LIJ site

(*continued*)

Table 2. *Continued*.

Haplotyping program	Website	Comments
MRC Rosalind Franklin Centre for Genomics Research	http://www.hgmp.mrc.ac.uk/Registered/Menu/alphabet.html	Registration required
		Numerous programs available
Power for Association with Errors	http://linkage.rockefeller.edu/pawe/	Calculates power and sample size in the presence of differing genotype error rates
PRL: Polymorphism Research Lab	http://polymorphism.ucsd.edu/	Additional information and links to all reviewed programs
The Wellcome Trust Centre for Human Genetics	http://archimedes.well.ox.ac.uk/pise/	Several additional programs with links to their sources

EM: Expectation maximisation algorithm; HF: Haplotype frequency estimate; MCMC: Markov Chain Monte Carlo algorithm, Bayesian.

Table 3. Haplotyping software for hypothesis testing and analysis.

Program name	Haplotyping algorithm	Key analysis feature(s)	Discrete outcome[a]	Continuous outcome[b]
Hypothesis				
CHAPLIN	ECM	Includes likelihood ratio statistic and score statistic for haplotype–phenotype analysis, uses permutation test to determine significance	Yes, case-control	No
		Includes AIC for model selection, does not accommodate covariates		
EH	EM	Test for LD for unrelated and in case-control	Yes, case-control	No
		Test for frequency difference between case-control under: H1 association, H2 association for all loci		
EHPLUS	EM	Improves on EH	Yes, case-control	No
		Model-free analysis and permutation test		
FASTEHPLUS	EM	Implements EH and EHPLUS test	Yes, case-control	No
		Significant speed improvements		
GENECOUNTING	EM	Compares overall and specific haplotype frequency between cases and controls	Yes, case-control	No
HAP^{H}	IP	Phylogeny based haplotyping method	Yes, case-control	Yes
		Uses information from phylogeny for analysis, includes parametric and non-parametric tests for qualitative and quantitative phenotypes		
HAPLO.STATS	EM	Score statistic for haplotype–phenotype analysis	Yes, binary, ordinal, & Poisson	Yes
		GLM for regression of trait on haplotype, adjustment for covariates and interaction		
HPLUS	EE + PL + EM	Compares haplotypes frequency between cases and controls, option to adjust for covariates, and interaction assessment	Yes, case-control	No
		Reports OR, confidence interval and identifies haplotype blocks		
LDSUPPORT	EM	Uses likelihood method to calculate risk of developing disease phenotype from diplotype configuration	Yes, case-control	No
PHASE v2.0	MCMC	Allows comparison of haplotype frequency between populations	Yes, case-control	No

(*continued*)

Table 3. *Continued*.

Program name	Haplotyping algorithm	Key analysis feature(s)	Discrete outcome[a]	Continuous outcome[b]
THESIAS	S-EM	Compares haplotypes frequency between cases and controls, survival analysis, option to adjust for covariates and interaction assessment	Yes, case-control, survival analysis	Yes
		Uses Chi-square statistics/t-test for analysis		
SAS Genetics	EM	Allows comparison of haplotype frequency between populations	Yes, case-control	Yes
		Haplotype trend regression (HTR) and several population genetic tests		
		TDT test for family data		
SNPEM	EM	Compares overall and specific haplotype frequency between cases and controls	Yes, case-control	No
		Includes batch feature for sliding windows analysis		
WHAP	EM	Uses SNPHAP for regression based haplotype association test on SNPs, provides beta estimates of effects	Yes, case-control	Yes
		Includes haplotype weighted likelihood analysis, permutation tests and sliding windows analysis		
Zaykin et al.	EM	Likelihood ratio statistic for haplotype–phenotype analysis	Yes, case-control	Yes
		Allows sliding windows analysis		
3locus.PAS	EM	Test for global disequilibrium, including pairwise and three way disequilibrium for an unrelated sample	No	No
Other analysis programs				
Arlequin v2.0/3.0	EM/ELB	Several population genetic tests		
Zou and Zhao	EM	Adjust haplotype frequency estimates for genotyping error		

[a] Qualitative phenotype.
[b] Quantitative phenotype.
AIC: Akaike's Information Criterion; EE: Estimating equation; ELB: Excoffier-Laval-Balding algorithm, Bayesian; ECM: Expectation Conditional maximisation algorithm; EM: Expectation maximisation algorithm; GLM: General Linear Model; IP: Imperfect phylogeny; MCMC: Markov Chain Monte Carlo algorithm, Bayesian algorithm; OR: Odds Ratio; PL: Partition ligation algorithm; S-EM: Stochastic EM algorithm.

Table 4. Description of programs designed for pooled samples.

Program name	Algorithm	Output[a]	Missing data[b]	Assumptions[c]	Key features	Limitations	Pool Size, MAX # Loci, Type	Platform	Ref.[d]
Pools2	Clark's/EM	HF/HA	N/A	None	Haplotype-tagging SNPs	Computationally slow	Pools of 2 individuals, practical limit, biallelic	PC	117
					Accommodates a large number of SNPs	Need to re-calculate several times to assure consistent results			
						EM issues			
LDPooled	EM	HF/HA	No	HWE	Calculates LD	LD impacts performance	Based on pools of 4 individuals, practical limit, biallelic	*	96
					SNPs or microsatellites	EM issues			
EHP.R	EM	HF	Yes	HWE	Tests haplotype-disease association	Variance increases with pool size, weaker LD and # loci	Pools of 4 individuals, practical limit, biallelic	PC/UNIX	98
					Assessment of haplotype frequency estimate accuracy	EM issues			
					Handles different types of missing data	Requires knowledge of S-Plus 6.0 or R			

[a] Program haplotype output, individual assignment, frequency estimates or both.
[b] Ability of program to accept missing data.
[c] Program assumptions.
[d] List of references.
*Could not determine from available data.
EM: Expectation maximisation algorithm; EM issues: May be sensitive to HWE departures, long run times, and non-global max (requiring multiple restarts); HF: Haplotype frequency estimate; HA: Individual haplotype assignment; HWE: Hardy–Weinberg equilibrium; LD: Linkage disequilibrium; PC: IBM compatible personal computer; UNIX: Runs on Unix operating system, including Linux, FORTRAN, Solaris and others.

the PL-EM program combines a partition-ligation (PL) strategy with the EM algorithm to allow haplotyping of hundreds of loci.[54–56] The HPLUS program combines the EM likelihood function with an estimating equation and the PL model to efficiently handle construction of large haplotypes with missing data.[55]

The second class of likelihood algorithms are based on Bayesian estimators and Bayesian-based numerical strategies, such as Gibbs sampling.[51,57–61] Bayesian methods use different models or prior assumptions to model haplotype frequencies, and as such can be tailored to different settings, thereby improving its accuracy. Bayesian haplotype analysis methods can be further subdivided into 'simple' and 'coalescent-based' methods. The simple methods make no assumption about the history of the populations from which samples of individuals have been drawn. Simple Bayesian programs include HAPLOTYPER and HAPLOREC. HAPLOTYPER uses a statistical method similar to EM.[57] HAPLOREC implements a Bayesian method using a Variable Length Markov Chain chain approach.[62] The coalescent-based Bayesian methods essentially take similarities between and among haplotypes into account. This class includes the widely-used program, PHASE. The latest version of PHASE (v2.0) incorporates an updated algorithm to improve accuracy and the PL algorithm to improve performance time.[59] A modified model, the neutral coalescent model, is implemented in SLHAP v1.0.[58] SLHAP v1.0 builds on PHASE v1.0 to include modifications to improve computation time and to accommodate missing data.[58] Finally, Arlequin (version 3.0) draws on the coalescent model, exploiting a relaxed definition for similar haplotypes in an adaptive window approach.[60]

Accuracy. The accuracy of available programs was assessed through consideration of published articles investigating haplotype frequency estimation and assignment accuracy, including comparisons to molecular and simulated haplotype data. The measurement of the accuracy of a haplotyping method necessitates a comparison, comparing observed haplotype assignments and/or frequency estimates to expected haplotypes. The 'gold standard' for comparison is DNA sequence-derived haplotype information. The advantage of using accurate molecular haplotype data is that no assumptions, guiding, for example, simulations, are specified. The accuracy of a specific program is not influenced or biased by assumptions imposed in simulated data. Additional testing, including the discrimination of program performance under a variety of situations and assumptions is facilitated with use of simulated data.

