Genetic Epidemiology 31: 659–671 (2007)

Understanding the Accuracy of Statistical Haplotype Inference with Sequence Data of Known Phase

Aida M. Andrés,[1]* Andrew G. Clark,[1] Lawrence Shimmin,[2] Eric Boerwinkle,[2] Charles F. Sing,[3] and James E. Hixson[2]

[1]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York*
[2]*Human Genetics Center, University of Texas Health Science Center, Houston, Texas*
[3]*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan*

Statistical methods for haplotype inference from multi-site genotypes of unrelated individuals have important application in association studies and population genetics. Understanding the factors that affect the accuracy of this inference is important, but their assessment has been restricted by the limited availability of biological data with known phase. We created hybrid cell lines monosomic for human chromosome 19 and produced single-chromosome complete sequences of a 48 kb genomic region in 39 individuals of African American (AA) and European American (EA) origin. We employ these phase-known genotypes and coalescent simulations to assess the accuracy of statistical haplotype reconstruction by several algorithms. Accuracy of phase inference was considerably low in our biological data even for regions as short as 25–50 kb, suggesting that caution is needed when analyzing reconstructed haplotypes. Moreover, the reliability of estimated confidence in phase inference is not high enough to allow for a reliable incorporation of site-specific uncertainty information in subsequent analyses. We show that, in samples of certain mixed ancestry (AA and EA populations), the most accurate haplotypes are probably obtained when increasing sample size by considering the largest, pooled sample, despite the hypothetical problems associated with pooling across those heterogeneous samples. Strategies to improve confidence in reconstructed haplotypes, and realistic alternatives to the analysis of inferred haplotypes, are discussed. *Genet. Epidemiol.* 31:659–671, 2007.

Key words: Kallekrein; KLK; haplotype reconstruction; phase; LD

Contract grant sponsor: NIH; Contract grant number: GM65509.
Aida M. Andrés present address is Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.
*Correspondence to: Dr. Aida M. Andrés, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50 Room 5527, Bethesda, MD 20892. E-mail: andresa@mail.nih.gov
Received 7 May 2006; Revised 8 August 2006; Accepted 25 September 2006
Published online 5 October 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20185

INTRODUCTION

There are compelling biological and statistical reasons to examine polymorphism at the level of groups of linked SNPs whose collective states comprise a haplotype. In studies that seek evolutionary inferences from DNA polymorphism data, extending the analysis from the polymorphic site to the haplotype may reveal the latent genealogical history of mutation and recombination events and present a richer picture of the structure of genetic variation. Haplotypes have thus become a common component of population genetic studies to understand mutational, recombinational, selective, and demographic dynamics of genomic regions [Tishkoff et al., 1996; Clark et al., 1998; Mateu et al., 2001; Sabeti et al., 2002; Kidd KK et al., 2004]. Moreover, haplotype data may be valuable for association studies, where the hope is that information captured by haplotypes will facilitate the recognition of genetic bases of phenotypic traits, including common diseases [Johnson et al., 2001; Clark, 2004].

Standard methods of surveying DNA variation are usually based on genomic amplification by PCR and provide multi-site genotypic information with unknown gametic phase. Unambiguous haplotypes can be determined experimentally by a number of methods, including single-cell amplification [Stephens et al., 1990], cloning of a

genomic region or PCR product [Martinez-Arias et al., 2001], allele-specific amplification [Clark et al., 1998], or somatic hybrid cell construction [Yan et al., 2000]. All these approaches are complicated and tedious, and are not suitable for large-scale studies of diversity. The one exception are hydatidiform moles, which can provide genome-wide haplotype information [Kukita et al., 2005], but this technique is limited to the exceptional samples presented by the moles themselves. Alternatively, haplotypes can be inferred from family data, and the use of father-mother-offspring trios is probably the most efficient strategy [Myers et al., 2005; The International HapMap Consortium, 2005; Marchini et al., 2006]. Family-based approaches, however, dramatically increase the number of assays and the cost of the study.

Statistical inference of haplotype phase from population data was first formally developed for pairs of loci by Hill [1974] using an EM algorithm. Clark [1990] recognized that multiple-site haplotypes could be inferred from unphased population samples using an ad hoc parsimony algorithm. This inference is possible primarily due to two attributes of genetic data: linkage disequilibrium (LD) and the surprisingly simple structured composition of haplotypes in populations as a result of the coalescent properties of allelic lineages [Patil et al., 2001; Gabriel et al., 2002]. Other algorithms were subsequently developed to infer (by maximum likelihood, ML) relative population haplotype frequencies based on genotype frequencies [Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995].

Recently a new generation of methods has been developed based on similar principles and implemented as ML [Kimmel and Shamir, 2005], Bayesian [Stephens et al., 2001; Lin et al., 2002; Niu et al., 2002; Excoffier et al., 2003], or Perfect and Imperfect Phylogeny algorithms [Chung and Gusfield, 2003; Halperin and Eskin, 2004], among others [see reviews from Niu et al., 2004; Salem et al., 2005]. As applicability and accuracy of the methods have improved and experience in their application has accumulated, statistical inference of haplotypes has become the strategy of choice to obtain haplotype information from unrelated diploid individuals.

Although haplotype reconstruction may have important practical application, thorough testing of its accuracy has been limited by a scarcity of data with known phase. This limitation has been partially overcome by the use of coalescent simulations, where haplotype reconstruction performs well [Stephens et al., 2001; Niu et al., 2002]. Nevertheless, the inference can be a more challenging problem in biological data, as shown by results on the limited available data with known phase. For example, a recent study by Stephens and Scheet [2005] compared several phasing methods on eight X-linked loci and showed large error rates for all algorithms. Kukita et al. [2005] showed that one particular measure of error [the *switch error*] can vary substantially among genomic regions. The HapMap data showed a high accuracy of several methods when family trios are used [Marchini et al., 2006], but the accuracy of reconstruction was substantially lower in unrelated individuals (as assessed using family-reconstructed haplotypes). These and additional studies have helped measure and compare performance of different methods of haplotype reconstruction. But they are based on long genomic regions with sparse SNP coverage and, in many cases, small sample sizes, all factors expected to negatively affect phasing accuracy. Little is known about the accuracy of the inference for specific genomic regions with large number of SNPs of diverse frequencies, in samples of individuals from several populations. That is, for most candidate-gene studies.

