MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 1066–1070 doi: 10.1111/j.1755-0998.2010.02843.x

COMPUTER PROGRAM NOTE

CVhaplot: a consensus tool for statistical haplotyping

ZU-SHI HUANG* and DE-XING ZHANG*†
**State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China, †Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*

Abstract

Haplotypes contain genealogical information and play a prominent part in population genetic and evolutionary studies. However, haplotype inference is a complex statistical problem, showing considerable internal algorithm variability and among-algorithm discordance. Thus, haplotypes inferred by statistical algorithms often contain hidden uncertainties, which may complicate and even mislead downstream analysis. Consensus strategy is one of the effective means to increase the confidence of inferred haplotypes. Here, we present a consensus tool, the CVhaplot package, to automate consensus techniques for haplotype inference. It generates consensus haplotypes from inferrals of competing algorithms to increase the confidence of haplotype inference results, while improving the performance of individual algorithms by considering their internal variability. It can effectively identify uncertain haplotypes potentially associated with inference errors. In addition, this tool allows file format conversion for several popular algorithms and extends the applicability of some algorithms to complex data containing triallelic polymorphic sites. CVhaplot is written in PERL and freely available at http://www.ioz.ac.cn/department/agripest/group/zhangdx/CVhaplot.htm.

Keywords: algorithm, ambiguous genotype, consensus vote, haplotype inference, haplotyping uncertainty, inconsistency

Received 25 September 2009; revision received 18 November 2009, 5 January 2010; accepted 16 January 2010

Introduction

A haplotype refers to any distinct set of nucleotide sites linked on the same chromosome that are inherited together as a unit. Thus, haplotypes contain genealogical information and are the key components of contemporary evolutionary and population genetic studies. As a consequence, statistical inference of haplotypes from population data has gained much attention in recent years, and a large number (>40) of statistical algorithms have been developed (for review, see Salem *et al.* 2005). The major driving force for such effort is the effectiveness of statistical haplotyping in reducing cost and labour in large-scale studies. However, haplotype inference is a complex statistical problem, showing considerable internal algorithm variability and among-algorithm discordance (Huang *et al.* 2009). Given the complexity of genetic data and the simplification of assumptions of statistical models, no single algorithm can closely approach the truth in every circumstance. For example, it is not infrequent to observe that the best solution inferred by an algorithm (i.e. the one being assigned a confidence probability close or equal to one) is not the true solution, or more than one inferred haplotype pairs have similar probability to be correct. Such hidden uncertainty may complicate and even mislead downstream analysis using the inferred haplotypes (Lin & Huang 2007). Therefore, a great effort is needed to increase the confidence of haplotype inference results.

Consensus strategy is one of the effective means for improving the performance of haplotype inference (Orzack *et al.* 2003; Niu 2004; Scheet & Stephens 2006; Kääriäinen *et al.* 2007; Huang *et al.* 2008). By combining matching inferrals from the same approach or competing approaches into a consensus solution, consensus

Correspondence: De-Xing Zhang, Fax: (+86) 10 6480 7232;
E-mail: dxzhang@ioz.ac.cn

techniques can filter out a great amount of noise signals in individual inferences. It has been shown by experimental verification that uncertain haplotypes associated with noise signals can all be inference errors (Huang *et al.* 2008). Basically, there are two complementary approaches in haplotype reconstruction that apply the consensus strategy. One approach limits its effort in the internal variability of individual algorithms (e.g. Orzack *et al.* 2003); the other approach places its emphasis on the discordance among algorithms (e.g. Huang *et al.* 2008). Approaches combining the two consensus strategies should substantially increase the confidence for statistical haplotyping results (Huang *et al.* 2009).

CVhaplot has been developed to realize such combined approach. It automates the consensus vote (CV) approach explored in Huang *et al.* (2008) (which evaluates the among-algorithm discordance in haplotype inference) while also considering the internal variability of individual algorithms.