Comparison of accuracy between haplotyping programs is a taxing venture, complicated by a variety of issues. A significant challenge is that most programs have not been directly compared with each other (Supplemental Table S-A provides a brief overview of retrieved articles that compared accuracy and performance of programs). Only a small set of programs are compared in each individual paper. Comparison of accuracy and performance of these select programs is often carried out with different datasets and under varying conditions.

A further challenge is that numerous measures have been used to assess accuracy, and these vary across publications, which are described in the reviewed literature. In brief, several measures of global accuracy of frequency estimates/assignments were found: discrepancy, error rate, mean square error (MSE), similarity index I_f and similarity index I_S, in addition to several measures comparing similarity of incorrect haplotype assignments to true haplotypes: hamming distance 'error rate H', similarity index I_G, single site error rate and switch accuracy (see Supplemental Table S-B for detailed accuracy definitions). Divergent results may be attributable to the method of accuracy measurement. Unfortunately, a comparison of the different accuracy measures was not identified in reviewed literature.

To illustrate this, a relatively simple example of four articles that all focus on comparing the PHASE (v1.0) program to EM-based programs is provided here. An original publication describing PHASE (v1.0) reported that the program outperformed other haplotyping methods, reducing MSE rates by more than 50 per cent relative to the HAPINFREX program and a program with a standard EM algorithm.[51] A subsequent comparison[35] between PHASE v1.0 and a standard EM program comparing accuracy, measured by discrepancy error rates, showed that average error rates did not differ statistically between EM-based methods and PHASE v1.0. This finding was seen across simulated and phase-known data.[35] In rebuttal, Stephens *et al.*[63] showed that PHASE v1.0 outperforms HAPLOTYPER and PL-EM, with lower error rates on data simulated to fit a coalescent model. The results were reversed when a dataset of molecular haplotypes was used, where HAPLOTYPER and PL-EM were comparable, with both outperforming PHASE v1.0.[57]

As this example demonstrates, characteristics inherent to a specific dataset whether molecular or simulated data, influence the performance and accuracy of a program. This may influence the perceived accuracy and performance of a haplotyping program. Moreover, the studies did not compare identical set of programs. Both Stephens *et al.*[51] and Zhang *et al.*[35] employed their own standard versions of the EM algorithm, which should be comparable but may not have identical specifications. A further challenge is that, while PL-EM is an EM-based program, it is one of several EM programs that have been modified to overcome performance problems of the EM algorithm, as discussed previously. Therefore, the improvement in the performance of the EM-based program, PL-EM, versus PHASE may not necessarily be generalisable to all EM-based programs. To overcome these problems, Stephens *et al*[59] compared their updated version of PHASE (v2.0) with several programs, using the same datasets and measures of accuracy as published comparisons of PHASE v1.0 to other programs.[57,58]

Overall, programs based on the Bayesian principles, EM algorithm and imperfect phylogeny performed similarly with

sequence-derived and simulated haplotype data. As shown previously,[31] no program or algorithm clearly distinguished itself from the rest. While Clark's intuitive method has shown utility, the present assessment of the literature suggests that other methods offer distinct advantages. The performance of all programs is affected by model assumptions and population genetic parameters. The impact of these assumptions is discussed below.

Assumptions. This section focuses on several common assumptions incorporated in haplotyping programs. Departures from or violations of these assumptions may affect program accuracy and performance. The assumptions are related to each other; violation of one assumption may lead to violation of a second. For ease of evaluation and discussion, each assumption is addressed separately. Program assumptions (HWE, LD, population history, etc) are noted in Tables 1 and S1.

Hardy-Weinberg equilibrium: as described in Tables 1 and 4, many programs — including all EM algorithm-based programs — assume HWE. Algorithms that assume HWE may be sensitive to departures from this assumption. Departures from HWE arise either from excess homozygosity or heterozygosity at a locus in a population. Measures evaluating departures from HWE have been shown to correlate with haplotype frequency estimation and assignment inference accuracy.[57] Increases in homozygosity tend to decrease the number of ambiguous individuals (ie individuals whose phase cannot be determined with certainty) and have been shown to have little impact on the accuracy of the EM-based method, as measured by the MSE.[49,64] By contrast, accuracy decreases with HWE departures resulting from increased heterozygosity. Comparing the performance of HAPINFREX, EM-DECODER, PHASE v1.0 and HAPLOTYPER in simulated data with varying HWE departures found that all methods showed increased error levels with excess heterozygosity.[57] HAPINFREX was most vulnerable to HWE departures, particularly underperforming in situations with low numbers of homozygotes. Performance improves rapidly with increasing proportions of homozygotes in a population.[57] In data with a significant proportion of homozygous individuals, HAPINFREX outperformed PHASE v1.0.[57] In an evaluation of HPLUS on simulated data with HWE departures, accuracy improved with increasing sample size, although little benefit was achieved with samples beyond 100 subjects.[55]

Linkage disequilibrium and recombination: research suggests that recombination hotspots — that is, chromosomal segments with high levels of recombination — tend to be separated by extended LD or haplotype 'blocks' exhibiting little recombination and strong LD. This structuring of LD blocks may be common in the human genome.[16-18,65] Highly variable recombination rates in a small genomic region may violate assumptions of the current coalescent-based programs;[51,58] however, all methods may have problems constructing haplotypes across regions with high levels of recombination[57,60] and low LD.[36] While a majority of programs do not make explicit assumptions about LD, the performance of both EM methods[29,36,48,64] and PHASE v1.0[51] has been shown to improve with increasing LD. Comparisons of the accuracy of PHASE v1.0, HAPLOTYPER and Arlequin v3.0, showed that accuracy was adversely affected by increases in the recombination rate.[60] Doubling in theta (θ) — that is, the mutation rate per locus — results in a 5–10 per cent decrease in accuracy for both Arlequin v3.0 and PHASE v1.0. By contrast, the global accuracy of HAPLTOYPER increased with theta in some situations.[60] In this comparison, Arlequin v3.0 demonstrated the highest accuracy in the presence of recombination, by using a sliding windows approach to phase loci. Performance measured by a similarity index for HPLUS declined with increasing number of single nucleotide polymorphisms (SNPs) for a simulated dataset with recombination, although this trend was not observed with MSE.[55]