To help assess the accuracy of phasing methods and further understand long-range haplotype patterns in humans, we produced a dataset of experimentally determined haplotype sequences of a 48 kb genomic region in 39 unrelated control individuals from two populations. The experimental design resembles most candidate-gene epidemiological studies: a single locus extensively screened for variants in several populations. Coalescent simulations were also used to assess phasing performance, in an attempt to better understand the factors primarily affecting the accuracy of statistical haplotype inference.

MATERIALS AND METHODS

SUBJECTS AND SEQUENCES

The sample for our study included 20 African Americans (AA) from Atlanta, Georgia, and 19 European Americans (EA) from Rochester, Minnesota, who provided written informed consent. In order to obtain haploid sequences, we produced hybrid somatic cell lines, each containing a single human chromosome 19. Briefly, the hybrid cell lines were generated by electrofusion

of a mouse embryonic fibroblast host cell line with human lymphoblastoid cell lines. The fusion cell products randomly lose human chromosomes, yielding hybrid cell lines with a human X chromosome and random retention of 0, 1, or 2 copies of other chromosomes [Yan et al., 2000; Douglas et al., 2001]. For every individual, cell lines monosomic for each homolog of chromosome 19 were selected by genotyping one short tandem repeat (STR) on both arms of the chromosome, known to be heterozygous by previous genotyping of the original lymphoblastoid sample [Shimmin et al., in preparation].

We obtained haploid sequences of known phase for a 48 kb region of chromosome 19 containing two genes of the Kallikrein family: Kallikrein 14 (KLK14 [OMIM *606135]) and Kallikrein 13 (KLK13 [OMIM *605505]). This region was chosen as part of a survey of long-range sequence variation on chromosome 19 that includes Kallikrein family members involved in physiological pathways related to cancer and cardiovascular disease. The region was amplified by PCR of overlapping fragments (1–3 kb) in the hybrid cell lines, and haploid sequences were obtained by direct nucleotide sequencing of both strands with internal primers (spaced 400–500 bp) using an ABI Prism 3730XL DNA Analyzer. DNA variants were identified by assembly and alignment of these sequences. We recorded STRs, SNPs, and insertion/deletion polymorphisms (indels), but only diallelic segregating sites (referred as SS throughout this paper) were considered for further analysis.

PHASING

Several overlapping datasets were considered: the whole region (KLK) and the two genes separately (regions KLK14 and KLK13), each considering all individuals together or splitting the sample by population of origin (AA and EA). The original dataset (''all SS'') contains all diallelic SS found in our sample. To assess the influence of low-frequency alleles in phasing, rare SS were removed in the ''common SS'' datasets (frequency in the sample <0.1). To evaluate the performance of phasing tagSNPs, the pairwise method was used to select tagSNPs from the common SS datasets (equivalent to ldSelect [Carlson et al., 2004]) with a threshold of 0.8, as implemented in Haploview [Barrett et al., 2005]. Table I summarizes the characteristics of each dataset.

Each dataset was analyzed as follows. The two known haplotypes from each individual were paired and phase information was erased by shuffling the two haplotypes of every individual, producing a dataset that mimics unphased genotypes. Phase was inferred using fastPHASE 1.1 [Scheet and Stephens, 2006], GERBIL 1.1 [Kimmel and Shamir, 2005], HAP 3.0 [Halperin and Eskin, 2004], and PHASE 2.1 [Stephens and Donnelly, 2003]. Haplotyper 1.0 [Niu et al., 2002] was not considered as it failed to produce results on our datasets. HAP 3.0 was run via the web server, the only access to the software available at present. For fastPHASE 1.1, GERBIL 1.1, and PHASE 2.1, we report the results of running programs at their default parameters, which were considered adequate for the authors to test performance [Kimmel and Shamir, 2005; Stephens and Scheet, 2005; Scheet and Stephens, 2006]. PHASE 2.1 was run three times with different randomization seeds for each dataset. To ensure that accuracy was not limited by running time, longer runs were allowed, with no improvement in performance observed (data not shown). For fastPHASE, the output that tries to maximize switch accuracy (switch.out) was chosen due to its general higher accuracy.

TABLE I. Description of real data datasets of the study

	Length(kb)	N	allSS	comSS	tagSS
KLK	48	39	401	197	110
KLK_AA	48	20	370	205	122
KLK_EA	48	19	233	169	72
KLK14	25	39	181	87	47
KLK14_AA	25	20	169	88	51
KLK14_EA	25	19	96	65	26
KLK13	23	39	220	110	62
KLK13_AA	23	20	201	117	69
KLK13_EA	23	19	137	104	45

N: number of individuals in the sample; allSS: number of SS in the complete dataset; comSS: number of SS in the common SS dataset (frequency $>=0.1$); tagSS: number of SS in the tag SS dataset.
KLK is the complete region containing both KLK14 and KLK13 genes. The complete datasets (KLK, KLK14, and KLK13) have both AA and EA individuals; _AA: datasets have only African-American individuals; and _EA: only European-Americans.

SAMPLE SIZE

A series of subsamples were created from KLK, KLK14, and KLK13 datasets to test for the poorly understood influence of sample size and population structure in phasing. For datasets from just one population (_AA and _EA datasets), subsamples contained 5–15 individuals from AA or EA origin, and the ''Mixed ancestry'' subsamples had

sizes ranging from 5 to 35, with half AA and half EA individuals (subsamples of uneven size contain, randomly, one additional AA or EA individual). Fifty subsamples were generated for each condition (each size and population combination), and haplotypes were reconstructed using PHASE 2.1 with default parameters (number of interactions = 100, thinning interval = 1, burn-in = 100, see PHASE 2.1 documentation for details), given the general better performance of this method in our biological dataset.

COALESCENT SIMULATIONS

We used coalescent simulations to assess the influence of different genetic factors on the performance of haplotype reconstruction. Datasets of 100 haplotypes were generated with MS [Hudson, 2002], a program that simulates sample data under the neutral coalescent with a given neutral mutation rate, effective size, recombination rate, and specified demographic history. We attempted to employ parameters that yielded datasets comparable in size and complexity to the KLK data. Two different demographic scenarios were simulated: demographic equilibrium and the non-equilibrium demographic best fit from Marth et al. [2004]. For AA, we assumed an increase in population size for 7,500 generations, from $N = 10{,}000$ to 18,000, and for EA, we assumed a split from the original African population 3,500 generations ago, coincident with a reduction in size from $N = 10{,}000$ to 2,000, followed by a constant population size of 2,000 for 500 generations, and finally an increase in size for 3,000 generations, from $N = 2{,}000$ to 20,000.

