# Haplotype inference from diploid sequence data: evaluating performance using non-neutral MHC sequences

DAVID H. BOS, SARA M. TURNER and J. ANDREW DEWOODY

*Bindley Bioscience Center and Department of Forestry & Natural Resources, Purdue University, West Lafayette, Indiana, USA*

**Bos, H., Turner, S. M. and DeWoody, J. A. 2007**. Haplotype inference from diploid sequence data: evaluating performance using non-neutral MHC sequences. — *Hereditas 144*: 228–234. Lund, Sweden. eISSN 1601-5223. Received January 16, 2007. Accepted August 23, 2007

The direct sequencing of PCR products from diploid organisms is problematic because of ambiguities associated with phase inference in multi-site heterozygotes. Several molecular methods such as cloning, SSCP, and DGGE have been developed to empirically reduce diploid sequences to their constitutive haploid components, but in theory these empirical approaches can be supplanted by analytical treatment of diploid sequences. Analytical approaches are more desirable than molecular methods because of the added time and expense required to generate molecular data. A variety of analytical methods have been developed to address this issue, but few have been rigorously evaluated with empirical data. Furthermore, they all assume that the sequences under consideration are evolving in a neutral fashion and assume a moderate number of heterozygous sites. Here, we use non-neutral major histocompatibility complex (MHC) sequences comprised of large numbers of heterozygous sites that are under strong balancing selection to evaluate the performance of the popular Bayesian algorithm implemented by the program PHASE. Our results suggest that PHASE performs admirably with non-neutral sequences of moderate length with numerous heterozygous sites typical of MHC class II sequences. We conclude that analytical approaches to haplotype inference have great potential in large-scale population genetic assays, but recommend groundtruthing analytical results using empirical (molecular) approaches at the outset of population-level analyses.

*David H. Bos, 715 W. State St., Pfendler Hall, West Lafayette, IN 47906, USA.* E-mail: dbos@purdue.edu

Genes of the major histocompatibility complex (MHC) play a primary role in the adaptive immune system and are known to impact the severity of various pathologies (HILL et al. 1991; CARRINGTON and O'BRIEN 2003). As such, MHC gene products have a direct effect on survivorship and fitness and have become a useful genetic marker for evolutionary biologists (EDWARDS and HEDRICK 1998). Often, evolutionary analyses entail genotyping a given MHC locus in a large sample of diploid individuals, but most laboratory methods cannot determine whether two or more segregating sites are in the cis- or trans-configuration (i.e. haplotype information; Fig. 1). This may not be problematic for protein-coding loci with limited polymorphism, but balancing selection maintains extremely high levels of stable polymorphism and heterozygosity in MHC genes (PARHAM and OHTA 1996). Thus, raw data sets of MHC sequences are often comprised of many segregating sites for which haplotype information is unknown. Phase-ordered haplotypes are nearly always desired, but the level of diversity at MHC loci makes recovery of these alleles a formidable task.

Empirical approaches exist to generate ordered haplotype data, but in vitro methods (e.g. cloning, DGGE or SSCP) are time-consuming, expensive, and/or may contain inaccuracies. Thus, analytical methods have been devised to secure these data in silico. The first algorithm for determining haplotype data was based on the parsimony principle (CLARK 1990), but subsequent methods focused on maximum-likelihood approaches (EXCOFFIER and SLATKIN 1995; HAWLEY and KIDD 1995). More recently, methods based on Bayesian statistics have been employed to reconstruct haplotype information from heterozygous diploid individuals (STEPHENS et al. 2001a; LIN et al. 2002; NIU et al. 2002; EXCOFFIER et al. 2003; ZHANG et al. 2006).

