Molecular Ecology (2008) **17**, 1930–1947 doi: 10.1111/j.1365-294X.2008.03730.x

Haplotype reconstruction for scnp DNA: a consensus vote approach with extensive sequence data from populations of the migratory locust (*Locusta migratoria*)

ZU-SHI HUANG,*† YA-JIE JI* 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, †Graduate University of the Chinese Academy of Sciences, Beijing 100049, China, ‡Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*

Abstract

Single copy nuclear polymorphic (scnp) DNA is potentially a powerful molecular marker for evolutionary studies of populations. However, a practical obstacle to its employment is the general problem of haplotype determination due to the common occurrence of heterozygosity in diploid organisms. We explore here a 'consensus vote' (CV) approach to this question, combining statistical haplotype reconstruction and experimental verification using as an example an indel-free scnp DNA marker from the flanking region of a microsatellite locus of the migratory locust. The raw data comprise 251-bp sequences from 526 locust individuals (1052 chromosomes), with 71 (28.3%) polymorphic nucleotide sites (including seven triallelic sites) and 141 distinct genotypes (with frequencies ranging from 0.2 to 25.5%). Six representative statistical haplotype reconstruction algorithms are employed in our CV approach, including one parsimony method, two expectation–maximization (EM) methods and three Bayesian methods. The phases of 116 ambiguous individuals inferred by this approach are verified by molecular cloning experiments. We demonstrate the effectiveness of the CV approach compared to inferences based on individual statistical algorithms. First, it has the unique power to partition the inferrals into a reliable group and an uncertain group, thereby allowing the identification of the inferrals with greater uncertainty (12.7% of the total sample in this case). This considerably reduces subsequent efforts of experimental verification. Second, this approach is capable of handling genotype data pooled from many geographical populations, thus tolerating heterogeneity of genetic diversity among populations. Third, the performance of the CV approach is not influenced by the number of heterozygous sites in the ambiguous genotypes. Therefore, the CV approach is potentially a reliable strategy for effective haplotype determination of nuclear DNA markers. Our results also show that rare variations and rare inferrals tend to be more vulnerable to inference error, and hence deserve extra surveillance.

Keywords: ambiguous genotype, consensus vote, flanking region of nuclear microsatellite DNA, *Locusta migratoria*, statistical haplotype reconstruction

Received 9 November 2007; revision received 16 January 2008; accepted 4 February 2008

Introduction

Molecular population genetic and evolutionary studies increasingly rely on genealogical data for a better understanding of the effects of both historical evolutionary processes and contemporary ecological factors on populations and species. In addition to mitochondrial DNA markers, which have been popularly employed in metazoan animals in particular, single copy nuclear polymorphic (scnp) DNA is becoming the marker of choice for more comprehensive investigations (Zhang & Hewitt 2003). Scnp DNA refers to any DNA segment that is present as a

Correspondence: De-Xing Zhang, Institute of Zoology, Chinese Academy of Sciences, Datun Road, Chaoyang District, Beijing 100101, China. Fax: (+86) 10 64807232; E-mail: dxzhang@ioz.ac.cn

single copy in the haploid genome and polymorphic in the population. For diploid organisms, there is a pair of homologous alleles at any scnp DNA locus, with the two sequence copies being either identical (homozygous) or nonidentical (heterozygous). The molecular description of the particular DNA segments at a locus in an individual is called its 'genotype', with each copy form being termed a 'haplotype'. From a population genetic viewpoint, the term haplotype refers collectively to a set of identical alleles in populations; from a genomic viewpoint, it refers to a distinct set of nucleotide sites linked on the same chromosome and inherited together in meiosis. Because it has the highest resolution and contains genealogical information, haplotype sequence has been the key ingredient in many fine-scale analyses in molecular evolutionary and population genetic studies. However, haplotype determination is a challenging task when using scnp DNA in diploid organisms, because for multisite heterozygotes (i.e. more than one heterozygous site along an scnp DNA sequence), current sequencing techniques provide ambiguous genotype data and haplotypes are not directly discernible. For example, for the following ambiguous genotype data from a diploid organism at two nucleotide sites — RY (here R signifies an A or a G, Y a C or T), there are two possible haplotype assignments (i.e. either AC and GT, or AT and GC). The more heterozygous sites occurring in an scnp DNA sequence, the more complicated is the haplotype assignment (an exponential increase of possible solutions). Note that both homozygotes and one-site heterozygotes have clear haplotype phases.

A number of experimental haplotyping methods have been developed to circumvent the difficulty (for review, see Zhang & Hewitt 2003). However, all these methods have some important limitations in efficiency, cost and labour, or safety when applied in large-scale studies. Each also has its individual advantages, and recent attempts show some promise for a reasonably high throughput (Tost *et al*. 2002; Hurley *et al*. 2005). Consequently, these experimental methods have not been widely employed. Alternatively, statistical methods for inferring and reconstructing haplotype phase have blossomed in recent years largely as a result of the increasing demands of genetic and genomic studies such as the Human International HapMap Project (International HapMap Consortium 2005), with the attraction that they are more practical and cost-effective compared to experimental approaches. There are two classes of statistical methods, namely haplotype assembly and haplotype reconstruction. Haplotype assembly, which is also called single individual haplotyping, is based on aligned sequences for the retrieval of a pair of haplotypes (Lancia *et al*. 2001; Lippert *et al*. 2002). Haplotype reconstruction infers haplotypes from unrelated genotype data collected from samples from natural populations or related genotype data from families and/or pedigrees. Although pedigree information can improve haplotype reconstruction (Rohde & Fuerst 2001; Becker & Knapp 2002; Schaid 2002; Lin *et al*. 2004), such information is rarely available for natural populations. Hence, we focus here on population-based haplotyping. Numerous statistical algorithms have been published for haplotype reconstruction during the past two decades, which can be approximately divided into parsimony-based and likelihood-based methods.

Parsimony-based methods. In his seminal paper, Clark (1990) proposed a rule-based haplotype-subtraction algorithm for finding the solution with the fewest orphans and fewest anomalous matches. The modified program, HAPINFERX, is computationally very efficient and used to be one of the most popular algorithms for haplotype reconstruction (e.g. Clark *et al*. 1998; Olsen & Schaal 1999; Antunes *et al*. 2002; Bartish *et al*. 2006). Recently, various extensions of parsimony algorithms have been developed (Gusfield 2003; Wang & Xu 2003; Lancia *et al*. 2004; Huang *et al*. 2005; Li *et al*. 2005; Brown & Harrower 2006), including those based on the 'perfect phylogeny' assumption (e.g. Gusfield 2002; Halperin & Eskin 2004). However, the pure parsimony approach for finding a minimum set of haplotypes that explains a given set of genotypes has been shown to be an NP-hard (and even APX-hard) problem (i.e. not solvable in polynomial time; Gusfield 2001; Lancia *et al*. 2004; Sharan *et al*. 2006).