In both cases, we simulated three sequence lengths ($L = 12.5$, 25, and 50 kb) and four recombination rates ($\rho = 0$, 1.1, 3, 6 for equilibrium; $\rho = 0$, 0.5, 1.5, 3 for demography). ρ was scaled for the equilibrium and demographic models to avoid an extreme effect of recombination in the demographic model, given the differences in effective population size (N_e) between the two methods (N_e is larger for the demographic model than for the equilibrium case for most of the simulated time). A difference between biological and simulated data is the presence (in the former) of missing data, in this case due to a failure to obtain high-quality sequence in some bases of the region. To mimic experimentally obtained data, missing data were randomly introduced with the same amount of missing SS per haplotype as the "per haplotype" average in the KLK data. A filter for allele frequency was also applied to produce data sets containing only common alleles (frequency $> = 0.1$).

The parameter conditions of the simulated data can be summarized as: Demographic model: equilibrium, bottleneck. Length of the region: 12.5 kb (100 SS), 25 kb (200 SS), 50 kb (400 SS). Recombination: $\rho = 0$, 1.1, 3, 6 (equilibrium) and $\rho = 0$, 0.5, 1.5, 3 (demography). Segregating sites: "all SS", "common SS". Missing data: absent, present.

For every parameter combination, 50 independent samples were obtained. Haplotypes were paired, shuffled, and phased with PHASE 2.1 under default parameters in all cases except for the longest sequences (400 SS, 50 kb), run for shorter runs (burn-in time of 50 instead of 100, see PHASE 2.1 documentations for details). These running parameters were enough to stabilize the algorithm for a dataset of these characteristics, as shown from the biological data. In simulated data, PHASE 2.1 was run just once per dataset.

PERFORMANCE METRICS

Performance of all methods was evaluated by several metrics that summarize diverse attributes of the accuracy of the process:

- The *Haplotype error rate (HE)*: average proportion of haplotypes incorrectly inferred (percentage of reconstructed haplotypes with, at least, one site erroneously assigned).
- The *single-site error rate (SSE)*: average proportion of ambiguous SS (that is, heterozygote SS in the individual) whose phase is incorrectly inferred.
- The *global single-site error rate (gSSE)*: average proportion of all SS whose phase is incorrectly inferred. Note that the denominator here is the total number of sites, regardless of them being ambiguous or not.
- The *Switch error (SwE)*, as defined by Stephens and Donnelly [2003], corresponds to 1 minus *Switch accuracy* in Lin et al. [2002]: average proportion of heterozygous positions misassigned relative to the previous heterozygous position. It shows whether errors in haplotype reconstruction are mainly due to the misassignment of isolated SS (high error), or of blocks of neighboring SS (low error).

Performance was computed for each real-data dataset. When a program was run several times,

the best result (lowest error) is reported. Performance was also computed for every sample size pseudo-dataset and every simulated dataset; we report the average performance among the 50 datasets simulated under identical conditions. In all cases, only non-missing SSs in a given individual are tested for correctness in that individual.

RESULTS

The comparison of haploid sequences in monosomic hybrid cell lines from 39 individuals of AA and EA origin showed 411 polymorphisms in the targeted 48 kb region of chromosome 19. These polymorphisms include nine repetitive regions (STRs), 55 indels, 346 diallelic SNPs, and one triallelic SNP. Only the 401 diallelic SS were considered for further analyses (indels and diallelic SNPs). The sequence is accessible under accession number EU091477, SNPs in dbSNPs, and haplotypes in http://superc.hg.med.umich.edu/micortex/publications.

HAPLOTYPE RECONSTRUCTION OF BIOLOGICAL DATA

The objective of this study is to assess the general accuracy of haplotype reconstruction in real datasets rather than to compare all available software. We focus on most widely used methods or most accurate algorithms according to available assessments. Readers interested in a more exhaustive comparison of programs, on a (X-linked) smaller dataset, are referred to Stephens and Scheet [2005].

The complete matrix of methods and error measures is presented in Table II, but the most general error measures, HE and SSE, are shown in Figure 1A for the different algorithms. PHASE generally shows the lowest error rates, although the advantages of PHASE are not absolute, as some cases find HAP or, more often, fastPHASE and GERBIL reconstructions to be more accurate.

When all SS are considered (Fig. 1A and panel A of Table I)], the ***haplotype error rate*** (***HE***) is high for all methods: close to 1 for the longest region (KLK) and between ~0.6 and ~0.9 for the shorter regions (KLK13 and KLK14). However, this is a very stringent error measure since the mis-call of a single site makes an incorrect haplotype. The actual proportion of sites miss-assigned, the ***single-site error rate*** (***SSE***), ranges from 0.086 to 0.367. The low value of ***switch error rate*** (***SwE***) indicates that errors tend to be clustered, and haplotypes are composed by 'blocks' of adjacent SS correctly assigned to the same haplotype but incorrectly combined with the rest of 'blocks' (switch blocks). Still, the average number of SwE events per haplotype are high enough that such ''switch blocks'' are on average small compared with the total number of SSs, and inferred haplotypes consist of many, short, switch blocks (results not shown).

An important factor limiting the accuracy of haplotype inference is the presence of rare SS. Indeed, their removal in the *common SS* dataset (frequency >= 0.1) increases accuracy based on the various error metrics (dotted lines in Figs. 1B–E, and panel B in Table II). These results suggest that haplotype reconstruction of common SS genotypes (like those genotyped from previously ascertained SNPs) will be more accurate than that of datasets containing rare SS (e.g. resequencing studies).

A common strategy in association testing to maximize information while reducing redundancy is the selection, among all SNPs in the sample, of tagSNPs that minimize the pairs of SNPs in high LD. The effects of this criterion of site selection are shown in Figures 1B–E, dashed lines, and in Table II, panel C. In general, HE is similar or lower for tag SS than for common SS datasets (consistent with the reduction in number of SS), but SSE is in fact slightly higher. Furthermore, SwE increases as a consequence of tag SNP selection compared with common SS dataset (Fig. 1F). This suggests that site errors will be more frequent and less clustered in tagSNPs than in non-selected SNPs. By specifically selecting SNPs in lower LD, it is not surprising that the confidence in phase of the remaining low-LD SNPs would be reduced.

It is important to point out that one of the most evident differences between methods is their relative speed. For example, for a short region (KLK13), considering only common SS, HAP webserver returned results within few minutes, fastPHASE ran for 349 s, GERBIL ran for 491 s, and PHASE ran for 1,018 s in an Intel Pentium 4 CPU 2.40 GHz Linux computer. As the region grows in SS, the speed differences escalate.