Bayesian methods consider the known genotypes of observed data (with ambiguous phase), and employ an initial guess of an individual's haplotype from a prior distribution of haplotypes. They then iteratively continue sampling from the prior distribution for all individual haplotypes until the frequency of inferred haplotypes maximizes the probability of obtaining the observed genotype data. Various methods have employed partition-ligation procedures introduced by NIU et al. (2002), and/or informative versus flat priors (STEPHENS et al. 2001a; ZHANG et al. 2006). Informative priors typically take the properties of a neutral coalescent process, wherein haplotypes "cluster" together, i.e. they predict that a new haplotype to be inferred will be similar to and can be obtained from an existing haplotype after applying a few mutations. A flat prior is naïve with regard to new haplotypes and assumes a priori that haplotypes are unrelated in terms

DOI: 10.1111/j.2007.0018-0661.01994.x

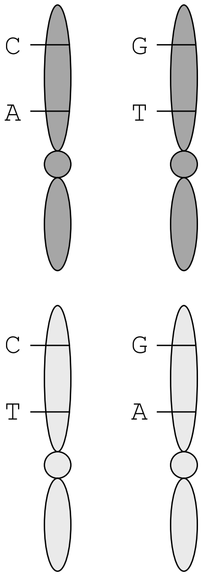

**Fig. 1.** Two possible phase configurations of a (diploid) double heterozygote. In the top pair of chromosomes, the C and A are in cis-configuration but C is trans to T; the situation is reversed in the bottom pair of chromosomes.

of sequence similarity and are randomly distributed in the prior.

Bayesian methods to infer haplotypes from unphased genotype data have been widely evaluated with simulated sequence data (STEPHENS et al. 2001a, 2001b; ZHANG et al. 2001; NIU et al. 2002; STEPHENS and DONNELLY 2003; ZHAO et al. 2003; TANG et al. 2006). In addition, some studies have used empirical human population data to evaluate haplotype inference (XU et al. 2002; ZHANG et al. 2006; LI et al. 2007). Data sets of assessments conducted so far often assume Hardy-Weinberg conditions, but some violations, most prominently the inclusion of samples admixed from multiple populations, have been conducted. We aim to more widely evaluate these methods in population genetic scenarios involving natural selection and population admixture separately and combined. We also use data where heterozygous sites are not simply bi-allelic, but display multiple alleles. Most in silico haplotyping programs permit the use of bi-allelic sites only; here, we use multi-allelic sites, and evaluate the performance of one of the few methods that can accommodate such heterozygosity.

In particular, the program PHASE relies on Bayesian logic and exhibits flexibility and effectiveness with neutral sequences (STEPHENS et al. 2001a). With PHASE, complex SNP data sets with three or more alleles can be analyzed, and two different priors are available (STEPHENS et al. 2001a). One prior approximates the neutral coalescent and the other is a naïve distribution with a "parent-independent mutation" pattern. The assumptions of both prior distributions are violated by natural selection, migration, fluctuation in population size, overlapping generations, and other common natural phenomena. Although simulated data that violate the neutral coalescent have been used to argue that PHASE performs more poorly than methods with flat prior distributions (NIU et al. 2002; XU et al. 2002), the authors of PHASE justify the informative prior as robust to model misspecifications, even when data with moderate levels of polymorphism contravene standard coalescent patterns (STEPHENS et al. 2001b; STEPHENS and DONNELLY 2003). For instance, the PHASE algorithm has a low error rate with data sets of ~50 individuals from multiple separate populations with 12–14 polymorphic sites (STEPHENS et al. 2001b).

In theory, both empirical and simulated data sets can be used to test haplotype estimation algorithms, but in practice empirical data have rarely been used for performance evaluations. Further, PHASE has never been tested using empirical data subject to strong natural selection with very high levels of polymorphism (e.g. MHC sequence data). Balancing selection creates a distribution of haplotype clusters that deviates markedly from the expectations of neutral evolution and the assumptions of standard neutral coalescent models (TAKAHATA and NEI 1990; PARHAM and OHTA 1996). Furthermore, elevated levels of heterozygosity and rates of nonsynonymous substitution ($d_N/d_S > 1$) typically result from balancing selection, and the former is known to cause problems with haplotype inference (CLARK 1990).

