Molecular Ecology (2009) **18**, 1556–1559

NEWS AND VIEWS

REPLY

Internal algorithm variability and among-algorithm discordance in statistical haplotype reconstruction

ZU-SHI HUANG,* YA-JIE JI* and DE-XING ZHANG*†

**State Key Laboratory of Integrated Management of Pest Insects and Rodents,* †*Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Datun Road, Chaoyang District, Beijing 100101, China*

The potential effectiveness of statistical haplotype inference makes it an area of active exploration over the last decade. There are several complications of statistical inference, including: the same algorithm can produce different solutions for the same data set, which reflects the internal algorithm variability; different algorithms can give different solutions for the same data set, reflecting the discordance among algorithms; and the algorithms per se are unable to evaluate the reliability of the solutions even if they are unique, this being a general limitation of all inference methods. With the aim of increasing the confidence of statistical inference results, consensus strategy appears to be an effective means to deal with these problems. Several authors have explored this with different emphases. Here we discuss two recent studies examining the internal algorithm variability and among-algorithm discordance, respectively, and evaluate the different outcomes of these analyses, in light of Orzack (2009) comment. Until other, better methods are developed, a combination of these two approaches should provide a practical way to increase the confidence of statistical haplotyping results.

Keywords: algorithm, ambiguous genotype, consensus vote, flanking region of nuclear microsatellite DNA, inconsistency

Received 16 November 2008; revision accepted 3 February 2009

Haplotype reconstruction is an important issue in the application of nuclear DNA markers in molecular ecology and evolutionary studies. Pure experimental approaches have encountered various difficulties when sample size is large such as in population genetic analysis (Zhang & Hewitt 2003). With the great increase of computer power, the potential effectiveness of statistical haplotype inference had made it an area of active exploration over the last decade. So far, more than 40 algorithms have appeared (for a recent review, see Salem *et al*. 2005), indicating the complex status and difficulty of statistical haplotype inference.

Correspondence: De-Xing Zhang, China. Fax: (+86) 10 64807232; E-mail: dxzhang@ioz.ac.cn

Several factors contribute to the complication of statistical inference, including: (i) the same algorithm can produce different solutions for the same data set, (ii) different algorithms can give different solutions for the same data set, and (iii) the algorithms per se are unable to evaluate the reliability of the solutions even if they are unique. The first problem reflects internal algorithm variability, i.e. the consistency of an algorithm. High consistency is an important requirement for an algorithm to be practically useful. The second problem reflects the discordance among algorithms, largely indicating the differences of the underlying theories and techniques which different algorithms employ. The third problem is a general limitation of all inference methods since, given the complexity of genetic data, no single simplified model can closely approach the truth in all circumstances. Statistical inference of haplotype and haplotype frequency needs to consider these problems, and in particular, among-algorithm discordance provides a yardstick to examine haplotype inference uncertainties (e.g. Saito *et al*. 2003; Kuittinen *et al*. 2008). In the absence of better methods, consensus strategy is an effective means for dealing with such problems and it has been successfully applied within a variety of fields, including economics, management, systematics, meteorology, and climatology (Araújo & New 2007). With regard to the field of biology, consensus techniques are commonly employed as a method of combining information from rival candidates or competing approaches, for example to define functional sequence motifs in molecular biology (e.g. Lewin 1987), or to summarize phylogenetic trees (reviewed in Bryant 2003).

The internal algorithm variability warrants further discussion in light of the comment of Orzack (2009) in this issue of *Molecular Ecology*. Most statistical algorithms require, either explicitly or implicitly, a test of consistency (e.g. p. 113, Clark 1990). Naturally, users should conduct multiple independent runs with different starting conditions to test the consistency of an algorithm before performing a formal analysis of their data. If significant inconsistency is observed, additional measures should be taken. Orzack *et al*. (2003) presented one of the most extensive attempts to examine the internal algorithm variability of Clark's rule-based approach (Clark 1990) under eight variations, and observed substantial differences in the average number of correctly inferred haplotype pairs among runs with different sample paths (both among and within variations). Given such a discrepancy, they adopted consensus methods to combine multiple inferrals to produce a set of unique solutions, and showed that the consensus set of inferrals is 'more accurate than the typical single set of inferrals chosen at random' (p. 915, Orzack *et al*. 2003). Then they used the consensus values to assess the reliability of the inferrals