The PL method used by HAPLOTYPER was shown to be insensitive to the presence of recombination hotspots, although extensive recombination may be problematic.[57] Accuracy improves when hotspots are used as the partition sites, however.[54,57] PL-EM allows users to specify the partition size, thereby allowing partitioning at the hotspot. Focusing on DNA segments in LD offers a method to overcome the challenges and errors related to haplotyping in the presence of recombination hotspots. Since the recombination hotspots are not known in advance, automating the identification of LD block boundaries, haplotyping within blocks may offer significant benefits[40,57] Several programs, notably HAP^{H}, SLHAP v1.0 and PHASE v2.0, have exploited this methodology. SLHAP v1.0[58] and HAP^{H} have been reported to improve the accuracy of inferred haplotypes. A related approach limits haplotype analysis to segments in LD. HAPLOREC based on the variable-length chains allows the program to obtain different length haplotype fragments in different regions, based on the LD strength.[62] A drawback of these methods is that it may lead to a loss of phase information.[66] PHASE v2.0 incorporates a separate algorithm to accommodate recombination, based on the method proposed by Fearnhead and Donnelly.[67]

Evaluation of linkage and recombination is an important first step in haplotype analysis. The HAP^{H} and HAPLOVIEW programs identify haplotype blocks in a graphical display. Data that contain recombination hotspots may pose a challenge to haplotyping software that assumes no recombination. Decreases in LD are correlated with increasing estimation error[36] and magnify the effects of genotyping error;[68] thus, although haplotyping with loci whose alleles are in low LD is important, haplotype estimates from such data may be unreliable. Further study in this area is required, particularly in situations of intermediate LD levels; the influence of LD level on accuracy and determination of the LD level that, if surpassed, improves accuracy. This is not trivial, especially if many loci are considered, each with varying degrees of LD by comparison with the others.

As one would expect, recombination leads to an increase in the number of haplotypes, including low frequency haplotypes that are difficult to estimate accurately.[36,49,53] Increasing sample size may improve haplotyping accuracy in the presence of high recombination.[39] Finally, analysing chromosome segments on either side of a recombination hotspot is most likely to be the only current viable option.[8]

Population evolutionary history: several programs impose assumptions on the evolutionary history of the populations from which samples have been obtained to improve program efficiency and accuracy and simplify haplotype analysis. The PHASE program is the best-known example of a program that incorporates a population evolutionary history model — in this case the coalescent model.[51,59] Moreover, the SLHAP v1.0[58] and Arlequin v3.0[60] programs are based on variants of the coalescent model. Several programs exploit the 'perfect phylogeny' concept. These programs (GPPH, DPPH and BPPH) are reported to be fast and accurate and to accommodate large numbers of markers.[40,42,43,45] The HAP^{H} program uses a relaxed model — imperfect phylogeny — to make the model more amenable to what is currently known about population evolutionary history.[44]

The benefit of incorporating an evolutionary model, such as the coalescent model, is to take advantage of similarities between haplotypes; it is thought to result in more accurate haplotypes than other methods.[51,59] The disadvantage is that the behaviour of alleles in the short-term evolution of chromosomes may violate the model, potentially leading to errors. By contrast, HAPLOTYPER, HAPINFREX and HAPAR impose no population evolutionary history assumptions. Program performance and accuracy may be affected when data fit or do not fit the program's population assumption. To illustrate, Stephens *et al.*[51] note that PHASE v1.0, by comparison with EM algorithm-based methods, would reduce error rates by 50 per cent when data fit the coalescent model. When compared to PL-EM, using similar data, the improvement in error rate was 26 per cent lower than that shown by Stephens *et al.* for data that fit the coalescent model.[54]

The coalescent model is appropriate for stable populations that have evolved over long periods of time, but is less suitable for populations with past gene flow, stratification and/or population migration. There is disagreement as to whether haplotyping programs based on the coalescence model are the most appropriate for accurate haplotyping.[35,51,57] Even when data do not fit the coalescent model, the performance of PHASE v1.0 is suggested to be no worse than that of EM methods.[63] Using simulated data that violate the coalescent model, Niu *et al.*[57] showed that HAPLTOYPER and EM-DECODER are more accurate than PHASE v1.0 and HAPINFREX. The decline in performance of PHASE v1.0 in at least one of the instances may have been due to insufficient updates rather than model assumptions.[59] The findings of Niu *et al.* were supported in a subsequent comparison of PHASE v1.0, HAPLOTYPER and Arlequin v3.0.[60] Arlequin v3.0 had the highest accuracy of the three programs when the coalescent model was violated. In a comparison of PHASE v1.0, HAPINFREX, HAPAR and HAPLOTYPER using data modelled to fit the coalescence model, PHASE v1.0 yielded the lowest error rate, followed by HAPAR.[39] The updated version of PHASE v2.0 demonstrated improved performance with molecular haplotype data, exceeding the performance of HAPLOTYPER, SLHAP v1.0 and the earlier version of PHASE.[59] An additional study assessed performance of PHASE v1.0, HAPAR and HAPLOTYPER using data simulated to fit the phylogeny model, an evolutionary model related to the coalescence model. The comparison found that PHASE v1.0 had the lowest error rate, followed by HAPAR and HAPLOTYPER. Error rates became similar for the three programs as sample size increased.[39] In summary, programs that assume a population evolutionary history of data should be used with care, since departures from model assumptions may have a significant impact on the accuracy of haplotype assignments and estimates. This should in no way detract from the utility and flexibility of these programs, but serves to illustrate that model assumptions should be considered when these programs are used.

Genotyping error. Genotyping error is a form of misclassification which can lead to deleterious effects on the power of association analyses,[69–72] LD measurements[69] and erroneous haplotype analysis.[60,68,73,74] The power of SNP association studies decreases with even relatively small genotyping error rates.[71] A similar trend may exist for haplotype association studies, although further examination is required. Sample size requirements of varying SNP error rates and power levels can be examined at the Power for Association with Error (PAWE) website[70,71] (see Tables 2 and S2).

Most genotyping errors are due to allelic dropout (missing data) and the inability to score heterozygotes, resulting in an increased proportion of homozygotes.[73,75] Non-random distributions of missing genotypes represent an error in genotype assignments. Programs that deal with missing data often do so by assuming that data are missing at random. Spurious haplotypes may be introduced if loci with genotype errors are included in haplotype analysis.[60] Error rates of 5 per cent may bias haplotype estimates by as much as 30 per cent.[72] Genotyping error leads to a substantial loss in haplotype accuracy, particularly when LD is low and many rare haplotypes exist.[74] Haplotyping methods that favour similar haplotypes may be less sensitive to genotyping error.[60] Recently, Zou and Zhao[72] introduced an EM-based program that corrects haplotype frequency estimates for known genotype error rates, although determining genotyping error can be difficult in unrelated populations.[76–78] A common strategy is to genotype a subset of the study population twice, to determine error rates. Genotyping as few as 25 individuals has been shown to be sufficient for determining genotyping error in a simulation study.[76] Testing assay specificity and HWE deviations of loci are established methods for reducing genotyping error rates.[79]

Finally, the accuracy and power of association analyses may be improved by incorporating genotyping uncertainty in haplotype inference to negate the effects of genotyping errors, as in GS-EM.[73]

Missing data. Current genotyping methods often result in missing data, owing to a variety of factors, including, for example, polymerase chain reaction dropouts, inability to score loci and systematic genotyping technology errors. Missing data complicate haplotype inference by increasing the difficulty and uncertainty of haplotype estimates. Missing data decrease the available information and may bias the haplotype assignment. The majority of programs score poorly in this area, as they are unable to accommodate any missing data (see Tables 1 and 4 for programs that accommodate missing data). Some of these programs deal with missing data by ignoring subjects with any missing marker data, leading to a loss of data. Most programs assume that missing data are missing at random (see the section above, on genotyping error).