SAMPLE SIZE AND POPULATION STRATIFICATION

Results above show that accuracy of haplotype inference is not negatively influenced by pooling AA

TABLE II. Accuracy of haplotype reconstruction algorithms on real data datasets

		panel A all SS				panel B common SS				panel C tag SS			
Data	Method	HE	SSE	gSSE	SwE	HE	SSE	gSSE	SwE	HE	SSE	gSSE	SwE
KLK													
	fastPHASE	0.923	0.294	0.059	0.076	0.949	0.273	0.093	0.057	0.949	0.213	0.070	0.092
	GERBIL	1	0.299	0.060	0.096	0.974	0.236	0.081	0.076	1	0.274	0.089	0.147
	HAP					0.949	0.266	0.090	0.079	0.949	0.305	0.107	0.128
	PHASE	0.974	0.264	0.053	0.102	0.897	0.203	0.067	0.068	0.846	0.228	0.078	0.093
KLK_AA													
	fastPHASE	1	0.360	0.083	0.119	1	0.259	0.095	0.066	1	0.289	0.106	0.118
	GERBIL	1	0.317	0.071	0.108	1	0.284	0.104	0.087	1	0.318	0.118	0.150
	HAP					1	0.340	0.129	0.150	1	0.392	0.146	0.265
	PHASE	1	0.320	0.074	0.131	1	0.192	0.073	0.072	0.950	0.214	0.079	0.101
KLK_EA													
	fastPHASE	1	0.301	0.050	0.070	0.947	0.283	0.106	0.055	0.947	0.272	0.100	0.113
	GERBIL	1	0.283	0.047	0.087	0.947	0.220	0.092	0.061	0.895	0.235	0.089	0.147
	HAP					0.947	0.274	0.105	0.086	1	0.254	0.098	0.229
	PHASE	0.947	0.249	0.041	0.098	0.947	0.245	0.086	0.063	0.947	0.305	0.111	0.166
KLK13													
	fastPHASE	0.897	0.185	0.039	0.079	0.821	0.229	0.090	0.063	0.769	0.211	0.072	0.138
	GERBIL	0.872	0.220	0.048	0.092	0.795	0.188	0.074	0.072	0.923	0.249	0.083	0.179
	HAP	0.923	0.302	0.067	0.132	0.846	0.239	0.092	0.078	0.718	0.205	0.079	0.107
	PHASE	0.872	0.202	0.041	0.097	0.744	0.154	0.058	0.060	0.795	0.179	0.053	0.108
KLK13_AA													
	fastPHASE	1	0.276	0.063	0.127	0.850	0.270	0.106	0.081	0.750	0.233	0.087	0.134
	GERBIL	0.900	0.276	0.064	0.091	0.850	0.212	0.080	0.078	0.850	0.224	0.084	0.123
	HAP	1	0.367	0.087	0.197	1	0.286	0.114	0.131	1	0.330	0.132	0.233
	PHASE	1	0.228	0.051	0.115	0.950	0.216	0.079	0.097	0.750	0.189	0.070	0.126
KLK13_EA													
	fastPHASE	0.895	0.254	0.049	0.088	0.632	0.156	0.061	0.040	0.737	0.167	0.055	0.129
	GERBIL	0.684	0.173	0.034	0.074	0.737	0.181	0.074	0.064	0.737	0.211	0.075	0.139
	HAP	0.842	0.256	0.049	0.105	0.789	0.174	0.071	0.056	0.737	0.200	0.071	0.157
	PHASE	0.842	0.149	0.024	0.075	0.684	0.190	0.067	0.055	0.579	0.162	0.054	0.107
KLK14													
	fastPHASE	0.769	0.129	0.024	0.092	0.769	0.120	0.035	0.084	0.641	0.119	0.042	0.113
	GERBIL	0.923	0.170	0.032	0.109	0.821	0.129	0.044	0.087	0.667	0.160	0.062	0.143
	HAP	0.949	0.260	0.048	0.186	0.744	0.156	0.051	0.077	0.744	0.168	0.056	0.122
	PHASE	0.641	0.102	0.016	0.107	0.564	0.089	0.026	0.061	0.667	0.093	0.028	0.096
KLK14_AA													
	fastPHASE	0.900	0.145	0.029	0.147	0.750	0.143	0.050	0.069	0.800	0.169	0.054	0.119
	GERBIL	0.950	0.277	0.058	0.143	0.950	0.264	0.097	0.122	0.900	0.221	0.086	0.128
	HAP	1	0.312	0.066	0.256	0.950	0.240	0.091	0.167	1	0.282	0.104	0.304
	PHASE	0.750	0.106	0.022	0.120	0.700	0.087	0.031	0.065	0.650	0.125	0.042	0.101
KLK14_EA													
	fastPHASE	0.737	0.083	0.014	0.051	0.737	0.113	0.036	0.082	0.789	0.167	0.069	0.178
	GERBIL	0.789	0.153	0.024	0.071	0.684	0.082	0.030	0.046	0.632	0.146	0.062	0.101
	HAP	0.895	0.227	0.035	0.180	0.737	0.189	0.065	0.114	0.895	0.216	0.085	0.190
	PHASE	0.789	0.167	0.028	0.089	0.632	0.108	0.047	0.040	0.474	0.086	0.042	0.071

HE: Haplotype error rate; SSE: SS error rate; gSSE: global SS error rate; SwE: Switch error rate (for a full explanation of error rates, see Materials and Methods section). Empty cells correspond to failures of the software to correctly run and return results on that dataset.

and EA samples. The pooling combines disparate population samples but increases sample size, and the effects of these two factors may be acting in different directions and canceling one another. The effects of sample size are slight once samples reach intermediate sizes (Fig. 2A), but the influence of population structure is more dramatic, with haplotypes in AA samples being harder to infer than in EA samples. Interestingly, accuracy in the combined sample (AA plus EA) is similar or slightly poorer than for the EA sample but clearly better than for the AA sample (Fig. 2A). Therefore, in this kind of structured sample, with AA and EA individuals and a relatively small sample size, the best global phasing strategy is probably to pool all individuals as a single sample, regardless of their origin.