MHC class I and class II genes are the best known examples of genes that evolve under strong balancing selection (HUGHES and YEAGER 1998). We examined the performance of PHASE when presented with ambiguous-phase genotypes from several different empirical MHC data sets. These data sets contain sample sizes comparable to many population studies and numbers of heterozygous sites typically found at MHC loci, but far exceed the numbers of heterozygous sites for which PHASE is known to perform well. We also test whether there is a significant difference in the performance of PHASE when using two different prior distributions, one based on neutral coalescence and the other naïve with regard to the expected similarity of haplotypes. In so doing, we test whether or not PHASE represents a viable alternative for evolutionary biologists studying highly polymorphic nuclear (diploid) genes.

## MATERIAL AND METHODS

We generated sequences from the MHC class IIβ gene of Atlantic salmon (*Salmo salar, Sasa-DAB*) and tiger salamanders (*Ambystoma tigrinum, Amti-DAB*). Genomic DNA of salamanders was isolated using a standard lysis procedure (Sambrook and Russell 2001) and fragments *Amti-DAB* were PCR amplified as described in Bos and DeWoody (2005). Atlantic salmon DNA was isolated from fin clips using a standard lysis procedure, and a portion of *Sasa-DAB* was isolated via PCR using primers and protocols previously described in Landry et al. (2001). In both cases, PCR primers amplified a ~260bp fragment of exon 2, which encodes the highly polymorphic peptide binding region (PBR) which is the target of natural selection (Hughes and Hughes 1995). These data sets are expressed class II genes that have classical immune function (Hordvik et al. 1993; Bos and DeWoody 2005).

Bidirectional sequence data were generated directly from PCR products of salamanders and salmon, resulting in ambiguous-phase genotype data for heterozygous individuals. We used Phred scores (Ewing et al. 1998) to quantitatively evaluate the quality of our data by calculating a mean Phred score across all nucleotides in a sequence; we calculated a grand mean by averaging across sequences in a species. Initially, secondary peaks that were >60% as intense as the upper peak were automatically scored using the "call secondary peaks" function in Sequencher 4.1 (Genecodes). In practice, we found that automated calling of secondary peaks was error prone (Nickerson et al. 1997; Weckx et al. 2005) so we manually confirmed and/or edited each chromatogram using Sequencher. To groundtruth the results from the Bayesian inference, we cloned and sequenced the PCR products following standard protocols (Sambrook and Russell 2001) and identified all haplotypes present in both data sets (accession no. DQ071905 – DQ071913). We sequenced between six and twenty clones per individual and in so doing identified both haplotypes in heterozygotes and confirmed the single haplotype (represented multiple times) in homozygotes. All tiger salamander sequences were cloned; data for *Sasa-DAB* was confirmed using the Luminex-100 flow cytometry platform (Vignali 2000; Itoh et al. 2005).

From the empirical data generated in our laboratory, we assembled five data sets (Table 1). All of the sampled salamanders (n = 30) were used in a single data set termed Amti-DAB. From the Atlantic salmon (n = 103), we generated ambiguous-phase genotype data via DNA sequencing directly from PCR product. We sub-sampled 40 of those individuals for one data set (Sasa-DAB$^1$). Three other data sets (Sasa-DAB$^{2-4}$) of 20 individuals each were also sub-sampled to create different data sets with varying degrees of individual-level heterozygosity. These data sets contained one (Sasa-DAB$^{3,4}$) or two (Sasa-DAB$^2$) homozygous individuals each. The distinguishing factor between Sasa-DAB$^3$ and Sasa-DAB$^4$ is that the homozygote individual in Sasa-DAB$^4$ is comprised of two copies of a haplotype that was not found in any other individuals, mimicking the brown trout (*Salmo trutta*) scenario detailed below.

The aforementioned data sets were each comprised of samples from a single interbreeding population. In addition to the empirical sequences generated in our laboratory, we assembled four more data sets from GenBank sequences. Two of these data sets comprise full-length trout cDNA sequences collected by Shum et al. (2001). From brown trout, we used a class II gene (*Satr-DAB*) data set collected from a single population. This data set has the highest ratio of heterozygous sites to sample size; it also contains only a single homozygote and the haplotype (allele) of that individual is not found elsewhere in the data set. From rainbow trout (*Oncorhynchus mykiss*), we used sequences of the *Onmy-DAB* class II gene from fifteen total sampled individuals collected from two distinct populations that share no haplotypes.