Accommodating missing data results in a performance decline, with increased memory requirements, longer run times and increased uncertainty. Several strategies have been proposed and implemented for dealing with haplotyping in the presence of missing data. The EM algorithm can be set to accommodate missing data; a discussion focusing on EM haplotyping and missing data is provided elsewhere.[80] Among EM-based programs, LOGINSERM_ESTIHAPLOE includes the option of ignoring individuals with missing data or of using them in haplotype inference, depending on research objectives,[80] whereas PL-EM allows users to specify the number of possible haplotype sets with a probability above a specific level.[54] By contrast, HAP^{H} ignores missing markers in haplotype construction, and uses a maximum likelihood method to infer missing allele(s) to match common haplotypes.[44] The accuracy of HAP^{H} was maintained with up to 10 per cent missing data. Arlequin v3.0 does not try to impute missing data in haplotype analysis, but rather ignores missing loci in the process.[60] This approach is sensitive to the amount of missing data, with small decreases in accuracy with up to 2 per cent missing data becoming more noticeable at 4 per cent. Moreover, the addition of a subset of individuals with large amounts of missing data (20 per cent) has been shown to have a detrimental effect on haplotype analysis on the larger group with complete data.[60]

A limitation of the original version of PHASE (v1.0) was that it could not accommodate missing data.[51] SLHAP v1.0, based on of PHASE v1.0's methods, includes modifications that allow accommodation of missing data.[58] The updated version of PHASE v2.0 was also adapted to accept missing data; phase at unknown positions is randomised and any missing genotypes are imputed with random guesses.[59] The HAPLOREC program also handles missing data by matching haplotypes with missing data to known haplotypes, although missing alleles are not imputed.[62] Finally, the performance of HAPLOTYPER was shown to be stable in the presence of missing data, although caution should be exercised when missing data are included.[57] Excellent discussions of the challenges of haplotyping with missing data are presented elsewhere.[57,81] The inclusion of individuals with too much missing data (>10 per cent) may have a detrimental effect on the reconstruction of phase of individuals without missing data. Finally, markers with non-random patterns of genotyping failure should be redesigned or dropped from the haplotyping set.[57,80]

Software characteristics. In this section, issues related to usability of programs are discussed. User-friendliness is an important issue in the selection of appropriate haplotyping programs, especially in terms of practical usability of programs. Relevant issues include computer system requirements, data format, interface, marker characteristics, run time and sample size.

Computer system requirements: as detailed in the 'platform' column in Tables 1 and 4, not all programs are available for use with all computer operating systems. The selection of a haplotyping program may necessitate investment in new computer equipment and training. Compiling programs to run on new operating systems poses similar challenges.

Data input format: unfortunately, there is no standard data input format. Nearly all of the programs use a unique data input format. Manipulating data from one format to work with another is cumbersome and difficult. HIT and HAPLOSCOPE are platform programs, incorporating several haplotyping programs in one interface. These programs facilitate comparisons of programs on the same datasets.

User interface: the interface is an important component of usability of a haplotyping program. Selection of a program will depend heavily on current knowledge or ability to invest time in learning about a computer system. The majority of identified programs are command prompt driven (see Tables 1 and 4). These interfaces tend to intimidate computer novices or non-computer scientists. Fortunately, several programs with a graphical user interface were identified, including: Arlequin, HAPLOVIEW, HAPLOSCOPE and HPLUS. Finally, individuals familiar with SAS and S-PLUS may be interested in the SAS Genetics module and HAPLO.STATS programs, respectively.

Marker characteristics: many of the widely-used haplotyping programs are limited to biallelic loci. Programs that accommodate multiallelic markers often experience longer run times. Allele frequency is an important consideration in the selection of markers. Low allele frequencies result in low frequency haplotypes that may have little value in explaining common disease variation.[49] Moreover, low frequency haplotypes, for a variety of reasons (eg sampling error, genotyping error, recombination and low LD), are difficult to estimate accurately.[29,30,36,49,50,53]

Output: in addition to haplotype frequency estimates and assignments, many programs provide measures for evaluating

the 'goodness of fit' of constructed haplotypes. A number of EM-based programs provide posterior probabilities of haplotype assignments, including GENECOUNTING, HPLUS, HAPLO.STATS, LDSUPPORT, MLOCUS, PL-EM and SNPHAP. Posterior probabilities are helpful for evaluation of haplotype assignment and any subsequent analyses. Moreover, the probabilities can be used to weight and evaluate assigned haplotypes and frequency estimates.[25,82] Determination and interpretation of posterior probabilities is difficult for programs that use pseudo-Gibbs samplers, including Arlequin, HAPLOTYPER and PHASE.[51,57,60] Finally, Arlequin, HAPLO^{H}, HPLUS and PL-EM provide the variance estimates for the estimated haplotype frequencies.

Run time: another issue in assessing the performance of haplotyping programs involves the programs' use of memory and demands on the central processing unit. Run time is also affected by the complexity of the haplotyping problem, which increases with the number of loci.[48,51] Although the present EM algorithm can theoretically handle an infinite number of polymorphic sites in a sample, it is limited in practice by its exponentially increasing memory requirements.[48,49] Moreover, EM methods may require multiple restarts to avoid local convergence and non-global optimum, increasing the time required to infer haplotypes.[48] Using a Gibbs sampler, PHASE v1.0 more efficiently determines phase than the EM algorithm and constructs haplotypes with a larger number of markers, although run times are lengthy.[51,58] PHASE has been universally recognised as having several useful features, but a very slow implementation.[51,55,58,60] In the original article describing PHASE v1.0, it took minutes to hours to run, whereas an EM program and HAPINFREX took seconds.[51] Among Bayesian-based programs, with 50 subjects and 14-119 loci, HAPLOTYPER estimated haplotypes in seconds, Arlequin v3.0 in minutes and PHASE v1.0 in hours.[60] In comparisons of several programs over complete datasets from Reich *et al.*,[16] HPLUS and HAPLOTYPER completed analysis in under one second, Arlequin v2.0 in less than one minute and PHASE v2.0 in 11 minutes.[55]

Additional comparisons suggest that programs that implement modified EM algorithms, such as SNPHAP and PL-EM, had shorter run times than PHASE v1.0 on large datasets. HAPLOREC has similar run times to the modified EM programs.[62] The updated version of PHASE (v2.0) improves program performance, although it was found still to be slower than the other programs.[59] The phylogeny programs (GPPH, DPPH, BPPH and HAP^{H}) have remarkably fast run times.[40,43–45] HAP^{H} was shown to run faster than both HAPLOTYPER and PHASE v1.0 in a variety of situations.[44] Run times for all programs increased in the presence of missing data and multiallelic markers.[54,60,62]

Sample *size*: both sample size and the number of loci are important components for the selection of haplotyping programs. Details on sample size and loci limits are listed in Tables 1 and S1. As sample size increases, both in terms of the number of markers and subjects, the run time increases. The accuracy of EM-based programs has been shown to improve with increasing sample size.[4,53] Likewise, the accuracy of HAPAR, HAPLOTYPER and PHASE v1.0 were also shown to improve with increasing sample size.[39] Accurate haplotyping of low frequency haplotypes improves with increasing sample size.[30]

While standard EM-based programs have no theoretical limit, in practice these programs are limited to fewer than 25 loci, due to memory and processing requirements.[48,49,51] HAPINFREX, likewise, has no practical size limits, although the program may fail to start with large numbers of markers.[37] The parsimony program, HAPAR, overcomes HAPINFREX limitations, with accuracy improving with increasing sample size.[39] Programs that accommodate large datasets often sacrifice performance. PL, a divide and conquer strategy, has been proposed as an effective method of dealing with the construction of large haplotypes.[57] This and similar schemes have been implemented in both EM-[54–56] and Bayesian-based programs.[57,59,60,62] These programs are able to handle large datasets, although performance varies (see run time discussion above).