Likelihood-based methods. Several expectation–maximization (EM) methods were reported independently by three research groups in 1995 (Excoffier & Slatkin 1995; Hawley & Kidd 1995; Long *et al*. 1995). They estimate population haplotype probabilities under the maximum likelihood (ML) principle based on assumption of Hardy–Weinberg equilibrium (HWE), so the resolution will maximize the probability of the observed genotype data. Subsequently, diverse EM programs (Hoehe *et al*. 2000; Clayton 2001; Qin *et al*. 2002; Li *et al*. 2003; Thomas 2003a, b; Excoffier *et al*. 2005; Kimmel & Shamir 2005; Eronen *et al*. 2006; Scheet & Stephens 2006; Landwehr *et al*. 2007) were developed to improve the computational efficiency by employing various more efficient techniques (e.g. considering only subsets of the polymorphic sites at a time and/or minimizing the number of potential haplotypes that must be considered). For example, GCHAP (Thomas 2003a, b) can quickly obtain maximum likelihood estimates (MLE) of haplotype frequencies using the gene counting method, detecting and eliminating all subhaplotypes whose MLE frequency from one round of EM is exactly zero at the step of progressively building haplotypes.

Bayesian methods, as alternative likelihood-based approaches, incorporate different models and prior assumptions, thus extending their application. Numerous Bayesian methods are available (Stephens *et al*. 2001; Lin *et al*. 2002;

Box 1 Abbreviations and terms

Ambiguous individual: any multisite heterozygote whose phase is unresolved.
Ambiguous genotype: the genotype of any ambiguous individual.
CV: consensus vote.
CV category: vote category established in the consensus vote approach where the multisite heterozygotes receiving the same number of votes were classified into the same category. Six categories were established in this study with I–IV having the number of votes 6, 5, 4, 3 and 2 (pooled), respectively, and H for homozygotes and S for one-site heterozygotes.
Genotyping error: refers to nucleotides originally scored from the direct sequencing of PCR products that were found to be in error based on subsequent sequencing of multiple clones of (independent) PCR products.
HID: haplotype-inferring discrepancy, i.e. the discrepancy of distinct haplotypes between two solutions.
HWD: Hardy–Weinberg disequilibrium.
IRD: individual-resolving discrepancy, i.e. the proportion (number) of individuals whose genotypes were resolved differently between two solutions.
Lost haplotypes: the distinct haplotypes that an approach fails to infer in its solution (i.e. the haplotype that was absent in the solution of an approach compared to the true solution).
NDH: number of distinct haplotypes in a solution.
NLCP: number of individuals whose haplotype pair is of *l*ow *c*onfidence *p*robability in a solution.
PL: partition–ligation.
Scnp: single copy nuclear polymorphic (DNA).
Uncertain individual: any multisite heterozygote without the highest consensus votes (here votes < 6).
Uncertain genotype: the genotype of any uncertain individual.
Variation: refers to the nucleotide state at a polymorphic site the frequency of which is lower than the most frequent state in population.

Niu *et al*. 2002; Excoffier *et al*. 2003; Eronen *et al*. 2004; Greenspan & Geiger 2004; Xing *et al*. 2006; Zhang *et al*. 2006; Ayers *et al*. 2007); among them, the coalescent-based PHASE (Stephens *et al*. 2001; Stephens & Donnelly 2003) currently appears to be most widely used (e.g. Ryynanen & Primmer 2004; Sotka *et al*. 2004; Taylor & Hellberg 2006). The state-of-the-art PHASE program is based on the important assumption of conditional distribution of the inferred haplotypes, that is, the next inferred haplotype tends to be either identical or similar to a haplotype that has already been observed or inferred. Another Bayesian program, HAPLOTYPER (Niu *et al*. 2002), employs the novel partition-ligation (PL) technique. The PL technique first resolves haplotypes within smaller segments of consecutive sites, and then links them into a complete haplotype either hierarchically or progressively. It is remarkably valuable because of its great computational efficiency and has been incorporated into other likelihood-based methods [e.g. PHASE and HAPLOREC (Eronen *et al*. 2004)].

Clearly, the boom in statistical methods for haplotype reconstruction should drive wide implementation of scnp markers in genetic studies of populations. Nevertheless, we need to answer a key question before any statistical haplotyping method is implemented: how can we be confident that the results deduced by statistical methods have acceptable reliability to safeguard subsequent data analyses, given the complexity of population genetic data? One common way to address this question is to combine statistical and experimental analyses to verify all inferred haplotypes and phased genotypes. Such combined approaches have been rather limited to date due to cost and labour (e.g. Clark *et al*. 1998; Antunes *et al*. 2002; Bartish *et al*. 2006). Theoretically, there is another more effective means of addressing this question. For statistical haplotype inference, every deduced haplotype and phase assignment is only a statistical solution, and the estimates of confidence suggested by such analyses can serve only as a reference (Xu *et al*. 2002; Sabbagh & Darlu 2005); thus, all inferrals bear some uncertainty. Also, because they differ in the underlying theories and techniques employed, different algorithms have different sensitivity to various genetic factors (Niu *et al*. 2002; Zhang *et al*. 2002; Niu 2004), and hence perform differently. Thus, solutions shared by independent statistical algorithms should be a good indication of their reliability if biases of different algorithms are independent. Therefore, if we compare the inference results of multiple algorithms on a given data set, the haplotypes and phase assignments that are completely agreed among all algorithms (thus with the highest consensus votes) should be most likely to be reliable; similarly, those with lower consensus votes should be less reliable, and hence need further experimental verification. We termed such an analysis 'consensus vote' (CV) approach (Box 1). If proven effective, such an approach would allow us to confine experimental verification to those individuals whose phase assignment really deserves confirmation while maintaining the quality of statistical inference.