(hereafter this procedure is referred to as 'consensus prediction'), and observed that with some of their variations 'as the consensus threshold or number of identical inferrals increases, there is a threshold value for which all inferrals are correct' (p. 921) (here and below, 'correct' means that an inferral is identical to the real type as determined by the experiment). Inferrals whose consensus values are greater than a given frequency threshold (e.g. ≥ 80%) may thus be considered as certain or nearly certain solutions. Essentially, Orzack *et al*.'s (2003) method allows identifying the most consistent solutions of an algorithm.

Huang *et al*. (2008) explored a consensus vote approach that, by inspecting the among-algorithm discordance, aims to evaluate the reliability of the inference solutions. They take into account the following considerations: First, since no algorithm is flaw-free (it has been demonstrated that no individual algorithm is free from inference errors in more than 10 popular algorithms examined; Huang *et al*. 2008), the inferred haplotypes and phase assignments can only be regarded as statistical solutions and are subject to errors; similarly, the estimates of confidence suggested by the algorithms can serve only as a reference (Xu *et al*. 2002; Sabbagh & Darlu 2005). Second, algorithms *based on different theoretical and technical frameworks* (that is, methodologically independent) are expected to behave independently and have different sensitivity to various genetic and other stochastic factors (Niu *et al*. 2002; Zhang *et al*. 2002; Niu 2004). Therefore, inferrals obtained from independent algorithms for any given ambiguous genotype can largely be regarded as independent statistical solutions. Third, given these considerations, if methodologically independent algorithms are included, and if these algorithms are fairly accurate, it is reasonable to expect that solutions approved by all algorithms are most likely reliable and subject to lower mean error than those from individual algorithms (cf. Araújo & New 2007).

Thus, there exist some fundamental differences between Orzack *et al*.'s (2003) consensus method and Huang *et al*.'s (2008) consensus vote approach. While Orzack *et al*.'s (2003) method focused on internal algorithm variability and aimed to identify the most consistent solutions of an algorithm, Huang *et al*. (2008) considered the among-algorithm discordance and identified the set of solutions approved by independent methods. Operationally, the effectiveness of Huang *et al*.'s approach depends on the methodological independence and accuracy of individual algorithms. The effectiveness of Orzack *et al*.'s (2003) consensus prediction procedure depends on the accuracy and the internal variability of an algorithm, as well as the frequency threshold value adopted (but it seems that no general principle exists for determining (i) the *a priori* frequency threshold value, and (ii) which of Orzack *et al*.'s variations work well in practice, even for simple data sets). Table 1 shows the results of Huang *et al*.'s (2008) data analysed with several variations

Table 1 Analysis of Huang *et al*.'s (2008) data following Orzack *et al*. (2003)

		NDH*		Error rate (%)†	
Methods	No. of iterations	Independent solutions	Consensus solution	Independent solutions	Consensus solution
Variation 1	1000	107.4 (4.7)	83	15.4 (1.5)	10.5
Variation 2	1000	91.1 (3.7)	83	14.2 (3.1)	10.5
Variation 2b	1000	85.5 (6.3)	82	11.1 (3.7)	10.0
Variation 3	1000	88.6 (1.3)	84	11.0 (0.4)	10.7
Variation 4b	1000	85.6 (6.7)	81	11.3 (4.4)	10.0
HAPINFERX-1	1000	72.0 (1.2)	69	7.3 (0.5)	6.7
HAPINFERX-2	1000	72.5 (2.3)	70	7.8 (1.8)	7.3
HAPINFERX-2b	1000	71.5 (2.7)	69	7.2 (2.1)	6.7
HAPINFERX-3	1000	73.0 (0)	70	7.2 (0.1)	7.3
HAPINFERX original	1000	71.6 (3.1)	69	7.3 (2.7)	6.9
PHASE	100	60.0 (0.3)	59	2.9 (0.2)	2.9
HAPLOTYPER	100	60.6 (0.8)	59	3.2 (0.3)	2.9
Arlequin-EM	100	69.0 (0)	66	6.1 (0)	6.1
Consensus vote	NA	NA	62	NA	2.7