Hypothesis testing. Haplotyping in and of itself is usually not the final outcome of interest. The research objective dictates which subsequent analyses are needed. This section will focus on programs that combine haplotyping with hypothesis testing in genetic association studies (see Table 3 and Supplemental Table S3). All haplotype reconstruction methods will encounter a degree of misclassification error or uncertainty in haplotype assignments.[7,81,83] If uncertainty of assignments is ignored in subsequent analyses, it can lead to biased parameter estimates and inflated false-positive rates for statistically-based hypothesis tests.[25,31,82,83] In situations where inferred haplotypes had high reliability, biased estimates were avoided, and found to be useful for hypothesis testing.[83] The imperfect phylogeny-based method in HAP^{H} has been shown to assign accurate haplotypes[62] and has recently been updated to include association analysis of discrete and continuous phenotypes, although the potential for bias exists, due to uncertainty of haplotype assignments. Several programs avoid this pitfall by comparing estimated haplotype frequencies between two groups,[84,85] that is, a case-control model, these include EH, EHPLUS, FASTEHPLUS, GENECOUNTING, PHASE v 2.0, SAS Genetics module and SNPEM. Fallin *et al.*[10] demonstrated the advantages of this approach using the SNPEM program.

This methodology has been extended to allow adjustment for covariates. The Zaykin[82] program uses a likelihood ratio test statistic for association analysis of haplotypes and phenotypes. HAPLO.STATS[86,87] and THESIAS[53] also include a test for interaction with covariates using a score and likelihood ratio statistic, respectively. The HPLUS program is limited to qualitative phenotypes, and it provides odds ratio estimates.[55,83] The THESIAS program has recently been

expanded to allow haplotype-based association analysis of survival outcomes.[88] Finally, Arlequin[60,89] incorporates numerous population genetics tests. Additional discussions on hypothesis testing with haplotypes are available.[82,86,90–94]

Web-based programs. Several web-based haplotyping programs were identified and are presented in Table 2 and supplemental Table S2. Web-based versions of haplotyping programs help researchers to circumvent many of the issues related to practical usability, discussed previously. Web-based programs negate the need for the researcher to learn a computer language(s), purchase computer hardware/software, install and maintain programs or to have to troubleshoot computer problems, thus allowing genetics researchers to focus on what they do best. Moreover, web-based programs usually employ graphical interfaces, allowing the computer layman easily to use a haplotyping program. Additionally, many of the identified web-based programs allow the user to select results sent via e-mail. Finally, additional websites were identified with links to programs, as well as the website for the supplemental tables, also presented in Table 2.

Haplotyping in pooled data

Haplotype analysis using pooled samples is possible, but requires that alleles are in strong LD, are severely limited to a small number of individuals and that only a few of the possible allele combinations are present.[95] This requires actual genotyping of individuals to determine which haplotypes exist in the population of interest before testing for differences in allele frequencies in the two pooled samples.[95,96] Three programs for pooled samples were identified, as well as one technique, none of which were web-based (see Table 4 and Supplemental Table S4). All of the programs are only compatible with pools of one to six individuals, in which each pool uniquely comprises cases or controls of unrelated individuals. There has been some discussion as to the number of individuals and SNPs that the pooling technique or algorithm can handle.[95–99] Pools of three to four individuals are optimal, in terms of accuracy and efficiency. Accuracy begins to decline beyond four individuals.[12] Zou and Zhao[72] point out that pooled samples are particularly susceptible to genotyping error and that consideration should be given to the impact of population stratification in pooled samples.

Discussion

While no single haplotyping program is ideal in all situations, this review found that currently available haplotyping programs should accommodate the research needs of most scientists. While the programs share many similarities, significant differences were observed in their ability to handle various data characteristics and population genetic parameters. Each program had its own unique combination of features and limitations. It is hoped that researchers interested in haplotype analysis will use this paper as a guide for selecting the haplotype analysis program(s) most suitable for their research needs. Moreover, it is anticipated that this review will be an impetus for additional testing, development and improvement of haplotyping software.

The selection of haplotyping programs should be based on the research needs and characteristics of the data to be used for analysis. These criteria include: research objectives, hypothesis testing, data assumptions, genotyping error, missing data and computer expertise to implement programs, if necessary. A suitable haplotyping program is one that generates the desired results (haplotype frequency estimates and/or assignments) and analyses. For hypothesis testing, several programs were identified that combine haplotype analysis with hypothesis testing, which should facilitate analysis. The accuracy of haplotyping programs varied under different assumptions and situations. It was found that deviations from assumptions often resulted in declines in the performance of haplotyping programs, therefore, an important step in selecting a haplotyping program is the evaluation of the assumptions inherent to collection of the data. This should identify programs that can accommodate limitations or departures from assumptions of the data.

Selection of the appropriate haplotyping programs should also take into account the usability of a program. Assessment of this criterion is challenging because usefulness depends on a number of sub-criteria, discussed previously. Web-based programs and those with graphical user interfaces will generally be the easiest to use and have the best usability. Unfortunately, only a short list of programs may suit the needs of researchers. The usability of a program will also depend heavily on the researcher's computer expertise. In summary, the choice of haplotyping program should be based on identifying research needs and selecting a haplotyping program most appropriate to accommodating those requirements. Awareness of program assumptions and limitations should be an important factor in the final decision.

All of the programs reviewed assume genetic homogeneity of individuals in study populations. In brief, the basis of this assumption is that all individuals in a study population share a similar population history. Inclusion of individuals with dissimilar population histories will result in incorrect haplotype estimates due to, for example, LD differences and allele frequency differences between the populations. As an example, consider a hypothetical population of 200 individuals: half being of African-American ancestry and half of European-American ancestry. The resulting haplotyping estimates will not be correct for either the African-American or European American groups. To obtain accurate haplotype estimates and assignments, the groups must be analysed separately. Further discussions on this topic are available elsewhere.[5,100–103]

The majority of the reviewed programs are actively maintained and updated regularly. Haplotyping analysis is a rapidly evolving field, with many new methods and programs

emerging. Programs that are reviewed here may be modified or even be completely revamped in the near future. Accurate and updated information on existing haplotyping programs will be maintained at http://polymorphism.ucsd.edu/HapSoftwareReview/. An important limitation of this project is that it relied on a review of literature to evaluate the programs. Therefore, it was not possible to validate the accuracy, performance and claims of all individual programs.

This review found that haplotype analysis programs have increased in number and have improved rapidly over the past decade. While existing haplotyping methods may accommodate research needs, many opportunities exist for improvement of haplotyping programs. In particular, improvements in accuracy (particularly for assignments), faster run time, accommodation of larger sample sets and loci, handling missing data, incorporating association testing and identification and adjustment of haplotype estimates in the presence of genotyping error. In addition, an emerging question is how to construct haplotypes across large genomic regions — especially with substantial numbers of loci. Available methods include programs that use a block-based approach, methods that build large haplotypes by adding one loci at a time (ie SNPHAP) or programs that use the PL approach (ie HAPLOTYPER, PL-EM). Future studies are necessary to directly evaluate the different measures of accuracy, assess the influence of varying of LD levels on accuracy and further assess the impact of departures of assumptions on program performance and accuracy. Ideally, future studies would evaluate several of the more commonly used programs in a standard fashion, allowing comparison across studies. This would facilitate comparison of programs and determination of the most appropriate program. Moreover, adoption of a universal data format would also be helpful. Finally, the use of a standardised phase-known dataset(s), which developers of haplotyping programs could assess for evaluating their programs, would assist in the selection, improvement and development of haplotyping programs. Potential sources include examples from the literature[4,18,65] and the HapMap project data (available at: www.hapmap.org).