We report here such a study on a large data set with no prior phase information (except for homozygotes and

one-site heterozygotes) from the migratory locust (*Locusta migratoria*), characterized using an scnp DNA marker. The genotype data from 526 locusts sampled across a broad geographical range were analysed employing six statistical haplotype reconstruction methods, including one parsimony method (HAPINFERX), two EM methods [GCHAP and ARLEQUIN 3.11-EM (Excoffier *et al*. 2005)], and three Bayesian methods (PHASE 2.1.1, HAPLOTYPER 1.0 and HAPLOREC 1.0). The inferred haplotypes and phase assignments were then experimentally verified by cloning. We aimed to (i) evaluate the performance of the CV approach and the six algorithms on extensive population genetic data, (ii) most importantly, examine the effectiveness of the CV approach for guiding experimental verification of uncertain inferrals, and (iii) infer some rules of statistical haplotype inference from large genetic data sets from natural populations.

Materials and methods

Population sampling

Thirty-one populations of the migratory locust were sampled during 1998–2005, with 28 of them from China covering the main distributional range of this insect in the country, and one each from Japan, France and Eritrea. These samples are ascribed to five *Locusta migratoria* subspecies (i.e. *L. m. manilensis*, *L. m. migratoria*, *L. m. tibetensis*, *L. m. cinerascens* and *L. m. migratorioides*). Locusts were randomly collected in the field, and preserved in absolute alcohol at 4–6 °C. Thirteen to 30 samples of each population were sequenced in this study with an scnp marker. Detailed information on the samples is given in Table S1 (Supplementary material).

Laboratory techniques

Genomic DNA was extracted from muscle of a hind leg using a modified phenol-chloroform method (Zhang & Hewitt 1998). The screening process of scnp DNA markers was similar to the protocol for development of an SNPSTR system by Mountain *et al*. (2002), but here we focused on the flanking regions of the highly variable microsatellite loci developed in our laboratory (Zhang *et al*. 2003). Polymerase chain reaction (PCR) primers were designed in the flanking regions of the eight microsatellite loci of the migratory locust using Oligo 6 (National Biosciences, Inc.) and synthesized by Sangon Biotech. Amplification from genomic DNA was performed using *Taq* DNA polymerase (HuaMei Biotech) on Perkin Elmer GeneAmp 9700 systems. Specific PCR products were obtained from the flanking regions of seven microsatellite loci. Pilot tests were carried out by sequencing 24 locusts at each of the seven loci on an ABI PRISM 3100 automated sequencer (Applied Biosystems). The downstream flanking region of the locus LmIOZc76 (hereafter referred to scnpc76) was chosen for the subsequent large-scale study because it is highly polymorphic, and most importantly, it lacks indel mutation in the region we investigated.

For each of the 526 individuals, a ~300 bp fragment of scnpc76 was amplified using the primers C76F144 (5′-TATCC TAAGCGTCTTCATCTC-3′) and C76B436 (5′-TGTTTGCCCG TCTTTGGTTATT-3′). PCR was performed in a total volume of 30μL containing 1× *Taq* reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTPs, 0.3 μM of each primer, 1.35 U of *Taq* DNA polymerase, about 50 ng of the template DNA. The temperature profile was as follows: 94 °C for 4 min, followed by 10 cycles of 95 °C for 15 s, 51 °C for 30 s and 72 °C for 10 s, then 27 cycles of 89 °C for 15 s, 49 °C for 30 s and 72 °C for 10 s, and finally an extension at 72 °C for 2 min. PCR products were checked on a 1.5% agarose gel containing 0.5 μg/mL of ethidium bromide. Purified PCR products were directly sequenced using the PCR primers with the ABI BigDye Terminators Cycle Sequencing Kit (version 2.0). Genotype data were carefully inspected at least three times by eye using SEQUENCHER (Gene Codes). Sequences immediately next to the sequencing primer often are of lower quality and were excluded. Heterozygous sites were identified as sites having double peaks clearly different from the baseline signal, and missing or over-accounting for such sites constitutes a potential source of error. When there was any doubt, the individual was then resequenced or sequenced from the complementary strand (about a quarter of multisite heterozygotes were sequenced from both strands in the initial analysis).

In total, 116 multisite heterozygotes (with 85 distinct genotypes) were verified by molecular cloning based on the results of the CV approach (see below). Two homozygotes were included as controls for experimental error (the other distinct homozygotes, they were resequenced at least once from independent genomic PCR, and no error or allele-drop was observed). High-fidelity PCR was performed in a total volume of 30 μL containing 1× *Pfu* reaction buffer (Stratagene), 0.2 mM of each dNTPs, 0.3 μM of each primer, 0.75 U of *Pfu* DNA polymerase (Stratagene), about 50 ng of the template DNA. The temperature profile was as follows: 94 °C for 2 min, followed by 10 cycles of 95 °C for 30 s, 51 °C for 30 s and 72 °C for 1 min, then 25 cycles of 89 °C for 30 s, 49 °C for 30 s and 72 °C for 1 min, and finally, an extension at 72 °C for 4 min. Purified *Pfu* PCR products were ligated into *Eco*RV-cut pBluescript II SK(+) phagemid (Stratagene) using T4 DNA ligase (Promega), and *Escherichia coli* TOP10 competent cells (Tiangen Biotech) were transformed according to the manufacturer's instructions and plated on standard white–blue selection Luria-Bertani (LB) plates. Colonies were transferred into 200 μL LB liquid medium and cultured overnight at 37 °C with shaking. Recombinant phagemid DNA was

extracted using an alkali method (Sambrook *et al*. 1989) and dissolved in 70 μL 10 mM Tris HCl (pH 8.0). Four microlitres of extract was used for subsequent PCR using the primers C76F144 and C76B436. Then, purified PCR products were sequenced as above. To control for allele-drop effect so that both alleles (haplotypes) were obtained and genotyping errors verified, three to seven independent clones were sequenced for each genomic *Pfu*-PCR assay, and about one-third of individuals were cloned twice from independent genomic PCR. In several cases, apparently 'recombinant' artefact sequences were observed in the replicate clones analysed.

Specific terms and basic summary statistics

Several specific terms employed in this study are defined as follows (see also Box 1). 'Variation' refers to the nucleotide state at a polymorphic site the frequency of which is lower than the most frequent state in population. 'Ambiguous individual' refers to any multisite heterozygote whose phase is unresolved, and the corresponding genotype is referred to as 'ambiguous genotype'. 'Uncertain individual' refers to any multisite heterozygote without the highest consensus votes, and the corresponding genotype is referred to as 'uncertain genotype'. 'CV category' refers to the vote category established in the CV approach containing the multisite heterozygotes receiving the same number of votes (see below). 'CV solution' refers to the solution inferred by the CV approach, and 'true solution' to the solution determined by molecular cloning experiments. 'Lost haplotypes' refer to the distinct haplotypes that were absent in the solution of an approach compared to the true solution. 'Genotyping errors' refer to nucleotides originally scored from the direct sequencing of PCR products that were found to be in error based on subsequent sequencing of multiple clones of (independent) PCR products. These include sites that were originally scored as a single peak and later found to be heterozygous. The following basic statistics were used to summarize our scnpc76 data: number (percentage) of polymorphic sites, number (percentage) of triallelic sites, number of homozygotes, number of one-site heterozygotes, number of multisite heterozygotes, frequencies of variations and heterozygosities of polymorphic sites.