Variations 1, 2, 2b, 3 and 4b were described by Orzack *et al*. (2003); HAPINFERX-1, -2, -2b and -3 are the corresponding HAPINFERX variations, respectively. Note that 'HAPINFERX original' refers to the unmodified HAPINFERX (Perl version) provided by Professor A. G. Clark, which differs from the algorithm presented in Clark (1990). Three other algorithms (PHASE, HAPLOTYPER, Arlequin-EM) are also included. Values in parenthesis are standard deviation. Here, the internal variability of an algorithm was indicated by the values of the standard deviation of NDH. About the data: the scnpc76 data contains 1052 chromosomes, with 70 polymorphic nucleotide sites, 63 distinct haplotypes and 138 distinct genotypes. There are 179 homozygotes, 15 one-site heterozygotes and 332 multisite heterozygotes. The length of the sequences is 251 bp. Genotyping-error-free scnpc76 data of *Locusta migratoria* was used here; this explains the slight difference of the statistics compared to those reported in Huang *et al*. (2008). It appears that consensus solutions with fewer number of inferred haplotypes are generally more accurate for the scnpc76 data (Spearman rank correlation coefficient, 0.988 for the individual error, d.f. = 9, $P < 0.001$) (this may be a general feature for regions with strong linkage disequilibrium). NA, not applicable.

*number of distinct haplotypes in a solution.

†inference error rate of individuals.