The final two data sets are comprised of full-length cDNA sequences of human class I *HLA-B* sequences obtained from GenBank. One of these (*HLA-B* Brazil)

Table 1. *Parameters of nine data sets used in PHASE analysis.*

| Data set | Amti-DAB | Sasa-DAB$^1$ | Sasa-DAB$^2$ | Sasa-DAB$^3$ | Sasa-DAB$^4$ | Onmy-DAB | Satr-DAB | HLA-B Kenya | HLA-B Brazil |
|---|---|---|---|---|---|---|---|---|---|
| Population sample size | 30 | 40 | 20 | 20 | 20 | 15 | 10 | 40 | 40 |
| Sequence length | 264 | 254 | 254 | 254 | 254 | 653 | 653 | 1065 | 1065 |
| Heterozygous sites | 40 | 38 | 38 | 38 | 40 | 56 | 82 | 100 | 74 |
| No. haplotypes | 8 | 6 | 6 | 7 | 8 | 14 | 12 | 18 | 10 |
| Sample heterozygosity | 0.733 | 0.725 | 0.900 | 0.950 | 0.950 | 0.667 | 0.900 | 0.925 | 0.875 |
| Ave. PBR $d_N/d_S$ | 2.94 | 6.02 | 6.02 | 6.03 | 6.03 | 2.60 | 2.00 | 4.50 | 3.47 |
| No. of source pops. | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |

was assembled from observed sequences from the Kaingang and Guarani tribes in Brazil (Belich et al. 1992). The other (*HLA-B* Kenya) was assembled from observed sequences of Kenyan highlander and Kenyan lowlander tribes (Kitts et al. 2003). For each of the human data sets, we used all haplotypes from two related populations to construct 40 hypothetical individuals by randomly joining two haplotypes, and the probability of each haplotype sampled corresponded to its observed frequency in the combined population. Such a scenario, which corresponds to admixture of two separate populations, represents a level of complexity beyond balancing selection. Therefore, *HLA-B* data sets violate the neutral coalescent not only with respect to selection, but migration as well.

Haplotypes (i.e. phase-ordered alleles) were determined using PHASE ver. 2.1, described in Stephens et al. (2001a) and Stephens and Donnelly (2003). Each ambiguous-phase genotype data set was run blind, prior to the empirical collection of phase-ordered alleles. Each data set was analyzed 100 times in PHASE, each with different random number seeds to ensure independence of runs; each run was conducted with 1000 iterations as a burn-in period followed by 1000 iterations each with 10 steps through a Markov chain (i.e. 10 × the number of iterations suggested by the program's default setting). We employed the MS model (no recombination) because it was a better fit to the data, and because of the tight physical linkage of our heterozygous sites. We also relaxed the assumption of a stepwise mutation model that was designed for microsatellite data. Each run was performed in duplicate using two different prior distributions (Stephens et al. 2001a). The first is termed the "coalescent prior" and uses the default prior where the Gibbs sampler draws from a distribution of haplotypes that informally approximates the distribution occurring through a neutral coalescent process. The second run utilizes the naïve Gibbs sampler where the distribution reflects "parent-independent mutation" that ignores relationships among haplotypes ("the naïve prior"). A major difference between these priors is the informal approximation of the coalescent process.

Empirically-determined genotypes were then compared to the analytical genotypes inferred by PHASE. We tested performance in two ways: 1) haplotype identification, which is defined as the proportion of runs for which the haplotypes (alleles) inferred from each data set are completely correct, and 2) genotype assignment, the proportion of individuals assigned each of their true diploid haplotypes (i.e. genotyped correctly).