Statistical haplotype reconstruction using the CV approach

Over 14 algorithms were tested for our data. HAPINFERX and SLHAP (Lin *et al*. 2002) were kindly provided by Dr Clark and Dr Lin, the remaining algorithms were downloaded from their websites (listed at the end of the main text). Because the size of the data set exceeded the maximum capacity of several algorithms (e.g. HAPLOTYPER), 30 homozygotes with an identical phase (being the most predominant genotype in our samples, with a percentage of 25.5%) were excluded at the initial stage of analysis (and readmitted later, see below). Thus, a data set of 496 individuals was analysed following these steps:

1 *Algorithm examination*. Several indices, i.e. the number of distinct haplotypes (NDH) in a solution, the number of individuals whose haplotype pair is of low confidence probability (NLCP) and the likelihood value of the solution, were used to examine the suitability of an algorithm to the data. Here, the threshold of low confidence probability is defined as follows: 0.05 for HAPINFERX (probability of the second solution), 0.5 for HAPLOREC, and 0.95 for the others. Some algorithms may assign more than one pair of haplotypes to a given individual with different levels of certainty when confronted with phase-ambiguous sites. We would like to emphasize that the NDH value should not be much larger than the number of polymorphic sites observed in the data (see Discussion). Note that some algorithms (e.g. HAPLOTYPER and GCHAP) employed a coding switch technique which encoded a triallelic site as two biallelic ones in order to handle triallelic sites in analysis.
2 *Consistency test*. Comparison between solutions from independent runs of the same algorithm was performed to examine the consistency of that algorithm, a practical way to further evaluate if an algorithm is suitable for a data set. Here, 'independent run' means separate runs using different initial random seeds (e.g. PHASE) or input order (HAPINFERX), or iteration numbers with equilibrium likelihood value (HAPLOREC). Two features were examined for this. One is the individual-resolving discrepancy (IRD), that is, the proportion (number) of individuals whose genotypes were resolved differently between two solutions. It verifies individual-resolving consistency. The other is the haplotype-inferring discrepancy (HID), that is, the discrepancy of distinct haplotypes between two solutions. $HID_{ab} = 1–[2\,k_i/(k_a + k_b)]$, where k_i is the number of identical distinct haplotypes between two solutions, and k_a and k_b are the number of distinct haplotypes in two solutions, respectively. HID evaluates haplotype-inferring consistency. In addition, all algorithms were inspected for inconsistencies between input and output files (i.e. heterozygous sites of some individuals in the input file were reported as homozygous in the output file), which were observed in some simulation studies (Excoffier *et al*. 2003). This was not observed in our study. Finally, six algorithms within three theoretical frameworks (HAPINFERX, ARLEQUIN 3.11-EM, GCHAP, PHASE 2.1.1, HAPLOTYPER 1.0 and HAPLOREC 1.0) were employed in our CV approach (see Discussion).
3 *Data grouping according to consensus votes*. We sorted homozygotes and one-site heterozygotes into two

categories, H and S, respectively. Then, for each ambiguous individual, we examined the sameness of six solutions. Each is the most likely solution with high confidence probability from the corresponding algorithm. Here, the solution with low confidence probability was not considered. We then conducted a consensus vote to choose the best-matched solution. We put equal weight to each algorithm, except for individuals having two primary haplotype pairs which obtained equal votes. In the latter case, we preferred to choose the non-HAPINFERX solution. This is because HAPINFERX is sensitive to input order (Clark 1990), and more sensitive to departure from HWE and genetic diversity level than other algorithms (Niu *et al*. 2002; Niu 2004). Individuals were grouped into different CV categories according to their consensus votes. Here, we ignored a special bias which HAPLOTYPER generated — a low confidence probability (0.5 or so) was given to nine individuals with the same genotype in the category with six votes. Such a bias occurring at multi-individual level is extremely rare, and disappeared when only a subset of the data was analysed.

Using the above methods, a set of grouped data was obtained, in which each individual bearing two defined or inferred haplotypes was classified into a certain category, and with the 30 homozygotes dropped out earlier (due to data set size constraint, see above) readmitted. The total data were partitioned into six categories: H, S, I, II, III, and IV. Category H is composed of homozygotes, and thus with clear phases; category S consists of one-site heterozygotes, the phase of which can be unambiguously deduced; categories I to III contain individuals with six to four votes, respectively; the last category (IV) comprises individuals with two and three votes. This also formed the CV solution of the total data.

Evaluation of the performance of the statistical methods and the CV approach

Besides the several diagnostic indices described above (e.g. NDH and NLCP) for individual algorithms, we performed pairwise comparisons both between the CV solution and those individual algorithms and between the six algorithms themselves by examining their IRD and HID. The two measures (IRD and HID) indirectly (compared to the CV solution) and directly (compared with each other) reflect the different performance of the six algorithms on our data.

After we obtained the true solution through experimental verification, the performance of haplotype inference by the six algorithms and the CV approach was examined through five accuracy measures. The first four are: (i) the error rate of individuals (the proportion of individuals whose haplotypes are not correct; Niu *et al*. 2002), (ii) the error rate of distinct genotypes (the proportion of distinct genotypes whose haplotypes are not correct), (iii) the switch error (a measure of the similarity between the estimated haplotypes and the true haplotypes from the switch accuracy defined by Lin *et al*. (2002), averaged over the ambiguous individuals in the sample), and (iv) the haplotype-inferring error (the errors of distinct haplotypes between the solution of the statistical approach and the true solution, calculated using the HID formula, see above). The switch error for an individual is defined as $sw/(n-1)$, where n denotes the number of heterozygous sites and sw the number of phase switches required to recover the correct haplotype pair from the reconstructed pair. The haplotype-inferring error considers both the incorrectly inferred haplotypes (not actually present) and the lost haplotypes (present but not inferred). The fifth measure is the mean squared error (MSE_{te}), a global measure of discrepancy of haplotype frequencies between the estimated and the true values for a data set (Fallin & Schork 2000). This allows comparison of the haplotype frequencies in the true solution (T_k) to the corresponding estimates (E_k) of the six algorithms and the CV approach. $MSE_{te} = [\Sigma_k(T_k - E_k)^2]/n$, where $k = 1 \ldots n$, where n is the number of distinct haplotypes in the true solution. Note that smaller values of these measures indicate a better performance. For simplicity and clarity, individuals with genotyping errors were excluded from analysis (this does not affect the conclusion obtained).