Implementation

CVhaplot has the following features: (i) File conversion: it can convert raw genotype data into input files of several algorithms, and directly output haplotype data after the analysis. This avoids the tedious and error-prone manual operation, being particularly convenient for large data set; (ii) Data recoding: it recodes each triallelic site as two biallelic ones to extend the applicability of several algorithms that were developed for analysing biallelic data; (iii) Consistency testing: it can evaluate the internal variability of individual algorithms using several diagnostic indices (Huang *et al.* 2008); (iv) CV solution inference: it infers CV solution by considering the confidence probability and among-algorithm discordance of the inferrals. Uncertain haplotypes are identified according to among-algorithm discordance; and (v) Data inspection: after the CV analysis, it identifies samples that show any mismatch between the inferred and original genotypes.

Workflow

CVhaplot consists of three Perl scripts: trans.pl, consistency.pl and CV.pl. A typical analysis involves three steps (Fig. S1). The first step is file conversion using the script trans.pl. This script also generates three batch files that allow automatically launching the relevant statistical haplotyping programs and performing multiple independent iterations at the users' disposal (independent iterations refer to different runs of an algorithm with different seed number or random input order). Second, the script consistency.pl examines the internal variability of individual algorithms [excepting GCHAP (Thomas 2003) and GERBIL (Kimmel & Shamir 2005), as they always generate unique solutions]. In addition, it helps to produce a HAPINFERX (Clark 1990) solution with high accuracy and consistency by generating an ensemble (consensus) from those independent iterations whose NDH (number of distinct haplotypes) values are among the smallest (Figs 1, 2 and S4) (see Orzack *et al.* 2003 for a similar approach). Finally, CV.pl performs the CV analysis. In addition to generate the CV solution, it reports the consensus vote information (Table 1 gives an example), including the confidence probability of inferrals from each algorithm, the vote number of the CV solution, the discordance among algorithms, etc. This helps users to have a closer inspection of haplotype uncertainty in the CV solution. This script also allows the user to control which solution of an algorithm is used in the analysis.

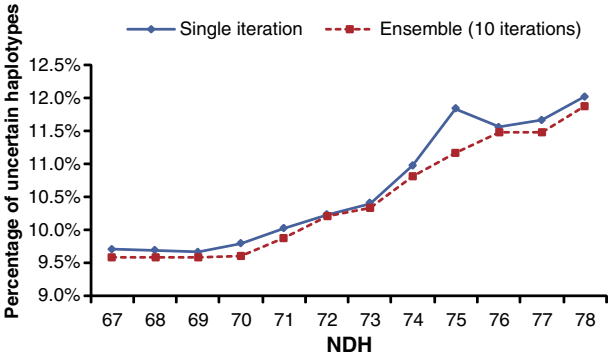

Fig. 1 Frequency distribution of uncertain haplotypes in the consensus vote (CV) solution on functions of the number of distinct haplotypes (NDH) of HAPINFERX inference. The ensemble summarized from as few as ten independent HAPINFERX iterations substantially (and sufficiently) improves the inference performance compared to using only a single iteration (see Fig. S4 for details). The smaller the NDH values of independent HAPINFERX iterations are, the more accurate the ensemble is (Spearman rank correlation coefficient, 0.993 for individual error, d.f. = 11, $P < 0.001$). Each data point represents 100 CV analyses of scnpc76 data.

Data conversion

Genotypic sequence data should be in sequential PHYLIP format as an input file for CVhaplot. The Perl script trans.pl can reformat the data into the input file formats of the following programs: PHASE (Stephens *et al.* 2001), HAPLOTYPER (Niu *et al.* 2002), HAPLOREC (Eronen *et al.* 2006), ARLEQUIN-EM (Excoffier *et al.* 2005), GCHAP, GERBIL, and HAPINFERX. Among them, HAPLOTYPER, GCHAP and GERBIL require to code triallelic sites as biallelic ones.