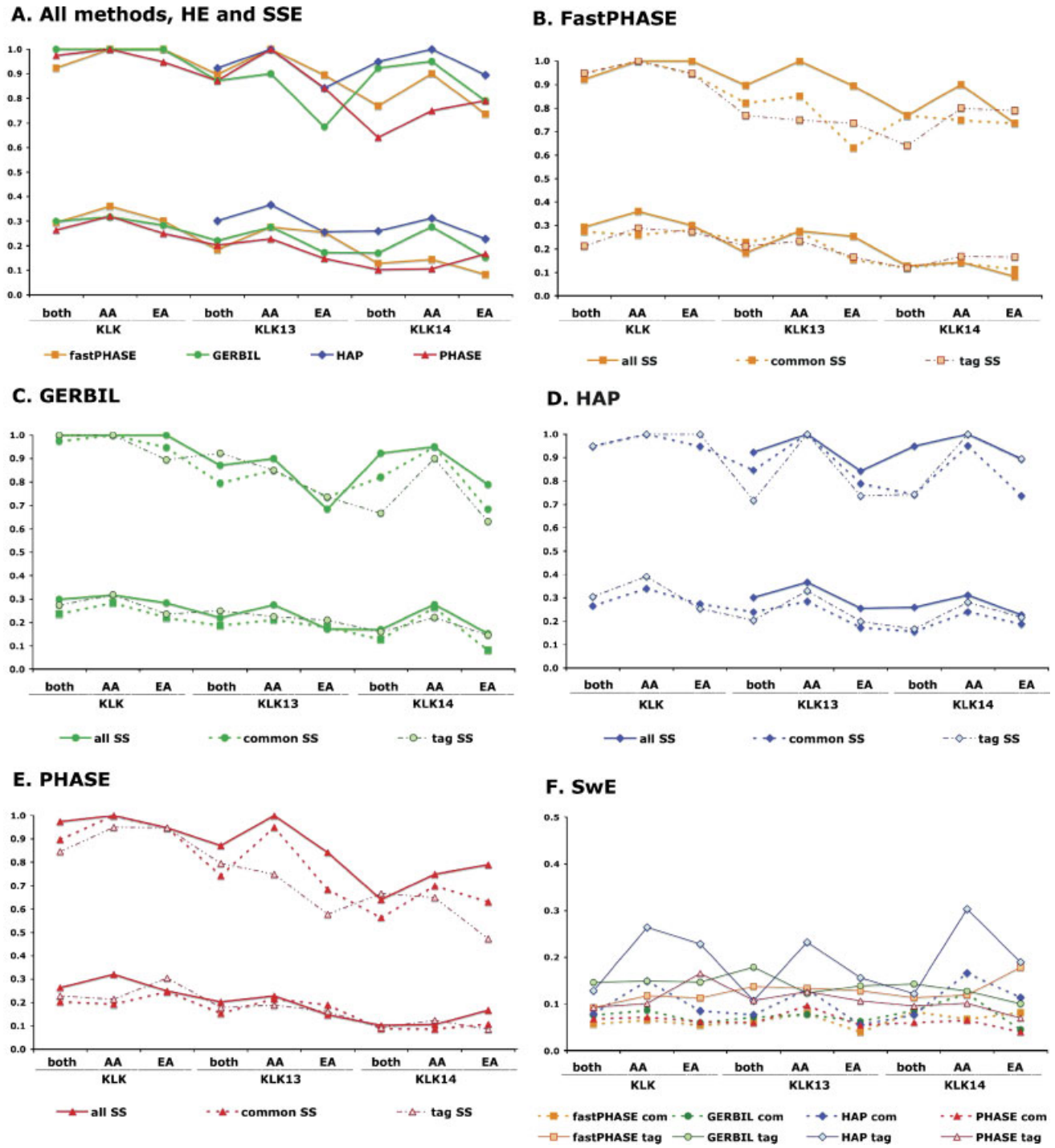

Fig. 1. (A) Comparison of the Haplotype error rate (upper) and SS error rate (lower) of datasets containing all SS, for the four programs. *X*-axis: dataset; *Y*-axis: error rate. Similar plots are obtained when considering common SS or tag SS datasets. (B, C, D, E) Haplotype error rate (upper lines) and SS error rate (lower lines) of datasets containing all SS, common SS, or tag SS, for the three programs (fastPHASE [B], GERBIL [C], HAP [D], PHASE [E]). Axes as in (A). (F) Switch Error Rate of datasets containing common SS or tagSS, for the four different programs. Axes as in (A).