Acknowledgements

We would like to thank the authors of the papers reviewed for making their programs readily available and for their helpful documentation. Aspects of this work were supported by NIH NHLBI grants HL54998-01, HL69758-01 and HL64777-03 awarded to N.J.S. R.M.S. is supported by NIH training grant T32 DA007315.

References

1. Lander, E.S., Linton, L.M., Birren, B. *et al.* (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
3. Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003), 'A vision for the future of genomics research', *Nature* Vol. 422, pp. 835–847.
4. Gibbs, R.A., Belmont, J.W., Hardenbol, P. *et al.* (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
5. Lander, E.S. and Schork, N.J. (1994), 'Genetic dissection of complex traits', *Science* Vol. 265, pp. 2037–2048.
6. Akey, J., Jin, L. and Xiong, M. (2001), 'Haplotypes vs. single marker linkage disequilibrium tests: What do we gain?', *Eur. J. Hum. Genet.* Vol. 9, pp. 291–300.
7. Judson, R. and Stephens, J.C. (2001), 'Notes from the SNP vs. haplotype front', *Pharmacogenomics* Vol. 2, pp. 7–10.
8. Judson, R., Stephens, J.C. and Windemuth, A. (2000), 'The predictive power of haplotypes in clinical response', *Pharmacogenomics* Vol. 1, pp. 15–26.
9. Bader, J.S. (2001), 'The relative power of SNPs and haplotype as genetic markers for association tests', *Pharmacogenomics* Vol. 2, pp. 11–24.
10. Fallin, D., Cohen, A., Essioux, L. *et al.* (2001), 'Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease', *Genome Res.* Vol. 11, pp. 143–151.
11. Bridges, Jr, S.L., Jenq, G., Moran, M. *et al.* (2002), 'Single-nucleotide polymorphisms in tumor necrosis factor receptor genes: Definition of novel haplotypes and racial/ethnic differences', *Arthritis Rheum.* Vol. 46, pp. 2045–2050.
12. Yang, Y., Swaminathan, S., Martin, B.K. and Sharan, S.K. (2003), 'Aberrant splicing induced by missense mutations in BRCA1: Clues from a humanized mouse model', *Hum. Mol. Genet.* Vol. 12, pp. 2121–2131.
13. Small, K.M., McGraw, D.W. and Liggett, S.B. (2003), 'Pharmacology and physiology of human adrenergic receptor polymorphisms', *Annu. Rev. Pharmacol. Toxicol.* Vol. 43, pp. 381–411.
14. Drysdale, C.M., McGraw, D.W., Stack, C.B. *et al.* (2000), 'Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 10483–10488.
15. Steinberg, M.H., Voskaridou, E., Kutlar, A. *et al.* (2003), 'Concordant fetal hemoglobin response to hydroxyurea in siblings with sickle cell disease', *Am. J. Hematol.* Vol. 72, pp. 121–126.
16. Reich, D.E., Cargill, M., Bolk, S. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature* Vol. 411, pp. 199–204.
17. Patil, N., Berno, A.J., Hinds, D.A. *et al.* (2001), 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21', *Science* Vol. 294, pp. 1719–1723.
18. Gabriel, S.B., Schaffner, S.F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
19. Ruano, G., Kidd, K.K. and Stephens, J.C. (1990), 'Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules', *Proc. Natl. Acad. Sci. USA* Vol. 87, pp. 6296–6300.
20. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L. *et al.* (1996), 'Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR', *Nucleic Acids Res.* Vol. 24, pp. 4841–4843.
21. Ding, C. and Cantor, C.R. (2003), 'Direct molecular haplotyping of long-range genomic DNA with M1-PCR', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 7449–7453.
22. Douglas, J.A., Boehnke, M., Gillanders, E. *et al.* (2001), 'Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies', *Nat. Genet.* Vol. 28, pp. 361–364.
23. Tost, J., Brandt, O., Boussicault, F. *et al.* (2002), 'Molecular haplotyping at high throughput', *Nucleic Acids Res.* Vol. 30, p. e96.
24. Kwok, P.Y. and Xiao, M. (2004), 'Single-molecule analysis for molecular haplotyping', *Hum. Mutat.* Vol. 23, pp. 442–446.
25. Schaid, D.J. (2002), 'Relative efficiency of ambiguous vs. directly measured haplotype frequencies', *Genet. Epidemiol.* Vol. 23, pp. 426–443.
26. Becker, T. and Knapp, M. (2002), 'Efficiency of haplotype frequency estimation when nuclear family information is included', *Hum. Hered.* Vol. 54, pp. 45–53.