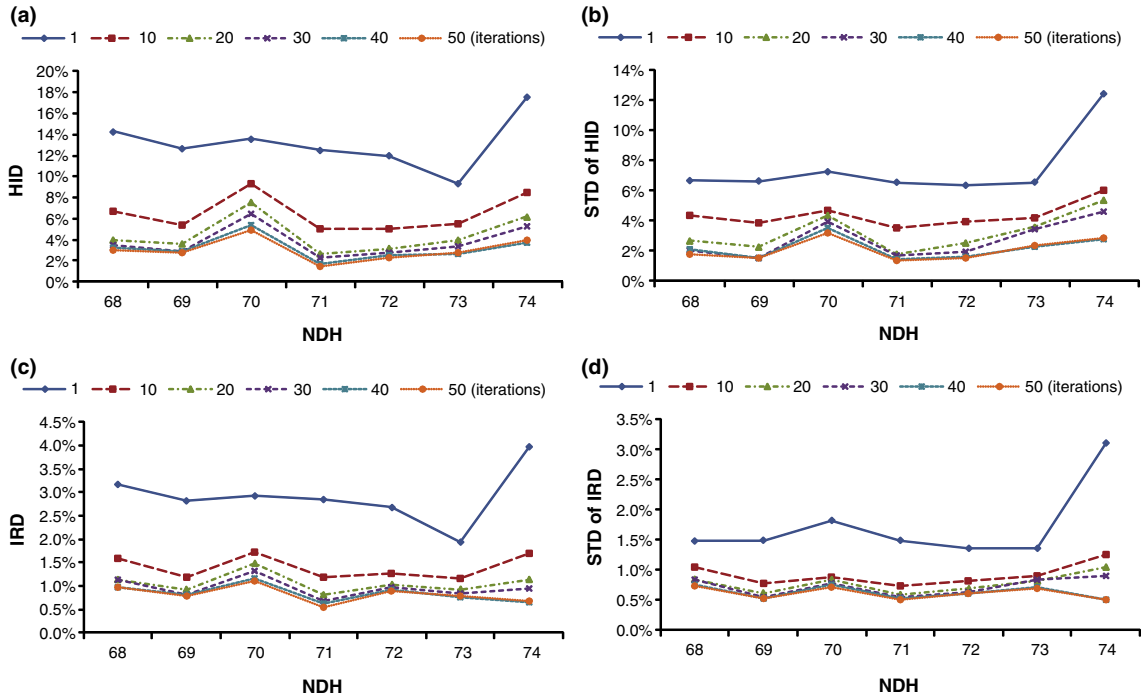

Fig. 2 Influence of iteration number on the consistency of the ensemble solutions of HAPINFERX inference. The consistency was examined by four criteria: (a) haplotype-inferring discrepancy (HID), (b) standard deviation (STD) of haplotype-inferring discrepancy (c) individual-resolving discrepancy (IRD), and (d) standard deviation of individual-resolving discrepancy. NDH is the number of distinct haplotypes in a solution. HID is the discrepancy of distinct haplotypes between two solutions. IRD is the proportion (number) of individuals whose genotypes were resolved differently between two solutions. In general, the larger the number of iterations is used, the more consistent the ensemble solution is. Each data point represents 100 independent single iterations or ensemble solutions.

Table 1 Summary of the output information of a consensus vote (CV) analysis using six algorithms

Number	Sample	Confidence probabilities*	Solution status†	Vote number of CV solution	Algorithm assignment group‡
1	IOZLm1	+ − − + − −	3 solutions	2.4 vote	(ARLEQUINEM HAPLOTYPER HAPINFERX) (GCHap PHASE) (HAPLOREC)
2	IOZLm2	+ + + + + −	Fully supported	5.7 vote	(ARLEQUINEM GCHap HAPINFERX HAPLOTYPER HAPLOREC PHASE)
3	IOZLm3	+ + + + − −	3 solutions	3 vote	(HAPLOREC HAPLOTYPER GCHap) (ARLEQUINEM HAPINFERX) (PHASE)
4	IOZLm4	+ + + + + +	Fully supported	6 vote	(ARLEQUINEM GCHap HAPINFERX HAPLOTYPER HAPLOREC PHASE)
5	IOZLm5	+ + + + + −	Fully supported	5.7 vote	(ARLEQUINEM GCHap HAPINFERX HAPLOTYPER HAPLOREC PHASE)
6	IOZLm6	+ + + + + +	Fully supported	6 vote	(ARLEQUINEM GCHap HAPINFERX HAPLOTYPER HAPLOREC PHASE)
7	IOZLm7	+ + + + + +	Homozygote	NA	NA