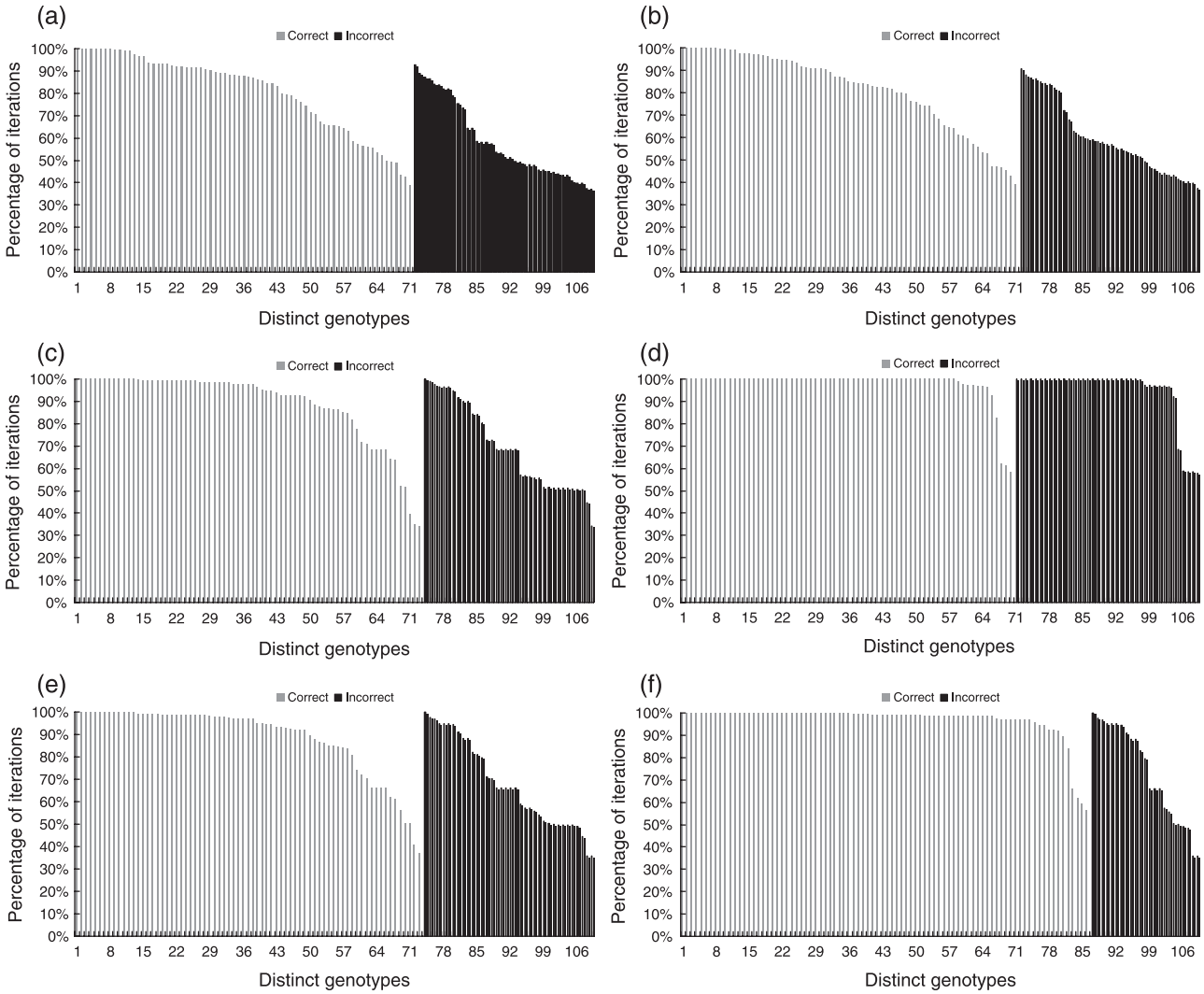

Fig. 1 Frequency distribution of the distinct genotypes of Huang *et al*.'s (2008) data inferred using different variations of Clark's method. (a–e) correspond to variations 1, 2, 2b, 3 and 4b described by Orzack *et al*. (2003), respectively, and f corresponds to Clark's HAPINFERX algorithm. Each variation was run with 1000 independent iterations (that is, 1000 different sample paths *sensu* Orzack). In each case, correctly inferred genotypes are arranged in descending order on the left (in grey) and the incorrectly inferred genotypes on the right (in black). If a consensus threshold of 50% is adopted, the consensus prediction solution will harbour many incorrect inferrals in all six cases. However, with a threshold of 80%, there is still a remarkable proportion of incorrect ones. Even if a strict consensus (threshold = 100%) is applied, the solution contains error in four cases (c–f).

of Clark's method as described by Orzack *et al*. (2003) and Clark's HAPINFERX algorithm (each with 1000 independent iterations, that is, 1000 different sample paths) and three other popular algorithms (PHASE, Stephens *et al*. 2001; HAPLOTYPER, Niu *et al*. 2002; Arlequin-EM, Excoffier *et al*. 2005). These results reveal that, in general, algorithms with higher internal variability (e.g. several variations of Clark's method) have higher error rates than algorithms with lower internal variability (e.g. PHASE). This is true both for a random run and for Orzack *et al*.'s (2003) consensus method. Therefore, one must be careful when applying Orzack *et al*.'s consensus prediction procedure in such situations. Figure 1 illustrates the frequency distribution of the distinct genotypes for the same analysis and clearly shows that if a frequency threshold of 80% is adopted, those solutions with a frequency ≥ 80% will harbour a remarkable proportion of incorrect inferrals in all six cases. Even if a strict consensus (threshold = 100%) is applied, the solutions remain non-error-free in four out of six cases. As a result, although consensus prediction can identify the most consistent solutions of an algorithm for a data set, it cannot necessarily mirror the reliability of the solutions.

In brief, the two approaches, Orzack *et al*.'s (2003) consensus method and Huang *et al*.'s (2008) consensus vote approach, are philosophically different (although the quotations given in Orzack's comment suggest them to be similar). In contrast to what Orzack (2009) has claimed in his comment, Orzack *et al*. (2003) did not propose the consensus vote approach described by Huang *et al*. (2008), neither conceptually nor technically.