27. Rohde, K. and Fuerst, R. (2001), 'Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information', *Hum. Mutat.* Vol. 17, pp. 289–295.
28. Wessel, J., Salem, R.M. and Schork, N.J., (2005), 'A comprehensive review of haplotyping software and methods for use with related individuals', *Hum. Genomics* (submitted).
29. Kitamura, Y., Moriguchi, M., Kaneko, H. *et al.* (2002), 'Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm', *Ann. Hum. Genet.* Vol. 66, pp. 183–193.
30. Tishkoff, S., Pakstis, A., Ruano, G. and Kidd, K. (2000), 'The accuracy of statistical methods for estimation of haplotype frequencies: An example from the CD4 locus', *Am. J. Hum. Genet.* Vol. 67, pp. 518–522.
31. Weale, M. (2004), 'A survey of current software for haplotype phase inference', *Hum. Genomics* Vol. 1, pp. 141–144.
32. Gusfield, D. (2004), 'An overview of combinatorial methods for haplotype inference', (*unpublished*).
33. Bonizzoni, P., Vedova, G., Dondi, R. and Li, J. (2003), 'The haplotyping problem: An overview of computational models and solutions', *J. Comput. Sci. Technol.* Vol. 16, pp. 675–688.
34. Orzack, S.H., Gusfield, D., Olson, J. *et al.* (2003), 'Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes', *Genetics* Vol. 165, pp. 915–928.
35. Zhang, S., Pakstis, A.J., Kidd, K.K. and Zhao, H. (2001), 'Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data', *Am. J. Hum. Genet.* Vol. 69, pp. 906–912.
36. Xu, C.F., Lewis, K., Cantone, K.L. *et al.* (2002), 'Effectiveness of computational methods in haplotype prediction', *Hum. Genet.* Vol. 110, pp. 148–156.
37. Clark, A. (1990), 'Inference of haplotypes from PCR-amplified samples of diploid populations', *Mol. Biol. Evol.* Vol. 7, pp. 111–122.
38. Clark, A.G., Weiss, K.M., Nickerson, D.A. *et al.* (1998), 'Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase', *Am. J. Hum. Genet.* Vol. 63, pp. 595–612.
39. Wang, L. and Xu, Y. (2003), 'Haplotype inference by maximum parsimony', *Bioinformatics* Vol. 19, pp. 1773–1780.
40. Gusfield, D. (2002), 'Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions', paper presented at the sixth annual conference on Research in Computational Molecular Biology (RECOMB), Washington, DC, April 18–21, available at http://www.csif.cs.ucdavis.edu/rgusfield/paperlist.html/.
41. Gusfield, D. (2003), 'Haplotype inference by pure parsimony', UC Davis Computer Science Engineering Technical Report CSE-2003-2, 24th January, pp. 1–10, available at http://www.cs.ucdavis.edu/research/tech-reports/2003/CSE-2003-2.pdf.
42. Chung, R.H. and Gusfield, D. (2003), 'Perfect phylogeny haplotyper: Haplotype inferral using a tree model', *Bioinformatics* Vol. 19, pp. 780–781.
43. Bafna, V., Gusfield, D., Lancia, G. and Yooseph, S. (2003), 'Haplotyping as perfect phylogeny: A direct approach', *J. Comput. Biol.* Vol. 10, pp. 323–340.
44. Halperin, E. and Eskin, E. (2004), 'Haplotype reconstruction from genotype data using imperfect phylogeny', *Bioinformatics*. Vol. 20, pp. 1842–1849.
45. Chung, R.H. and Gusfield, D. (2003), 'Empirical exploration of perfect phylogeny haplotyping and haplotypers', in: Warnow, T. and Zhu, B. (eds), *Lecture Notes in Computer Science*, Springer, Big Sky, MT, pp. 5–19.
46. Long, J., Williams, R. and Urbanek, M. (1995), 'An E-M algorithm and testing strategy for multiple-locus haplotypes', *Am. J. Hum. Genet.* Vol. 56, pp. 799–810.
47. Hawley, M. and Kidd, K. (1995), 'HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes', *J. Hered.* Vol. 86, pp. 409–411.
48. Excoffier, L. and Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.* Vol. 12, pp. 921–927.
49. Fallin, D. and Schork, N.J. (2000), 'Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data', *Am. J. Hum. Genet.* Vol. 67, pp. 947–959.
50. Saito, M., Saito, A. and Kamatani, N. (2002), 'Web-based detection of genotype errors in pedigree data', *J. Hum. Genet.* Vol. 47, pp. 377–379.
51. Stephens, M., Smith, N. and Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *Am. J. Hum. Genet.* Vol. 69, pp. 906–914.
52. Clayton, D. (2001). 'SNPHAP a program for estimating frequencies of haplotypes of large numbers of diallelic markers from unphased genotype data from unrelated subjects, ver 1.0', available at http://www-gene.cimr.cam.ac.uk/clayton/software/.
53. Tregouet, D.A., Escolano, S., Tiret, L. *et al.* (2004), 'A new algorithm for haplotype-based association analysis: The stochastic-EM algorithm', *Ann. Hum. Genet.* Vol. 68, pp. 165–177.
54. Qin, Z.S., Niu, T. and Liu, J.S. (2002), 'Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 71, pp. 1242–1247.
55. Li, S.S., Khalid, N., Carlson, C. and Zhao, L.P. (2003), 'Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms', *Biostatistics* Vol. 4, pp. 513–522.
56. Barrett, J. and Daly, M.J., (2004), HAPLOVIEW ver. 2.04 documentation, available at http://www.broad.mit.edu/personal/jcbarret/haplo/.
57. Niu, T., Qin, Z.S., Xu, X. and Liu, J.S. (2002), 'Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 70, pp. 1242–1247.
58. Lin, S., Cutler, D.J., Zwick, M.E. and Chakravarti, A. (2002), 'Haplotype inference in random population samples', *Am. J. Hum. Genet.* Vol. 71, pp. 1129–1137.
59. Stephens, M. and Donnelly, P. (2003), 'A comparison of Bayesian methods for haplotype reconstruction from population genotype data', *Am. J. Hum. Genet.* Vol. 73, pp. 1162–1169.
60. Excoffier, L., Laval, G. and Balding, D. (2003), 'Gametic phase estimation over large genomic regions using an adaptive window approach', *Hum. Genomics* Vol. 1, pp. 7–19.
61. Lin, S., Chakravarti, A. and Cutler, D.J. (2004), 'Haplotyping and missing data Inference in nuclear families', *Genome Res.* Vol. 14, pp. 1624–1632.
62. Eronen, L., Geerts, F. and Toivonen, H. (2004), 'A Markov chain approach to reconstruction of long haplotypes', *Pac. Symp. Biocomput*, available at http://www.cs.helsinki.fi/group/genetics/haplotyping.html/.
63. Stephens, M., Smith, N.J., Donnelly, P. *et al.* (2001), 'Reply to Zhang *et al.*', *Am. J. Hum. Genet.* Vol. 69, pp. 912–914.
64. Single, R.M., Meyer, D., Hollenbach, J.A. *et al.* (2002), 'Haplotype frequency estimation in patient populations: The effect of departures from Hardy–Weinberg proportions and collapsing over a locus in the HLA region', *Genet. Epidemiol.* Vol. 22, pp. 186–195.
65. Daly, M.J., Rioux, J.D., Schaffner, S.F. *et al.* (2001), 'High-resolution haplotype structure in the human genome', *Nat. Genet.* Vol. 29, pp. 229–232.
66. Goldstein, D.B., Ahmadi, K.R., Weale, M.E. *et al.* (2003), 'Genome scans and candidate gene approaches in the study of common diseases and variable drug responses', *Trends Genet.* Vol. 19, pp. 615–622.
67. Fearnhead, P. and Donnelly, P. (2001), 'Estimating recombination rates from population genetic data', *Genetics* Vol. 159, pp. 1299–1318.
68. Zou, G. and Zhao, H. (2003), 'Haplotype frequency estimation in the presence of genotyping errors', *Hum. Hered.* Vol. 56, pp. 131–138.
69. Akey, J.M., Zhang, K., Xiong, M. and Jin, L. (2003), 'The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium', *Mol. Biol. Evol.* Vol. 20, pp. 232–242.
70. Gordon, D., Finch, S.J., Nothnagel, M. and Ott, J. (2002), 'Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms', *Hum. Hered.* Vol. 54, pp. 22–33.
71. Kang, S.J., Gordon, D. and Finch, S.J. (2004), 'What SNP genotyping errors are most costly for genetic association studies?', *Genet. Epidemiol.* Vol. 26, pp. 132–141.