NA, not applicable.
*'+' denotes high probability, i.e. the probability is higher than the threshold value of that algorithm; '−' the low probability. Low probability leads to weighting down the inferral of an algorithm. The listing order of algorithms is given in the output file.
†The number of solutions among algorithms for each sample. 'Homozygote' means the sample is homozygous with clear haplotype phase. 'Fully supported' means one single solution approved by all algorithms was obtained. '3 solutions' means three different solutions were obtained, each being supported by one or more algorithms.
‡Refers to a set of algorithms that gave identical inference results. Here, algorithms of the same group are bracketed together, and those underlined are algorithms whose solution is identical to the CV solution.

Consensus rules

The following weighting rules are employed in CVhaplot besides the inferring rules described in Huang *et al.* (2008). First, inferrals are weighted according to their confidence probability, which is a probability measure of haplotype uncertainty computed by individual algorithms. If the probability estimate is lower than a threshold value (i.e. 0.7), the inferral will be assigned a low weight (e.g. 0.7), and otherwise a high weight (i.e. 1). Second, the cumulative vote value for each inferral is used as the indicator of the reliability of the inferral. Third, the CV solution only consists of inferrals with the highest vote value.

Performance

The performance of the program was tested extensively using sequence data from *Locusta migratoria* (Huang *et al.* 2008). The data set is comprised of 1052 chromosomes and free from genotyping error, bearing 70 polymorphic sites and 63 distinct haplotypes. A robust performance was confirmed by comparing inference results of CVhaplot with those of manual analysis and the true haplotype data obtained from laboratory experiment. Figures S2 and S3 demonstrate that compared to PHASE, the CV approach generally shows a higher accuracy, although it includes some algorithm with high internal variability (e.g. HAPINFERX). Most data points in Fig. S2 are lower than 2.9%, which is the average error rate of 100 PHASE iterations (Fig. S3).

Compared to the preliminary version reported in Huang *et al.* (2008), the present version of CVhaplot (ver. 2.01) has implemented several important functions and thus greatly improved its performance and flexibility (see the program manual for details). Here, in the following paragraphs, we discuss two key aspects.

Internal algorithm variability was considered

CVhaplot now has options allowing users to run multiple independent iterations for each algorithm to examine their internal variability on the genotype data. Among the algorithms explored in CVhaplot (i.e. HAPINFERX, PHASE, HAPLOTYPER, ARLEQUIN-EM, HAPLOREC, GCHAP and GERBIL), the algorithm HAPINFERX usually displays the highest internal variability (Orzack *et al.* 2003; Huang *et al.* 2008, 2009), that is, different HAPINFERX runs tend to give quite different inference results. This leads to higher error rate in general (Huang *et al.* 2009).

To further examine this issue, 100 000 iterations of HAPINFERX with different input order were performed for the scnpc76 data of *Locusta migratoria*. It is known that the NDH value of a solution is a good indicator of the accuracy of the solution for DNA regions with no (or weak) recombination (Orzack *et al.* 2003; Huang *et al.* 2008, 2009). Short nuclear DNA fragments of a few hundreds base pairs employed in population genetic studies are often assumed to be free from recombination, but this needs verification. Therefore, iterations of HAPINFERX were grouped into different categories according to their NDH values, and then the CV analyses performed separately for different categories. Figure 1 shows that the uncertain haplotypes identified by CVhaplot could increase by more than 20% when the HAPINFERX iterations with large NDH values were used in the CV analysis. Clearly, excessive inference errors produced by individual algorithm can significantly affect the overall performance of the CV approach and hence require constant vigilance.