## RESULTS

We generated sequences directly from the PCR products of n = 30 tiger salamanders and n = 103 Atlantic salmon. PCR amplicons were cloned into a T-vector, and we sequenced a total of 280 clones from tiger salamanders and 677 clones from Atlantic salmon. For *A. tigrinum*, the grand mean Phred score was 42.75 (n = 50 sequences; SE = 1.04) and for *S. salar* the grand mean was 51.81 (n = 81 sequences; SE = 0.62). Phred scores of 30 or more mean the associated probability of error at a particular site is 0.001 (Gibson and Muse 2001). Thus, our raw sequences were high-quality by objective, quantitative measures.

In total, our own work generated roughly 273 kb of sequence data; we also collected bioinformatic sequence data for four additional data sets. These data sets represent a variety of complexity in regards to sample size, numbers of known alleles (found in homozygote individuals), number of heterozygous sites, etc. (Table 1). Consistent with typical MHC class I and class II genes, the PBR of all data sets show the signature of balancing selection in terms of elevated substitution rate ratios ($d_N/d_S$).

When we ran the ambiguous-phase genotype data through the program, the haplotypes present in Sasa-DAB[1] and Amti-DAB data sets were correctly identified in every case (Fig. 2). Furthermore, the genotypes of each individual were faithfully reconstructed. PHASE performed these functions consistently, with 100% of the program runs resulting in the correct outcome. The correct results for these data were obtained using both the coalescent prior and the naïve prior. Analysis of Sasa-DAB[2–4] produced similar results despite different gene pools and levels of zygosity. In all cases, the coalescent prior and the

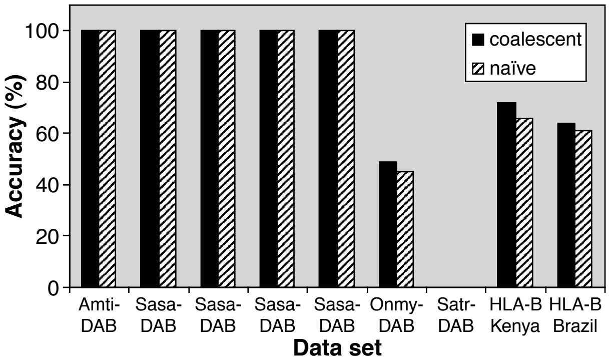

**Fig. 2.** Results of 100 independent runs of PHASE. The performance of two different prior distributions on nine data sets are compared. One hundred independent runs were performed to compare the number of times the haplotypes were inferred without error. Genotype assignment correlated perfectly with haplotype inference.

naïve prior produced identical results. Recall that the salmon and salamander data were collected from single panmictic populations; thus, they serve as a baseline for comparison to other data sets.

The Onmy-DAB data set was comprised of MHC class II genotype data from two different populations that have no haplotypes in common (Shum et al. 2001). Like the Amti-DAB and Sasa-DAB$^{1-4}$ data sets, Onmy-DAB violates the coalescent assumption of neutrality. Unlike the salamander and salmon data sets, Onmy-DAB also violates the coalescent assumption of panmixia and thus presents a more complex problem for PHASE. Not surprisingly, the performance of PHASE was worse; only ~50% of the runs resulted in accurate reconstruction of the haplotypes using the coalescent prior (Fig. 2). Performance was worse using the naïve prior, although the difference was not significant ($\chi^2 = 0.316$; $p > 0.5$). For both priors, all individuals were genotyped correctly in runs where all haplotypes were correctly identified.

Heterogeneous results across multiple independent runs could result from failure of the Markov chain Monte Carlo procedure to find convergence. We therefore increased the number of run iterations from 150 000 to 375 000 and performed ten independent runs of PHASE. The results indicated that the algorithm still did not consistently converge to the same answer over the ten runs (8 correct; 2 wrong). We selected the preferred outcome as the one with the highest average value for the goodness of fit and in all cases, the runs with the correct haplotypes had higher goodness of fit than the incorrect runs. Thus, average goodness of fit values were highly informative with regard to the accurate inference of haplotype.

The Satr-DAB data set was collected from a single population and apparently represented a difficult scenario for haplotype identification. For these data all runs resulted in at least one error, so that none of the runs accurately identified every haplotype (Fig. 2) regardless of the prior (coalescent or naïve) or number of iterations. Because of the errors in haplotype identification, errors were also seen in genotype assignment because the latter is predicated on the former.