Analysis of the grouped data

Separate HWE tests (Guo & Thompson 1992) were performed for each polymorphic site on genotype data using ARLEQUIN 3.11 (Excoffier *et al*. 2005). Genetic diversity within populations was measured by nucleotide diversity (π) and haplotype diversity (h) calculated using ARLEQUIN 3.11. Correlation analysis was performed using STATISTICA 6.0 (StatSoft).

The amount and quality of phase information extracted from the genotype data varied considerably among categories. To characterize the sensitivity of the CV approach to genetic diversity within populations, we examined the correlations between genetic diversity of populations and the increments in category I of resolved distinct genotypes, distinct haplotypes and polymorphic sites in the 31 studied populations. In addition, for inspecting the effect of the number of heterozygous sites on haplotype reconstruction in multisite heterozygotes, we examined the cumulative resolutions of resolved individuals and resolved genotypes as the functions of the number of heterozygous sites and the CV categories (I–IV).

In many haplotype reconstruction studies using real data (e.g. in linkage analysis), rare variations (defined as those whose frequency at the relevant polymorphic site

is ≤ 5%) were excluded because they contributed very little information to the whole data. But here we included them since rare variations in population genetic studies provide important information about the evolutionary and demographical history of the species. Thus, it is essential to understand the effect of rare variations on haplotype reconstruction. We inspected the patterns of distribution of heterozygous sites in the four CV categories (I–IV) on two frequency thresholds of 1% and 5%.

We have written a Perl script CVhaplot to automatically perform the CV analysis described above. It is freely available from the website for noncommercial purpose (http://www.ioz.ac.cn/department/agripest/group/zhangdx/ZhangDX_E.htm).

Results

Characteristics of the raw scnpc76 data set

The length of the scnpc76 sequences used in our analysis is 251 bp (sites 182–432 of the GenBank sequence AJ558251), with 71 (28.3%) nucleotide sites being polymorphic in the total data, of which seven are triallelic (9.9%) and the rest biallelic (90.1%). All polymorphic changes are nucleotide substitutions. The average probability that an individual is heterozygous at a nucleotide site is 0.021, that is, an individual is expected to be heterozygous at one of every 48 sites at this locus. The number of heterozygous sites in heterozygotes varies from one to 18 (but 10 and 11 not being observed). Our data also revealed that sites that host variations with higher population frequency are more likely to be in Hardy–Weinberg disequilibrium (HWD), with an excess of homozygosity. This considerably decreases the complexity of our data. The detailed information (including frequencies of variations, heterozygosities of polymorphic sites and HWD at polymorphic sites) is provided in Table S2 (Supplementary material).

Among the 526 individuals sequenced, there are 179 homozygotes, 15 one-site heterozygotes and 332 multisite heterozygotes. The most predominant homozygotes have a frequency of 25.5% (134 individuals). The raw data contain 141 distinct genotypes with their frequencies ranging from 0.2% (one individual) to 25.5% (134 individuals), of which 17 are from homozygous, nine from one-site heterozygous, and 115 from multisite heterozygous (Fig. S1, Supplementary material). A large proportion (87.8%) of ambiguous genotypes occurred at extremely low frequency (< 1%). Twenty-four distinct haplotypes were unambiguously defined from the first two categories (H and S), which is an essential prerequisite for starting the program HAPINFERX.

Haplotype reconstruction using the CV approach

The number of distinct haplotypes (NDH) inferred by the six algorithms varies from 67 to 78. The CV solution consists of 68 distinct haplotypes, which were partitioned into six categories (H, S, I–IV) according to the characteristics of the raw genotype data and their consensus votes (Table 1). The 24 haplotypes determined directly and unambiguously from the sequences of homozygotes (category H) and one-site heterozygotes (category S) account in frequency for 91.4% of haplotypes of the entire sample (1052 chromosomes). The frequency distributions of the 68 haplotypes of the CV solution are shown in Fig. S1. An L-shaped highly skewed distribution is evident, with a long tail containing the inferred haplotypes. Haplotype 1 has a much higher frequency (46.2%, 486 out of 1052

Table 1 Summary statistics on the consensus vote approach and experimental verification over the CV categories

CV category	H	S	I	II	III	IV	Overall
Consensus votes	NA	NA	6	5	4	3 & 2	NA
Individuals	179	15	265	32	25	10	526
Number of distinct genotypes	17	9	64	22	20	9	141
Total distinct haplotypes	17	15	30	28	26	13	68
Number of category-unique haplotypes*	17	7	11	11	16	6	68
Frequency (%) of category-unique haplotypes in the total sample	85.6	5.8	4.3	1.8	1.8	0.7	100
Cloned individuals†	2¶	0	50	25	21	8	106
Genotyping error	NA	NA	6	6	1	1	14
Individual-resolving error‡	NA	NA	0	40.0%	28.6%	25.0%	5.8%
Error rate of haplotypes§	NA	NA	0	28.6%	53.3%	25.0%	19.3%

NA, not applicable. *Refers to haplotypes that were newly observed in a category (i.e. not observed in the previous categories). †For each category-unique haplotype in categories I–IV, at least one individual was verified via molecular cloning and sequencing. ‡Error rate of individuals. §Denotes the proportion of incorrectly inferred haplotypes in the solution of the approach. ¶For the other distinct homozygotes, they were resequenced at least once from independent genomic PCR, and no error or allele-drop was observed.

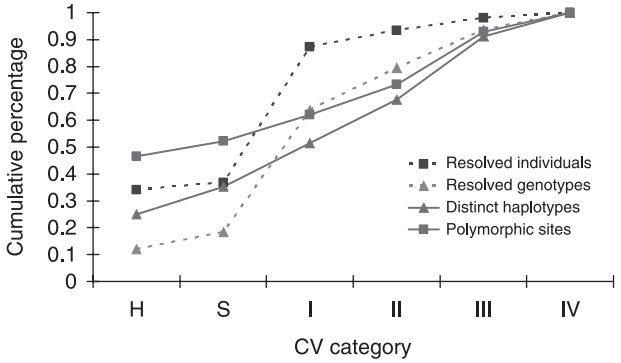


Fig. 1 Cumulative resolution of statistical inference over the CV categories using the CV approach. The cumulative percentages of resolved individuals, resolved genotypes, distinct haplotypes and polymorphic sites were plotted against the six CV categories.

haplotypes) than the others. The remaining 67 haplotypes vary in frequency from 6.7 to 0.1%. Fifty-three haplotypes, including all 44 haplotypes inferred by the CV approach, are extremely rare (frequency below 1%), with 30 haplotypes occurring only once (28 inferred). Experimental verification confirmed that the CV approach has the best overall performance (Table 2), with more than 96% of individuals assigned correct haplotypes, and the accuracy of haplotype-inferring was the highest (80%). It also has the lowest switch error.