Ensemble solution was introduced

Although HAPINFERX manifests high internal algorithm variability, it remains an important algorithm to be included in the CV analysis because of its theoretical and technical distinctness (Huang *et al.* 2008) and popularity. The deficiency of HAPINFERX can be overcome by choosing a HAPINFERX solution with high accuracy and consistency, e.g. an appropriate ensemble solution. Here, ensemble solution refers to a consensus solution summarized from multiple independent iterations of an algorithm. Figure 2 shows that ensembles from independent iterations can efficiently reduce the internal variability of HAPINFERX inferences, being more consistent than single iterations. Ensembles summarized from as few as ten HAPINFERX iterations (see the following paragraphs for more details) can substantially improve the performance of the CV analysis, reducing considerably the proportion of uncertain haplotypes in the CV solution (Fig. S4A). It also reduces the variance of the CV approach (Fig. S4B). Note that HAPINFERX ensemble solutions with high accuracy can adequately remove solutions with large NDH values, thus effectively reducing uncertain haplotypes associated with high internal variability (Fig. 1). Therefore, we recommend that users should first identify a good ensemble solution from independent HAPINFERX iterations and then use it in the CV analysis. This technique is equally applicable to any other algorithms if internal variability becomes a concern (see Orzack *et al.* 2003).

In practice, a rather robust frequency distribution of NDH can be obtained from ≥100 independent HAPINFERX iterations (Fig. S5). Therefore, 100 independent iterations are generally sufficient for producing a HAPINFERX ensemble solution with high accuracy. This is achieved in two steps. First, ten or more iterations with

the smallest NDH values are chosen from the 100 iterations; then they are used to generate the more accurate ensemble solution by simple consensus technique.

Conclusions

CVhaplot effectively automated the consensus vote approach for haplotype inference introduced in Huang *et al*. (2008), allowing sensible identification of uncertain haplotypes potentially associated with inference errors. The present version (ver. 2.01) fully considered internal algorithm variability and among algorithm discordance, and employed the technique of ensemble solution to further improve its overall performance. In addition, it also facilitates file format conversion for several popular algorithms, and extends the applicability of some algorithms to complex data containing triallelic polymorphic sites. Given the importance of haplotype determination in population genetic and evolutionary studies using nuclear DNA sequences (Zhang & Hewitt 2003), this tool should promote the employment of nuclear DNA markers in these research areas.

Acknowledgements

We thank J.S. Liu, Z.S. Qin and D. Clayton for their valuable suggestions of coding triallelic sites as biallelic sites, B.G. Xie for his technical supports in programming, Julie B. Hébert for comments on the program package, and Vincent Castric, the subject editor, for valuable suggestions. This work was supported by the Natural Science Foundation of China (grant nos 30730016, 30870360), the Knowledge Innovation Program of the Chinese Academy of Sciences (grant no. KZCX2-YW-428), and the Ministry of Science and Technology of China (grant no. 2006CB805901).

References

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

Eronen L, Geerts F, Toivonen H (2006) HaploRec: efficient and accurate large-scale reconstruction of haplotypes. *BMC Bioinformatics*, **7**, 542.

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 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.

Huang ZS, Ji YJ, Zhang DX (2009) Internal algorithm variability and among-algorithm discordance in statistical haplotype reconstruction. *Molecular Ecology*, **18**, 1556–1559.

Kääriäinen M, Landwehr N, Lappalainen S, Mielikäinen T (2007) *Combining Haplotypers*. Technical Report C-2007-57, Department of Computer Science, University of Helsinki.

Kimmel G, Shamir R (2005) GERBIL: genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 158–162.

Lin DY, Huang BE (2007) The use of inferred haplotypes in downstream analysis. *American Journal of Human Genetics*, **80**, 577–579.

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, 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.

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.

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. *American Journal of Human Genetics*, **78**, 629–644.

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.

Thomas A (2003) GCHap: fast MLEs for haplotype frequencies by gene counting. *Bioinformatics*, **19**, 2002–2003.

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

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 The flow chart of CVhaplot analysis.

Fig. S2 Error rate of the CV solution on functions of the number of distinct haplotypes (NDH) in HAPINFERX inferences.

Fig. S3 Error rate of PHASE inference measured by the occurrence of incorrect inferrals (haplotypes) in 100 PHASE iterations.

Fig. S4 Importance of using HAPINFERX ensemble solution in CV analysis.

Fig. S5 Frequency distribution of the number of distinct haplotypes (NDH) in HAPINFERX inference.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.