Actually, from a further scrutiny of the literature, it appears that the 'consensus vote' strategy based on multiple algorithms with different underlying statistical theories was conceptually first suggested by Niu (2004) as a possible way to increase the confidence of statistical inference results (p. 339). As a general remark on the conclusions of Orzack *et al*. (2003), the data sets they analysed are rather small. For example, the ApoE data involve only 9 (10) polymorphic sites, 17 distinct haplotypes and 47 ambiguous genotypes, and the Drysdale *et al*. (2000) data consist of only 13 single nucleotide polymorphisms, 12 distinct haplotypes, and 79 ambiguous genotypes. Both data sets represent fairly simple situations in haplotype reconstruction (cf. Huang *et al*.'s data: 70 polymorphic sites, 63 distinct haplotypes, 138 distinct genotypes). Therefore, some of the conclusions achieved in Orzack *et al*. (2003) merit further tests against more complex circumstances. Considering this and other issues discussed above, and since Huang *et al*. (2008) has not adopted Orzack *et al*.'s (2003) consensus strategy in their approach at all, we did not devote a section to discuss Orzack *et al*.'s (2003) results and conclusions in Huang *et al*. (2008) as the paper is already unusually long. However, we have cited Orzack *et al*. (2003) in our paper (p. 1943) for one of their observations that we believed to be important, although Orzack himself considers it as 'a small technical point' (Orzack 2009).

In summary, Orzack *et al*.'s (2003) consensus method and Huang *et al*.'s (2008) consensus vote approach are complementary. Given the existence of some great inconsistency with certain haplotyping algorithms, users are reminded that multiple independent runs should be performed to inspect internal algorithm variability. Orzack *et al*. (2003) set an excellent example to follow. When methodologically independent algorithms are considered, and the most consistent solutions for individual algorithms are included, approaches considering among-algorithm discordance, such as the consensus vote strategy explored by Huang *et al*. (2008), should substantially increase confidence for statistical haplotyping results.

Acknowledgements

We are grateful to an anonymous referee for his/her insightful comments and suggestions on an earlier version of the manuscript. We are in debt to Brent Emerson for comments and linguistic corrections on the manuscript and Nolan Kane for valuable suggestions. This research was supported by the Natural Science Foundation of China (grant nos 30870360 and 30730016), the CAS Knowledge Innovation Program (grant no. KZCX2-YW-428) and MOST grant no. 2006CB805901.

References

Araújo MB, New M (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42–47.

Bryant D (2003) A classification of consensus methods for phylogenetics. In: *Bioconsensus* (eds Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS), pp. 163–184. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence, Rhode Island.

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, **7**, 111–122.

Drysdale CM, McGraw DW, Stack CB *et al*. (2000) Complex promoter and coding region beta (2)-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proceedings of the National Academy of Sciences*, **97**, 10483–10488.

Excoffier L, Laval G, Schneider S (2005) ARLEQUIN version 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Huang ZS, Ji YJ, Zhang DX (2008) Haplotype reconstruction for scnp DNA: a consensus vote approach with extensive sequence data from populations of the migratory locust (*Locusta migratoria*). *Molecular Ecology*, **17**, 1930–1947.

Kuittinen H, Niittyvuopio A, Rinne P, Savolainen O (2008) Natural variation in *Arabidopsis Iyrata* vernalization requirement conferred by a *FRIGIDA* indel polymorphism. *Molecular Biology and Evolution*, **25**, 319–329.

Lewin B (1987) *Genes III*, 3rd edn. John Wiley & Sons, Chichester, UK.

Niu T (2004) Algorithms for inferring haplotypes. *Genetic Epidemiology*, **27**, 334–347.

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, **70**, 157–169.

Orzack SH (2009) The consensus approach to resolving heterogeneity among haplotype inferrals: a comment on Huang *et al*. (2008). *Molecular Ecology*, **18**, 1553–1555.

Orzack SH, 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*, **165**, 915–928.

Sabbagh A, Darlu P (2005) Inferring haplotypes at the NAT2 locus: the computational approach. *BMC Genetics*, **6**, 30.

Saito K, Miyake S, Moriya H *et al*. (2003) Detection of the four sequence variations of *MDR1* gene using TaqMan MGB probe based real-time PCR and haplotype analysis in healthy Japanese subjects. *Clinical Biochemistry*, **36**, 511–518.

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

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Xu CF, Lewis K, Cantone KL *et al*. (2002) Effectiveness of computational methods in haplotype prediction. *Human Genetics*, **110**, 148–156.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.

Zhang J, Rowe WL, Struewing JP, Buetow KH (2002) HAPSCOPE: a software system for automated and visual analysis of functionally annotated haplotypes. *Nucleic Acids Research*, **30**, 5213–5221.

doi: 10.1111/j.1365-294X.2009.04147.x