Haplotypes of the *HLA-B* data set of humans from Brazil were comprised of samples from two populations that shared five haplotypes (Belich et al. 1992). In our simulations, three different haplotypes were found in a homozygous state among five samples. For these data, haplotype identification was accurate in most runs using the coalescent prior whereas accuracy was slightly reduced using the naïve prior. The difference between the number of error-free results among the priors was nonsignificant ($\chi^2 = 0.840$; $p > 0.3$). Where the haplotypes were incorrectly inferred, the genotypes of individuals also contained errors. Under both priors, the runs with the error-free results always had a higher average goodness of fit than the runs with errors.

The *HLA-B* data from Kenyan populations were created in similar fashion as the *HLA-B* data from Brazil. The empirical data contained numerous haplotypes which were shared among the original source populations, but none of the haplotypes exceeded a frequency of 0.15. Most of the PHASE runs were error free when the coalescent prior was used, whereas the naïve prior was again slightly (but not significantly; ($\chi^2 = 0.192$; $p > 0.5$)) less accurate than the coalescent prior. As with other data sets, most of the errors were the creation of more haplotypes than actually existed in the input data.

## DISCUSSION

### *Performance of PHASE*

Our study uses empirical, non-neutral data sets to test the performance of PHASE algorithms. Major violations of the assumptions of the neutral coalescent occurred in each of our nine data sets, either by way of natural selection, population structure, or both (Table 1). Furthermore, each of these data sets challenged PHASE with empirical data comprised of large numbers of heterozygous sites. Our results suggest that the Bayesian approach of PHASE is robust even when polymorphism is high and certain assumptions of the prior distribution are violated. PHASE perfectly reconstructed genotypes from the Atlantic salmon and tiger salamander data sets. In both cases, haplotypes were accurately identified and individual genotypes were correctly assigned when using either a prior that informally assumes an underlying coalescent process or a prior that does not. There was no significant difference between the performance of the two priors, even when zygosity differed (*Sasa* data sets; Fig. 2). These results suggest PHASE can accurately determine haplotypes and genotypes using complex data under a variety of scenarios.

Roughly two-thirds of the runs with empirical data sets were correct regardless of the underlying prior. On the other hand, PHASE failed to accurately identify the alleles present in the brown trout data set, regardless of the prior distribution used for analysis. The brown trout data contain very high numbers of heterozygous sites compared to sample size and only a single homozygote. Previous work has shown that increasing levels of heterozygosity leads to inferior performance of this and other haplotyping methods

(Clark 1990; Niu et al. 2002). We subsampled the Atlantic salmon data to create data sets with varying population-level heterozygosity, and in all those samples PHASE functioned well. Therefore, we conclude that in addition to population-level heterozygosity, the relationship between haplotype-level heterozygosity and sample size also strongly influences performance of in silico haplotyping methods. This is evident when one compares algorithm performance using the Satr-DAB and Sasa-DAB$^4$ data sets (Fig. 2, Table 1), both of which have only a single homozygote that shares no haplotypes with any other sampled individual. Although PHASE is capable of accurately reconstructing haplotypes under a variety of conditions that violate assumptions of underlying models, the performance will obviously vary depending on the size and makeup of a particular data set.

In most cases, PHASE performed admirably by reconstructing error-free haplotypes. One emerging trend from our analysis is that PHASE performed more poorly on longer sequences. These longer sequences consist of cDNAs, which in some cases may represent large genomic sequences (Shum et al. 2001). This provides more opportunity for recombination to occur simply due to sequence length. In addition *HLA-B* genes are known to experience higher rates of recombination compared to other genes of the human class I and class II MHC genes (Jakobsen et al. 1998). Therefore, the poorer performance of PHASE on longer sequences may be explained by the fact that we relied on a model that does not account for possible recombination. For that reason, we ran 10 iterations of PHASE on the two HLA-B datasets using a model that accounts for recombination. PHASE correctly predicted haplotypes in 7 out of 10 and 8 out of 10 runs for the HLA-Brazil and HLA-Kenya datasets, respectively, which is similar to the rate of success for the model with no recombination. Therefore we cannot ascribe the poorer performance of PHASE to recombination when using longer sequences. More likely, the performance is related to the dynamic between haplotype complexity, sample heterozygosity and sample size.