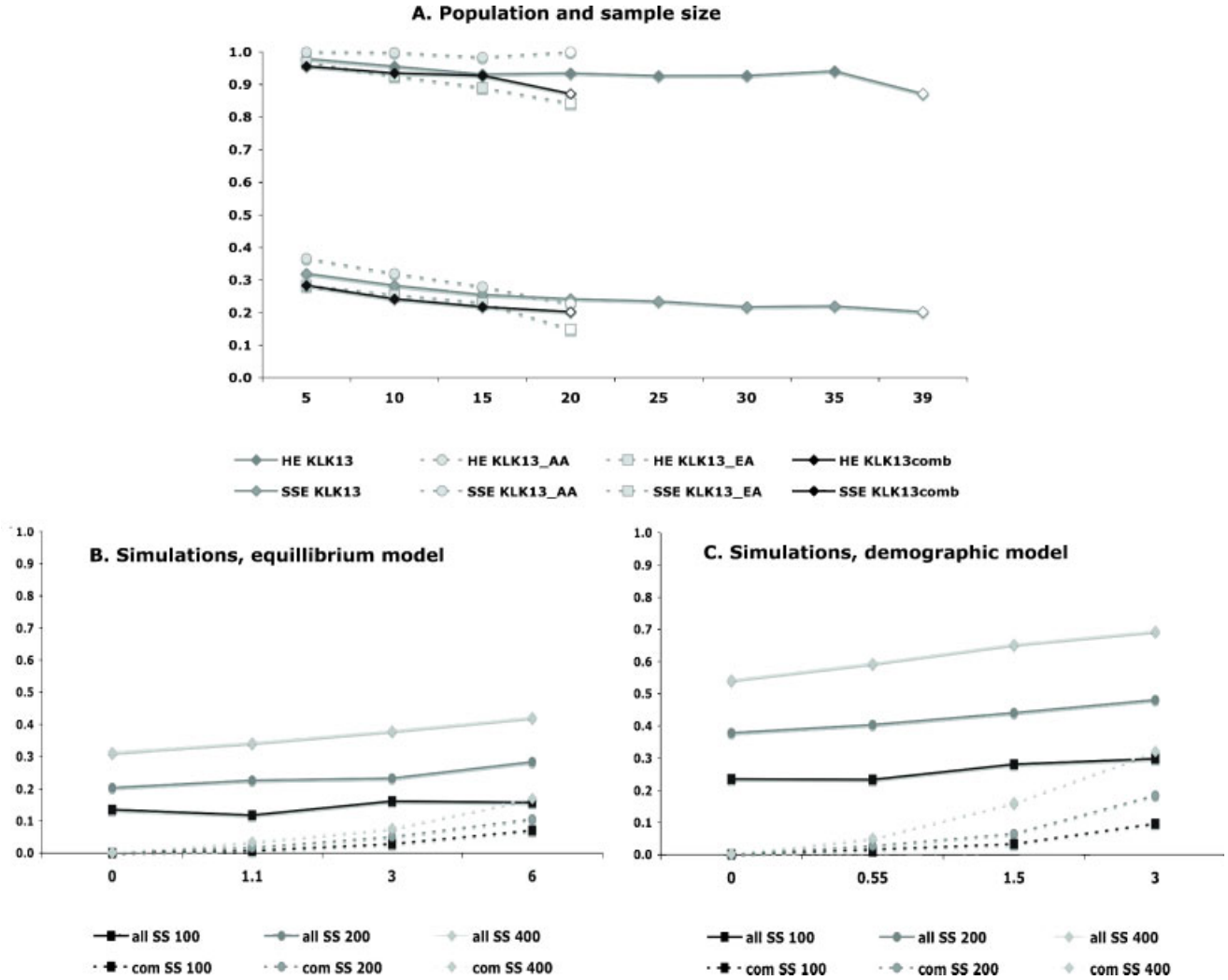

Fig. 2. (A) Haplotype error rate (upper) and SS error rate (lower) of haplotype reconstruction with PHASE 2.1 for simulated datasets of different sample sizes and population structure. *X*-axis: number of individuals in the sample; *Y*-axis: error rate. KLK13_AA: sample containing *X* individuals from African-American (AA) origin; KLK13_EA: sample containing *X* individuals from European-American (EA) origin; KLK13: sample containing *X* individuals from both populations (*X*/2 (AA)+*X*/2 (EA)); KLK13comb: sample containing 2*X* individuals, *X*(AA)+*X*(EA). The contrast of this last category with KLK13_AA and KLK13_EA illustrates the effect of phasing all individuals as a single sample, as opposed to reconstructing haplotypes separately by population. Note that points with white background (size 39 for KLK13, and size 20 for _AA and _EA) correspond to the error of the single best PHASE run on the original dataset. The rest of points correspond to the average error when phasing 50 pseudodatasets of the corresponding size. (B and C) Haplotype error (*Y*-axis) for coalescent simulations performed under the equilibrium (B) or demographic (C) models with different recombination rates (*X*-axis). Results are shown for different lengths of the segment (here indicated as number of SS), considering all SS or common SS. All haplotype reconstruction was performed with PHASE 2.1. The two graphs are not directly comparable due to inequality of population structure, long-range Ne, and ρ differences (see Materials and Methods).

HAPLOTYPE RECONSTRUCTION OF SIMULATED DATA

Phasing performance may be influenced by a number of attributes of the data, including segment length, number of SS, missing data, allele and haplotype frequencies, and mutation, recombination, and gene conversion rates. The relative influence of some of those factors was assessed using coalescent simulations. PHASE was the method of choice given its generally higher accuracy in the KLK region, but we expect comparable results with all methods.

HE is plotted by recombination rate and length of the fragment in Figures 2B and C. Figure 2B shows accuracy of haplotype reconstruction in an

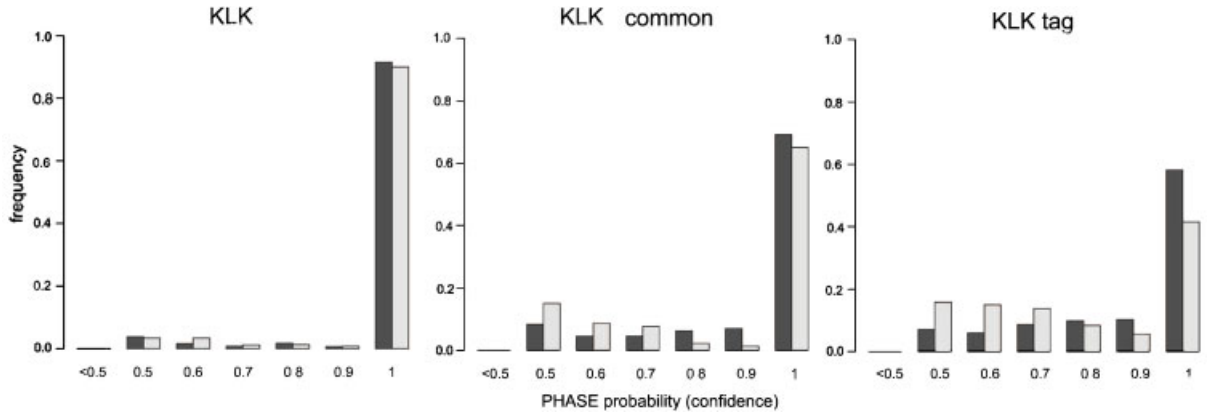

Fig. 3. Frequency of PHASE estimated confidence for a given site (*probability* of the site, see PHASE 2.1 documentation) for correctly and incorrectly assigned sites. *X*-axis: estimated confidence of the algorithm to individual sites (the *probability* of the site). Probability ranges from 0.5 to 1. *Y*-axes: relative frequency. Dark gray represents sites correctly assigned to haplotype (performance = 1) and light gray represents sites incorrectly assigned to haplotype (performance = 0). All values were calculated considering the information of three runs of the program for every dataset.

equilibrium population, and Figure 2C in samples simulated under a more realistic demographic model for humans. The Figures are not directly comparable due to intrinsic differences in population structure, N_e and ρ, but their inspection reveals some interesting features. The most striking observation is that haplotype inference is more accurate for simulated than for biological data of similar characteristics (note that, in length and SS density, the 25 kb dataset (200 SS) mimics KLK13/KK14, and the 50 kb region (400 SS) the complete KLK). This difference indicates that additional demographic or biological factors not accounted for in the simulations (like recurrent mutation, heterogeneous recombination, or gene conversion) interfere with haplotype reconstruction.

Demographic history has a large impact in phasing: introducing in the simulations a complex demographic scenario substantially erodes accuracy, even in the absence of recombination. This demographic effect is mainly due to the influence of low-frequency alleles (see dotted lines in Figs. 2B and C), as expected by the nature of the demographic scenario simulated (population expansions for both AA and EA, that increase the proportion of rare SS).