The CV approach has the unique ability to identify uncertain inferrals (those in categories II-IV). Figure 1 shows the cumulative resolution of individuals, genotypes, haplotypes and polymorphic sites over the CV categories. The sharp increase in the percentage of resolved individuals (from 36.9 to 87.3%) and resolved genotypes (from 18.4 to 63.8%) is most notable from S to I, despite the only modest increase of distinct haplotypes (from 35.3 to 51.5%) and polymorphic sites (from 52.1 to 62.0%). Experimental verification confirmed that there is no error for inferrals in category I (Table 1).

The sensitivity of the CV approach to levels of genetic diversity was investigated. Here, we characterized genetic diversity within populations by nucleotide diversity (π) and haplotype diversity (h). The values of these two measures of the total sample are 0.0243 and 0.775, respectively. In the 31 populations studied, nucleotide diversity (π) ranges from 0.0119 to 0.0355, and haplotype diversity (h) from 0.437 to 0.938. There was a significant correlation between π and h in populations (Pearson correlation coefficient, 0.807, d.f. = 30, $P < 0.001$). There were significant negative correlations between the increments of the resolved genotypes/haplotypes/polymorphic sites from S to I and the levels of haplotype diversity (or nucleotide diversity) (Fig. S2A–C, Supplementary material. Pearson correlation coefficients: with haplotype diversity, –0.478/–0.488/–0.484, $P = 0.007/0.005/0.006$ for genotypes/haplotypes/polymorphic sites, respectively, d.f. = 30; with nucleotide diversity –0.500/–0.427/–0.437, $P = 0.004/0.017/0.014$ for genotypes/haplotypes/polymorphic sites, respectively, d.f. = 30). In addition, statistical inference errors occurred significantly more frequently in populations with high haplotype diversity (Fig. S3, Supplementary material; Pearson correlation coefficients, 0.531/0.409, $P = 0.002/0.022$ for individuals/genotypes, respectively, d.f. = 30), and significant correlations between inference error rates and the levels of haplotype diversity (Fig. S2D–E) were observed (Pearson correlation coefficients, 0.507/0.382, $P = 0.004/0.034$ for individuals/genotypes, respectively, d.f. = 30).

Table 2 Performance and reliability of the consensus vote approach and the solutions of the six statistical algorithms with the scnpc76 data

	Error rate (%)					
Methods	Individuals*	Genotypes†	Switch‡	Haplotype-inferring§	Haplotypes¶	MSE** (E-06)
Consensus vote	*3.6 (5.8)*	*8.7 (11.0)*	*1.5*	*20.0*	*19.3*/20.7	2.8
HAPINFERX	6.1 (9.9)	15.9 (20.0)	3.3	26.8	30.8/22.4	9.7
PHASE 2.1.1	3.9 (6.4)	9.5 (12.0)	1.7	22.4	22.4/22.4	2.9
HAPLOTYPER 1.0	*3.6 (5.8)*	*8.7 (11.0)*	1.6	*20.0*	*19.3*/20.7	2.6
HAPLOREC 1.0	4.1 (6.7)	10.3 (13.0)	2.3	20.7	23.8/*17.2*	2.8
ARLEQUIN 3 .11-EM	5.3 (8.6)	12.7 (16.0)	2.7	24.4	26.2/22.4	6.8
GCHAP	3.7 (6.1)	11.9 (15.0)	1.9	28.2	28.8/27.6	*2.1*

Values in italics indicate the best performance value. *Inference error rate of individuals. Value in parenthesis is the value if only the ambiguous individuals were considered. †Inference error rate of genotypes. Value in parenthesis is the value if only the ambiguous genotypes were considered. ‡Switch error, see the text. §Haplotype-inferring error, i.e. the errors of distinct haplotypes between the solution of a statistical approach and the true solution. It considers both the incorrectly inferred haplotypes (not actually present) and the lost haplotypes (present but not inferred). ¶Haplotype-inferring error shown as the proportion of incorrectly inferred haplotypes in the solution of the approach and the proportion of lost haplotypes compared to the true solution. **Mean squared error.