Notably, individual genotype assignment by PHASE was perfect whenever haplotype identification was robust; the only time we encountered errors in genotype assignment was when there were problems in haplotype identification. The use of one prior distribution over the other in these data sets resulted in no significant difference in the number of times the results were error-free. Equally encouraging is that for data sets in which runs resulted in both correct and incorrect results, goodness-of-fit rankings resulted in selection of an error-free result for all but the *HLA-B* data from Kenya. Further, in cases where consistency cannot be achieved, we also find support for using the results of the run with the highest average goodness of fit (Stephens et al. 2001a).

### *Data collection*

Uneven peak heights, differential amplification of haplotypes, and/or *Taq* fidelity errors commonly lead to both type I and type II errors in secondary peaks on a chromatogram (Parker et al. 1996; Nickerson et al. 1997). Further, there remains a compelling tradeoff between type I and type II errors that may lead to falsely inferred SNPs when calling heterozygous sites (Weckx et al. 2005). This can be especially problematic for MHC class I or class II loci, where numerous polymorphic sites are found in close proximity. Such incorrect data input obviously leads to errors in haplotype assessment; because SNP typing and sequence data are not error-free, we suggest empirical groundtruthing of the haplotypes (alleles) inferred with PHASE before proceeding with population-level analyses. Such groundtruthing is routine with empirical methods of haplotyping that rely on laboratory techniques (e.g. SSCP). We urge caution and care in the strongest possible terms when collecting phase-unordered sequence data from diploid or polyploid organisms, because our experience suggests that most haplotype inference errors can be traced to problematic raw data.

In summary, our analyses indicate that both PHASE priors – neutral coalescent and naïve – work well with DNA sequences of MHC class II data generated directly from PCR products. These two prior distributions have not been evaluated head-to-head in the scientific literature, and here we show their performance is comparable across a variety of different empirical data sets. Furthermore, their performance is good even when rapidly evolving, non-neutral sequences are used with levels of polymorphism commonly seen in MHC data sets. This is fortunate for evolutionary biologists and geneticists who study highly polymorphic diploid genes.

*Acknowledgements* – We thank the DeWoody lab group for their input and assistance. We especially thank Rod Williams, Cory McCormick, and Roger Greil (Lake Superior State University's Aquatic Lab) for their help in procuring samples and performing lab work. This project was supported by a National Research Initiative Grant no. 2004-35205-14188 from the USDA Cooperative State Research, Education, and Extension Service and by a National Science Foundation grant (DEB-0514815) to J.A.D. This manuscript is ARP contribution no. 2006-17843 from Purdue University.

## REFERENCES

Belich, M. P., Madrigal, J. A., Hildebrand, W. H. et al. 1992. Unusual HLA-B alleles in two tribes of Brazilian indians. – Nature 357: 326–329.

Bos, D. H. and DeWoody, J. A. 2005. Molecular characterization of major histocompatibility complex class II alleles in wild tiger salamanders (*Ambystoma tigrinum*). – Immunogenetics 57: 775–781.

Carrington, M. and O'Brien, S. J. 2003. The influence of *HLA* genotype on AIDS. – Annu. Rev. Med. 54: 535–551.

Clark, A. G. 1990. Inference of haplotypes from pcr-amplified samples of diploid populations. – Mol. Biol. Evol. 7: 111–122.

Edwards, S. V. and Hedrick, P. W. 1998. Evolution and ecology of MHC molecules: from genomics to sexual selection. – Trends Ecol. Evol. 13: 305–311.

Ewing, B., Hillier, L., Wendl, M. C. et al. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. – Genome Res. 8: 175–185.

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

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

Gibson, G. and Muse, S. V. 2001. A Primer of genome science. – Sinauer.

Hawley, M. E. and Kidd, K. K. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-state haplotypes. – J. Hered. 86: 164–167.