Missing data is an additional factor affecting haplotype inference, expected to reduce accuracy simply by adding to the uncertainty. According to the simulations, the effect of missing data is not dramatic, at least for PHASE and when missing sites appear at random (results not shown). Moreover, our experience analyzing the biological KLK data during the genotyping finishing process (that basically reduced missing data) suggests that this is not a critical variable for the phasing methods considered here.

PHASING CONFIDENCE

Having a good estimator of the confidence of phase calls would be useful in order to incorporate uncertainty information in subsequent haplotype analysis. PHASE attempts to assess confidence by reporting the posterior probability of each phased site, and here we examine the extent to which this information reflects actual performance. The use of confidence to predict site accuracy seems impractical at this point: most SS (correct or erroneous) have a probability of 1, and the distribution of probability frequencies overlaps enough to complicate simple inferences of accuracy based on reported probability (Fig. 3).

DISCUSSION

Statistical inference of haplotypes from multilocus genotypes is the most common method to retrieve haplotype information lost during multisite genotyping. By testing the ability of inference methods to reconstruct true haplotypes from shuffled phase-known data, we find some sobering messages for those seeking to use inferred haplotypes for subsequent analysis. For example, in the KLK region (48 kb long and 400 SS), inferred haplotypes contain an average of 25–30% misassigned ambiguous SS. Errors are typically clustered and performance can be improved by

removing rare SS or considering shorter regions, but even then the accuracy is considerably low.

Very recently, Marchini et al. [2006] assessed the accuracy of statistical haplotype inference of unrelated individuals using haplotypes inferred from HapMap family data. That study concludes that phasing accuracy is high even for unrelated individuals. Several facets of that dataset differ from ours, including the origin of phase information (since family-reconstructed haplotypes contain some sites of unresolved phase), ascertainment of polymorphisms (HapMap SNPs are ascertained ~1 SNP every 5 kb), sample size, and origin (Marchini et al. [2006] considered a single, larger, European sample), and number of regions (that study considered 100 regions). Despite those differences, the two studies reveal similar results. gSSE (*Incorrect genotype percentage* in Marchini et al. [2006]), is very similar between the two studies, and SwE, in this study matches Marchini et al. [2006] for the common SS datasets and it is up only by a factor of 2 when considering all SS. Nevertheless, conclusions differ mainly due to differential use of error measures. We use SSE as a measure of single site error, which represents the percentage of ambiguous sites incorrectly assigned to haplotype. In contrast, Marchini et al. [2006] chose gSSE, which represents the percentage of incorrectly assigned sites among all sites, ambiguous or not. Given the large number of homozygote sites at a given individual, gSSE is necessarily lower than SSE (see Table II). In our opinion, SSE is a better measure of accuracy (since it is independent of site frequency) and it better reflects the uncertainty in subsequent analysis introduced by haplotype reconstruction (since heterozygote sites in individuals are the ones that discriminate among their haplotypes).

Besides the origin of phase information and ascertainment of SNPs, our data differ from recent assessments [Stephens and Scheet, 2005; Marchini et al., 2006; Scheet and Stephens, 2006] in that our sample contains both AA and EA individuals, exactly as one finds in association tests in the US. According to our results, this mixed ancestry does not negatively affect the phasing process. In fact, our results suggest that in such a mixed ancestry (AA and EA) sample probably the best phasing results are obtained by reconstructing haplotypes on the combined sample, at least for small samples. Such pooling increases sample size, especially beneficial in small samples. In addition, the presence of EA chromosomes in the sample may help in the reconstruction of AA haplotypes, as variability outside of Africa is mostly a subset of African variation [Tishkoff et al., 1996; Reich et al., 2001; Gabriel et al., 2002; Kidd, 2004]. Moreover, pooling of populations may result in deviations from Hardy-Weinberg equilibrium toward excess of homozygosity. Even if deviations from equilibrium violate the assumptions of the methods, the algorithms seems robust to such deviations [Stephens et al., 2001] and, in cases of increased homozygosity, such deviations may improve haplotype inference by reducing the percentage of heterozygote sites and increasing LD.

Demographic history influences accuracy of haplotype inference not only by stratifying populations, but also because recent expansion has resulted in a high proportion of rare SNPs and haplotypes, which hamper the reconstruction. This issue can be overcome by simply not considering rare SS, but such data truncation is not desirable in many re-sequencing efforts where discovery of all variation may be a key assumption of statistical models to be applied to the data. Past action of natural selection can also negatively affect haplotype reconstruction, either by increasing the coalescence time of chromosomes (by balancing selection) or by creating local genealogies similar to those of population expansions (by positive selection).

It is important to note that our study is centered on a single genomic region and two specific human populations, and the generality of these results is unclear. Nevertheless, LD and haplotype structure do not seem unusual for this regions [Shimmin et al., in preparation]. Moreover, accuracy in the KLK region is similar to the average of 100 regions from Marchini et al. [2006] and the 134 regions from Kukita et al. [2005], suggesting that this region is probably representative of the genome. Note that both of these studies considered ascertained SNPs, at intermediate frequencies and low densities. The concordance of that accuracy with our data, very dense in SS, suggests that, contrary to previous expectations, increasing the density of the SNPs will not drastically improve haplotype reconstruction.

Unfortunately, simulation results show that the most important factors affecting haplotype reconstruction are, besides length of the region, those over which the investigator has little control (including the number of sites, demographic history of the population, or recombination rate), while elements that researchers can easily modulate (sample size or stratification of that sample) are somewhat less influential. Moreover, the

reconstruction is considerably less accurate in real data than in simulations, revealing that additional factors not accounted for in our simulations may be hampering the inference. These include heterogeneity in mutation and recombination rates, gene conversion, and recurrent mutation, or a more complex demographic history than that considered here. The observation that haplotype inference could be even more difficult for studies based on tagSNPs is especially troublesome, given the extensive anticipated use of tagSNPs in association studies.