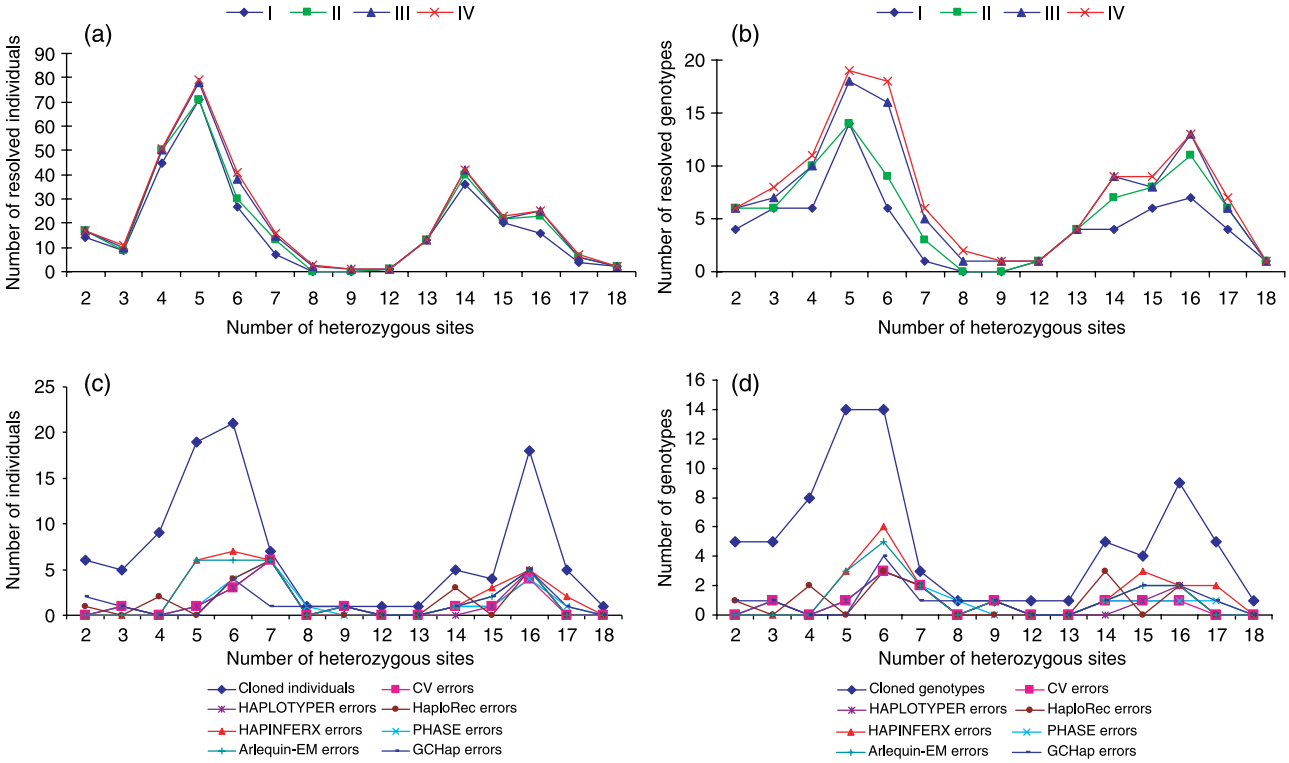
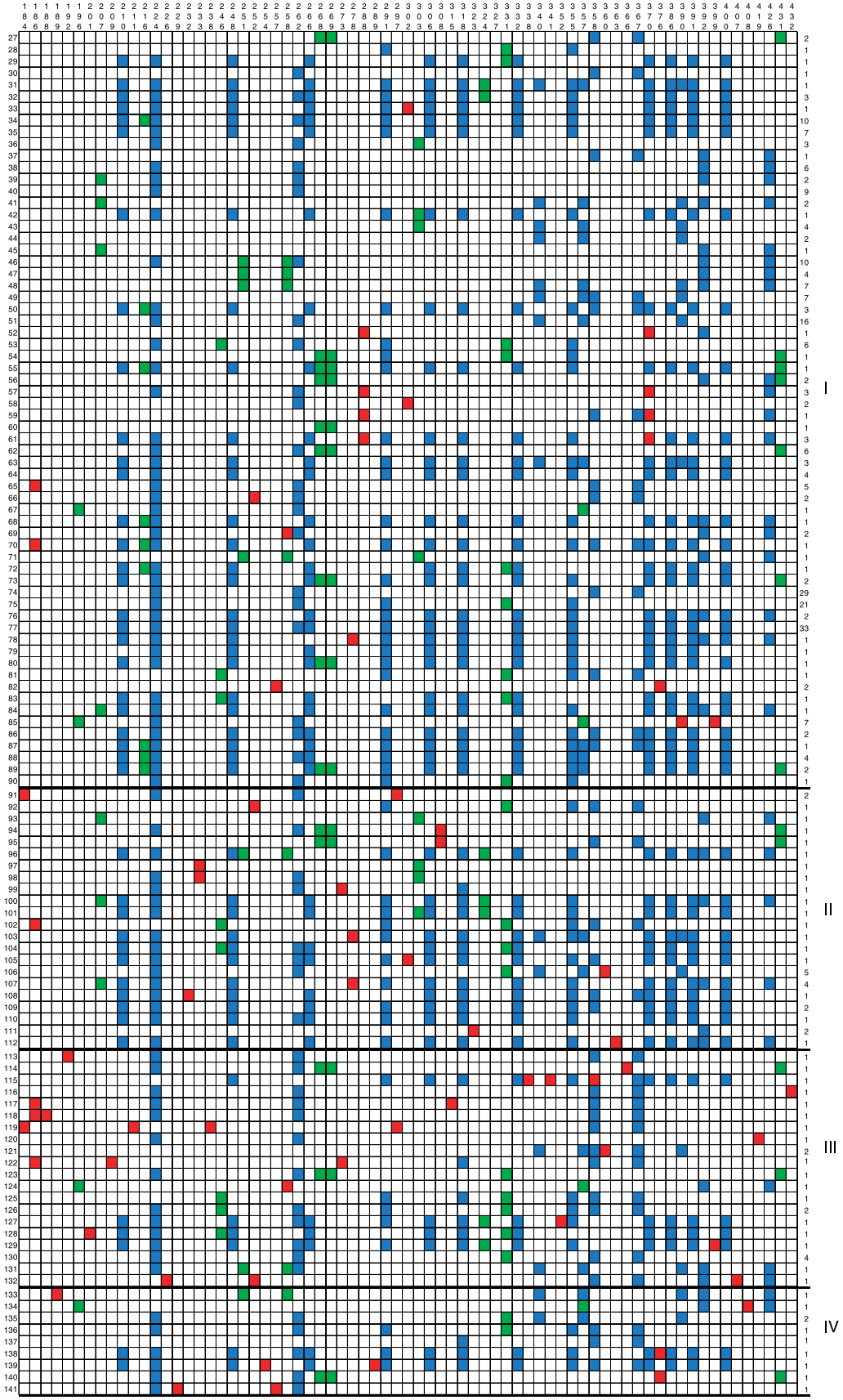


Fig. 2 Frequency distribution of the resolved individuals, resolved genotypes and inference errors on functions of the number of heterozygous sites. (a) Comparison of the cumulative frequency distribution of the resolved individuals among the CV categories (I–IV). (b) Comparison of the cumulative frequency distribution of the resolved genotypes among the CV categories (I–IV). (c) Comparison of the distribution of the inference errors of ambiguous individuals among the six inference methods and the CV approach. (d) Comparison of the distribution of the inference errors of ambiguous genotypes among the six inference methods and the CV approach. Note that the dark blue line in C and D, as the reference line, plotted the total numbers of individuals (genotypes) that were experimentally verified by high-fidelity *Pfu* PCR cloning and sequencing; the other lines, the number of observed errors of the seven different approaches.

Figure 2(a–b) shows the cumulative frequency distribution of the resolved individuals and resolved genotypes as the functions of the number of heterozygous sites and the CV categories (I–IV). All cumulative distributions approximate a bimodal distribution. The number of uncertain individuals (genotypes) is proportional to the number of ambiguous individuals (genotypes) but not to the number of heterozygous sites. Also, the inference error rate does not increase with the increase of the number of heterozygous sites (Fig. 2c–d. Spearman rank correlation coefficients, –0.079/–0.083, $P = 0.779/0.769$ for individuals/genotypes, respectively, d.f. = 14). It is interesting to observe the bimodal distribution of heterozygotes in regard to the number of heterozygous sites they carry. This is owing to the presence of two highly diverged lineages of haplotypes in the migratory locust, presumably a result of ancient population isolation.