72. Zou, G. and Zhao, H. (2004), 'The impacts of errors in individual genotyping and DNA pooling on association studies', *Genet. Epidemiol.* Vol. 26, pp. 1–10.
73. Kang, H., Qin, Z.S., Niu, T. and Liu, J.S. (2004), 'Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 74, pp. 495–510.
74. Kirk, K.M. and Cardon, L.R. (2002), 'The impact of genotyping error on haplotype reconstruction and frequency estimation', *Eur. J. Hum. Genet.* Vol. 10, pp. 616–622.
75. Ewen, K.R., Bahlo, M., Treloar, S.A. *et al.* (2000), 'Identification and analysis of error types in high-throughput genotyping', *Am. J. Hum. Genet.* Vol. 67, pp. 727–736.
76. Rice, K.M. and Holmans, P. (2003), 'Allowing for genotyping error in analysis of unmatched case-control studies', *Ann. Hum. Genet.* Vol. 67, pp. 165–174.
77. Sobel, E., Papp, J.C. and Lange, K. (2002), 'Detection and integration of genotyping errors in statistical genetics', *Am. J. Hum. Genet.* Vol. 70, pp. 496–508.
78. Douglas, J.A., Skol, A.D. and Boehnke, M. (2002), 'Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data', *Am. J. Hum. Genet.* Vol. 70, pp. 487–495.
79. Hosking, L., Lumsden, S., Lewis, K. *et al.* (2004), 'Detection of genotyping errors by Hardy-Weinberg equilibrium testing', *Eur. J. Hum. Genet.* Vol. 12, pp. 395–399.
80. Gourraud, P.A., Genin, E. and Cambon-Thomsen, A. (2004), 'Handling missing values in population data: Consequences for maximum likelihood estimation of haplotype frequencies', *Eur. J. Hum. Genet.* Vol. 12, pp. 805–812.
81. Chiano, M.N. and Clayton, D.G. (1998), 'Fine genetic mapping using haplotype analysis and the missing data problem', *Ann. Hum. Genet.* Vol. 62, pp. 55–60.
82. Zaykin, D.V., Westfall, P.H., Young, S.S. *et al.* (2002), 'Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals', *Hum. Hered.* Vol. 53, pp. 79–91.
83. Zhao, L.P., Li, S.S. and Khalid, N. (2003), 'A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies', *Am. J. Hum. Genet.* Vol. 72, pp. 1231–1250.
84. Zhao, J.H., Curtis, D. and Sham, P.C. (2000), 'Model-free analysis and permutation tests for allelic associations', *Hum. Hered.* Vol. 50, pp. 133–139.
85. Xie, X. and Ott, J. (1993), 'Testing linkage disequilibrium between a disease and marker locus', *Am. J. Hum. Genet.* Vol. 53, pp. 1107.
86. Schaid, D.J., Rowland, C.M., Tines, D.E. *et al.* (2002), 'Score tests for association between traits and haplotypes when linkage phase is ambiguous', *Am. J. Hum. Genet.* Vol. 70, pp. 425–434.
87. Lake, S.L., Lyon, H., Tantisira, K. *et al.* (2003), 'Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous', *Hum. Hered.* Vol. 55, pp. 56–65.
88. Tregouet, D.A. and Tiret, L. (2004), 'Cox proportional hazards survival regression in haplotype-based association analysis using the stochastic-EM algorithm', *Eur. J. Hum. Genet.* Vol. 12, pp. 971–974.
89. Schneidler, S., Roessli, D. and Excoffier, L. (2002), 'Arlequin version 2.001: A software for population genetics data analysis', Genetics and Biometry Laboratory, University of Geneva, Switzerland.
90. Chiano, M.N. and Clayton, D.G. (1998), 'Genotypic relative risks under ordered restriction', *Genet. Epidemiol.* Vol. 15, pp. 135–146.
91. Epstein, M.P. and Satten, G.A. (2003), 'Inference on haplotype effects in case-control studies using unphased genotype data', *Am. J. Hum. Genet.* Vol. 73, pp. 1316–1329.
92. Satten, G.A. and Epstein, M.P., (2004), 'Comparison of prospective and retrospective methods for haplotype inference in case-control studies', *Genet. Epidemiol.* Vol. 27, pp. 192–201.
93. Lin, D.Y. (2004), 'Haplotype-based association analysis in cohort studies of unrelated individuals', *Genet. Epidemiol.* Vol. 26, pp. 255–264.
94. Stram, D.O., Leigh Pearce, C., Bretsky, P. *et al.* (2003), 'Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals', *Hum. Hered.* Vol. 55, pp. 179–190.
95. Sham, P., Bader, J.S., Craig, I. *et al.* (2002), 'DNA pooling: A tool for large-scale association studies', *Nat. Rev. Genet.* Vol. 3, pp. 862–871.
96. Ito, T., Chiku, S., Inoue, E. *et al.* (2003), 'Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data', *Am. J. Hum. Genet.* Vol. 72, pp. 384–398.
97. Inbar, E., Yakir, B. and Darvasi, A. (2002), 'An efficient haplotyping method with DNA pools', *Nucleic Acids Res.* Vol. 30, p. e76.
98. Yang, Y., Zhang, J., Hoh, J., Matsuda, F. *et al.* (2003), 'Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 7225–7230.
99. Wang, S., Kidd, K.K. and Zhao, H. (2003), 'On the use of DNA pooling to estimate haplotype frequencies', *Genet. Epidemiol.* Vol. 24, pp. 74–82.
100. Clayton, E.W. (2002), 'The complex relationship of genetics, groups, and health: What it means for public health', *J. Law Med. Ethics* Vol. 30, pp. 290–297.
101. Zhao, H., Pfeiffer, R. and Gail, M.H. (2003), 'Haplotype analysis in population genetics and association studies', *Pharmacogenomics* Vol. 4, pp. 171–178.
102. Ziv, E. and Burchard, E.G. (2003), 'Human population structure and genetic association studies', *Pharmacogenomics* Vol. 4, pp. 431–441.
103. Deng, H.W. (2001), 'Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits', *Genetics* Vol. 159, pp. 1319–1323.
104. Terwilliger, J. and Ott, J. (1994), *Handbook for Human Genetic Linkage*, Johns Hopkins University Press, Baltimore, MD.
105. Zhao, J.H. and Sham, P.C. (2002), 'Faster haplotype frequency estimation using unrelated subjects', *Hum. Hered.* Vol. 53, pp. 36–41.
106. Zhao, J.H., Lissarrague, S., Essioux, L. and Sham, P.C. (2002), 'GENECOUNTING: Haplotype analysis with missing genotypes', *Bioinformatics* Vol. 18, pp. 1694–1695.
107. Thomas, A. (2003), 'GCHap: Fast MLEs for haplotype frequencies by gene counting', *Bioinformatics* Vol. 19, pp. 2002–2003.
108. Thomas, A. (2003), 'Accelerated gene counting for haplotype frequency estimation', *Ann. Hum. Genet.* Vol. 67, pp. 608–612.
109. Krawczak, M. (1994), 'HAPMAX documentation', available at http://www.uni-kiel.de/medinfo/mitarbeiter/krawczak/download/hapmax.txt.
110. Zhang, J., Rowe, W.L., Struewing, J.P. and Buetow, K.H. (2002), 'HapScope: A software system for automated and visual analysis of functionally annotated haplotypes', *Nucleic Acids Res.* Vol. 30, pp. 5213–5221.
111. Wang, X. (2003), 'HIT: A haplotype inference testbed', Department of Electrical and Computer Engineering — CAPSL, University of Deleware, 17th March.
112. Higashi, Y., Higuchi, H., Kido, T. *et al.* (2003), 'SNP analysis system for detecting complex disease associated sites', paper presented at the IEEE Computer Society Bioinformatics Conference (CSB 2003), Stanford, CA, available at http://csdl.computer.org/comp/proceedings/csb/2003/2000/00/2000toc.htm/.
113. Long, J.C. (1999), 'Multiple locus haplotype analysis, version 3.0', Software and documentation distributed by the author; section on Population Genetics and Linkage, Laboratory of Neurogenetics, NIAAA, National Institutes of Health, Bethesda, MD.
114. Zhang, P., Sheng, H., Morabia, A. and Gilliam, T.C. (2003), 'Optimal step length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping', *BMC Bioinformatics* Vol. 4, p. 3.
115. Czika, W., Yu, X. and Wolfinger, R.D. (2003), 'An introduction to genetic data analysis using SAS/genetics', available at http://support.sas.com/rnd/papers/sugi27/genetics.pdf.
116. Purcell, S. and Sham, P. (2003), 'WHAP: SNP haplotype analysis package, ver 2.04', available at http://www.genome.wi.mit.edu/~shaun/whap.
117. Hoh, J., Matsuda, F., Peng, X. *et al.* (2003), 'SNP haplotype tagging from DNA pools of two individuals', *BMC Bioinformatics* Vol. 4, p. 14.