In principle, reconstruction of haplotypes can be considerably improved by the addition of extrinsic evidence that facilitates the statistical inference. Experimental determination of ambiguous phase by allele-specific amplification can dramatically improve phasing, even when limited to a small number of SNP pairs and individuals [Clark et al., 1998]. Unfortunately, this method is not suitable for large-scale studies because experiments are individually designed along the phasing process, the technique is methodologically complex, and careful interpretation of agarose gels post-PCR is required. An alternative is to obtain haplotype information from pedigree data [Schaid, 2002; Schouten et al., 2005]. For example, by genotyping mother-father-child trios, the HapMap project considerably improved many of its haplotype inferences [The International HapMap consortium, 2005; Marchini et al., 2006]. The disadvantage of this strategy is that the use of trios triples the study sample size (increasing costs) and requires access to family members, which may be unavailable. An additional possibility would be to use a set of ''known'' haplotypes as 'predefined haplotypes' to help phasing genotype population data. These could be obtained from HapMap data (for CEPH and Yoruban), from sequence or genotyping of monosomic cell lines for candidate loci, or from the application of novel techniques to obtain phase information of long genomic regions [Kukita et al., 2005; Raymond et al., 2005].

Regardless of the method employed for phasing, uncertainty of the reconstruction should ideally be incorporated in subsequent analysis, especially in association testing. This does not seem a straightforward solution even if analyses were to integrate the uncertainty information provided by phasing software, given the complex relation between accuracy of phase inference and confidence reported by the algorithm. This relationship may be improved by new methods like fastPHASE, but at the price of lower accuracy [Scheet and Stephens, 2006 and this study].

In association studies, probably the simplest solution would be to treat the phase information as implicit in unphased genotypes, avoiding explicit haplotype phase inferences. The use of unphased genotype data has been proposed for LD analyses [Weir and Cockerham, 1989; Schaid, 2004] and for disease association mapping [Clayton et al., 2004; Morris et al., 2004], and these studies demonstrate that using unphased genotype data may have similar power and less error-associated problems than haplotype-based methods. In other cases, like in population-genetic studies where the structure of haplotypes is the object of interest, showing that results are not dependent on the phasing method could support their robustness. In all cases, if haplotypes must be reconstructed, it would be wise to focus exclusively on regions of high LD (where phasing is more accurate) and to avoid reconstructing haplotypes across very long regions unless only extremely close SS are to be considered (e.g. in sliding window approaches).

Haplotype reconstruction is a valuable statistical tool that plays an essential role in a wide variety of genetic studies. It is important to recognize the extraordinary improvement of the methods over the last 15 years, to the point where highly complex inferences provide useful results. It is equally important, though, to face their limitations. Haplotype reconstruction based solely on genotype data remains a challenge, and in many cases the underlying biology is just too complex to be completely predicted by statistical algorithms. In order to reduce the effect of haplotype inaccuracies in subsequent analysis, some possible strategies include the introduction of external haplotype information, the restriction of inferences to specific regions of high LD, or the explicit accommodation of a distribution of admissible haplotypes to test robustness of subsequent inferences that use haplotype information.

NOTE ADDED IN PROOF

The publication of this manuscript was delayed over the typical span of the journal's publication time due to difficulties in the submission of data to public databases. Due to difficulties in submitting the complete data, sequence is accessible through GenBank, SNPs through dbSNPs, and haplotype

information through our website, http://superc.hg.med.umich.edu/micortex/publications.

ACKNOWLEDGMENTS

The authors would like to thank Bret Payseur, Scott Williamson, and Kevin Thornton for useful discussions, and Sergi Castellano for computational help. Jian Li, Scott Williamson, Bret Payseur, and two anonymous reviewers provided valuable comments on the manuscript.

REFERENCES

Barrett JC, Fry B, Maller J, Daly MJ, Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. Am J Hum Genet 74:106–120.

Chung RH, Gusfield D. 2003. Perfect phylogeny haplotyper: haplotype inferral using a tree model. Bioinformatics 19: 780–781.

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

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333.

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

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415–428.

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. 2001. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28: 361–364.

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

Excoffier L, Laval G, Balding D. 2003. Gametic phase estimation over large genomic regions using an adaptive window approach. Hum Genomics 1:7–19.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Halperin E, Eskin E. 2004. Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics 20: 1842–1849.

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

Hill WG. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Kidd KK, Pakstis AJ, Speed WC, Kidd JR. 2004. Understanding human DNA sequence variation. J Hered 95:406–420.

Kimmel G, Shamir R. 2005. GERBIL: genotype resolution and block identification using likelihood. Proc Natl Acad Sci USA 102:158–162.

Kukita Y, Miyatake K, Stokowski R, Hinds D, Higasa K, Wake N, Hirakawa T, Kato H, Matsuda T, Pant K, Cox D, Tahira T, Hayashi K. 2005. Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. Genome Res 15:1511–1518.

Lin S, Cutler DJ, Zwick ME, Chakravarti A. 2002. Haplotype inference in random population samples. Am J Hum Genet 71:1129–1137.

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

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P, for the International HapMap Consortium. 2006. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78:437–450.

Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372.

Martinez-Arias R, Calafell F, Mateu E, Comas D, Andres A, Bertranpetit J. 2001. Sequence variability of a human pseudogene. Genome Res 11:1071–1085.

Mateu E, Calafell F, Lao O, Bonne-Tamir B, Kidd JR, Pakstis A, Kidd KK, Bertranpetit J. 2001. Worldwide genetic analysis of the CFTR region. Am J Hum Genet 68:103–117.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Niu T. 2004. Algorithms for inferring haplotypes. Genet Epidemiol 27:334–347.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Raymond CK, Subramanian S, Paddock M, Qiu R, Deodato C, Palmieri A, Chang J, Radke T, Haugen E, Kas A, Waring D, Bovee D, Stacy R, Kaul R, Olson MV. 2005. Targeted, haplotype-resolved resequencing of long segments of the human genome. Genomics 86:759–766.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES.

2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Salem RM, Wessel J, Schork NJ. 2005. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2:39–66.

Schaid DJ. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.

Schouten MT, Williams CK, Haley CS. 2005. The impact of using related individuals for haplotype reconstruction in population studies. Genetics 171:1321–1330.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

Stephens JC, Rogers J, Ruano G. 1990. Theoretical underpinning of the single-molecule-dilution. SMD. method of direct haplotype resolution. Am J Hum Genet 46:1149–1155.

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

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387.

Weir BS, Cockerham CC. 1989. Complete characterization of disequilibrium at two loci. In: Feldman MW, editor. Mathematical Evolutionary Theory. Princeton, NJ: Princeton University Press. p 86–110.

Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, Lynch HT, Chadwick RB, de la Chapelle A, Berg K, Eshleman JR, Yuan W, Markowitz S, Laken SJ, Lengauer C, Kinzler KW, Vogelstein B. 2000. Conversion of diploidy to haploidy. Nature 403:723–724.