Hill, A. V. S., Allsopp, C. E. M., Kwiatkowski, D. et al. 1991. Common West African HLA antigens are associated with protection from severe malaria. – Nature 352: 595–600.

Hordvik, I., Grimholt, U., Fosse, V. M. et al. 1993. Cloning and sequence analysis of cDNAs encoding the MHC class II β chain in Atlantic salmon (*Salmo salar*). – Immunogenetics 37: 437–441.

Hughes, A. L. and Hughes, M. K. 1995. Natural selection on the peptide binding regions of major histocompatibility complex molecules. – Immunogenetics 42: 233–243.

Hughes, A. L. and Yeager, M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. – Annu. Rev. Genet. 32: 415–435.

Itoh, Y., Mizuki, N., Shimada, T. et al. 2005. High-throughput DNA typing of HLA-A, -B, -C, and -DRB1 loci by a PCR-SSOP-Luminex method in the Japanese population. – Immunogenetics 57: 717–729.

Jakobsen, I. B., Wilson, S. R. and Easteal, S. 1998. Patterns of reticulate evolution for the classical class I and II HLA loci. – Immunogenetics 48: 312–323.

Kitts, A., Faolo, M. and Helmberg, W. 2003. The major histocompatibility complex database, MHCdb.

Landry, C., Garant, D., Duchesne, P. et al. 2001. 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). – Proc. R. Soc. Lond. B 268: 1279–1285.

Li, S., Cheng, J. and Zhao, L. 2007. Empirical vs Bayesian approach for estimating haplotypes from genotypes of unrelated individuals. – BMC Genetics 8: 2.

Lin, S., Cutler, D. J., Zwick, M. E. et al. 2002. Haplotype inference in random population samples. – Am. J. Hum. Genet. 71: 1129–1137.

Nickerson, D., Tobe, V and Taylor, S. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. – Nucleic Acids Res. 25: 2745–2751.

Niu, T., Qin, Z. S. and Liu, J. S. 2002. Bayesian classification for genome-wide control using multiple unlinked haplotype blocks. – Am. J. Hum. Genet. 71: 578–578.

Parham, P. and Ohta, T. 1996. Population biology of antigen presentation by MHC class I molecules. – Science 272: 67–74.

Parker, L. T., Zakeri, H., Deng, Q. et al. 1996. AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns. – Biotechniques 21: 694–699.

Sambrook, J. and Russell, D. W. 2001. Molecular cloning: a laboratory manual. – Cold Spring Harbor Press.

Shum, B. P., Guethlein, L. A., Flodin, L. R. et al. 2001. Modes of salmon MHC class I and II evolution differ from the primate paradigm. – J. Immunol. 166: 3297–3308.

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

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

Stephens, M., Smith, N. J. and Donnelly, P. 2001b. A reply to Zhang et al. – Am. J. Hum. Genet. 69: 912–914.

Takahata, N. and Nei, M. 1990. Allelic genealogy under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. – Genetics 124: 967–978.

Tang, H., Coram, M., Wang, P. et al. 2006. Reconstructing genetic ancestry blocks in admixed individuals. – Am. J. Hum. Genet. 79: 1–12.

Vignali, D. A. 2000. Multiplexed particle-based flow cytometric assays. – J. Immunol. Meth. 243: 243–255.

Weckx, S., Del-Favero, J., Rademakers, R. et al. 2005. novoSNP, a novel computational tool for sequence variation discovery. – Genome Res. 15: 436–442.

Xu, C. F., Lewis, K., Cantone, K. L. et al. 2002. Effectiveness of computational methods in haplotype prediction. – Hum. Genet. 110: 148–156.

Zhang, S., Pakstis, A. J., Kidd, K. K. et al. 2001. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. – Am. J. Hum. Genet. 69: 906–912.

Zhang, Y., Niu, T. and Liu, J. S. 2006. A coalescence-guided hierarchical Bayesian method of haplotype inference. – Am. J. Hum. Genet. 79: 313–322.

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. 72: 1231–1250.