A closer inspection of 115 ambiguous genotypes shows that rare variations (frequency ≤ 5%) can influence haplotype reconstruction of genotypes (individuals) bearing such variations, with the resolution of genotypes tending not to be accepted by all algorithms. Figure 3 shows the distributional patterns of heterozygous sites in the four CV categories (I–IV). Clearly, there are relatively less rare variations in category I than in the other three categories (II–IV). This is especially true for those with extremely rare variation (frequency < 1%, red boxes in Fig. 3). The extremely rare variations occurred 45% less frequently, and 2.5 times, 3.3 times and 2.4 times more frequently than expected in the CV categories I, and II, III and IV, respectively, and this

Fig. 3 Distribution patterns of the heterozygous sites of the 115 ambiguous genotypes over the four CV categories (I–IV). The Arabic numbers on the top are the nucleotide positions of the polymorphic sites with reference to the GenBank sequence AJ558251. The Arabic numbers on the left-hand side are the serial number of the ambiguous genotype (note: genotypes 1–26 are in the category H and S). The Arabic numbers on the right-hand side are the frequency of the correspondent genotype in the total sample. The roman numerals on the right-hand side indicate the CV categories to which the corresponding genotypes belong. Category I is composed of 64 distinct genotypes (No. 27–90), the other three categories include 22, 20 and nine distinct genotypes, respectively. A blue box represents the common heterozygotes (frequency > 5%) at the site, a green box the rare heterozygotes (1–5%), and a red box the extremely rare heterozygotes (< 1%).



uneven distribution is statistically significant ($\chi^2 = 90.14$, d.f. = 3, $P < 0.001$). Moreover, 22 of the 23 singleton variations fell into the categories II to IV (the only remaining one in S). Our experimental study revealed that 83.3% of inference errors resulted from rare variations, with 33.3% from extremely rare variations (frequency < 1%). In addition, the phase of 26.3% heterozygous sites with singleton variation was incorrectly inferred.

Experimental verification

Five independent clones have been sequenced for each of the two homozygote controls and no error was observed, indicating that the high-fidelity PCR and cloning system was reliable. Among the 116 ambiguous individuals (with 85 genotypes) verified by cloning experiment (Table 1), statistical inference errors were observed in 33% of uncertain individuals (categories II–IV), and no such errors in individuals of category I. Overall, 20 haplotypes and 67 haplotype pairs inferred from the CV approach were confirmed by experiments, and 14 new haplotypes and 18 new haplotype pairs uncovered. Thus, after experimental verification, the final numbers of segregating sites, distinct haplotypes and distinct genotypes for the locust data were reduced to 70, 63 and 138, respectively (The EMBL database accession numbers of the 63 haplotype sequences are AM889405–AM889467). The highest frequency of any haplotype in our data is 45.5% (479 out of 1052 haplotypes). The top four common haplotypes (those with a frequency > 5%), the top nine common haplotypes (those with a frequency > 2%) and the first 24 unambiguous haplotypes account for 65.0%, 78.8% and 91.3% of haplotypes of the entire sample, respectively. Sequencing bias (refers to nonhuman-mediated errors) is the main contributor of genotyping errors. The overall genotyping error is about 2.7% for our raw data (that is, 2.7% of total individuals encountered some genotyping error); such quality is essential for a successful statistical haplotype reconstruction and frequency estimation (Kirk & Cardon 2002; Niu 2004). This amount of genotyping errors does not change the significance of statistical test results, nor does it modify the conclusions. Therefore, for the sake of clarity and simplicity, we do not include them in the following discussion unless otherwise specified (but their effects were considered in our analyses).

Discussion

The performance of statistical haplotype reconstruction methods

The six algorithms all performed quite well in either haplotype assignment to individuals or estimation of haplotype frequencies (Table 2). In general, more than 90% of ambiguous individuals were assigned correct haplotypes. The accuracy of the estimation of haplotype frequencies can be seen from the small values of the mean squared error (MSE). Such a performance is due to several particular proprieties of the data, including: (i) large sample size (526 individuals), (ii) a significant proportion of individuals with clear phase (34% homozygotes and 2.9% one-site heterozygotes), with 24 distinct haplotypes defined, (iii) a much smaller number of distinct haplotypes (63) compared to the total (1052), and highly skewed frequency distribution (0.1%–45.5%, L-shaped), (iv) HWD at variable sites with excess homozygosity, and (v) moderate haplotype diversity of the entire sample (0.775). These archetypal characteristics (Fallin & Schork 2000; Niu *et al*. 2002; Eronen *et al*. 2006) make the scnpc76 data most favourable for statistical haplotype inference. Nevertheless, in this case, haplotype-inferring appears to be more problematic than individual-resolving in statistical haplotype reconstruction (Tables 2–4), as indicated by the generally larger haplotype-inferring discrepancy (HID) than individual-resolving discrepancy (IRD), both among independent runs of the same algorithm (Table 3) and among different algorithms (Table 4). The haplotype-inferring error is around 20% (and can be as higher as 28.2% for GCHAP, for example; Table 2). Much of the bias in haplotype frequency estimation results from common haplotypes (frequency > 5%), especially from the most frequent one (data not shown).

The solutions of the six algorithms produce different features, as shown in Tables 3 and 4. For example, the different values of NLCP suggest that these algorithms could not always assign confident haplotypes to a varying portion of ambiguous individuals. While HAPINFERX and HAPLOREC appeared to be sensitive to the running conditions, PHASE, HAPLOTYPER, ARLEQUIN-EM and GCHAP showed little sensitivity. Inference consistency was inspected in five of the six algorithms (HAPINFERX, PHASE, HAPLOTYPER, ARLEQUIN-EM and HAPLOREC) by comparing solutions of independent runs (GCHAP did not show any difference between runs). The consistency of the last four algorithms (i.e. except HAPINFERX) was found to be fairly good, with the values of the discrepancy measures being very small (< 1% for IRD and < 5% for HID). However, input order resulted in inconsistency among solutions of HAPINFERX.

Our results also demonstrated that no individual algorithm is free from inference errors. For example, PHASE is one of the most popular algorithms due to its excellent performance (e.g. Stephens & Donnelly 2003; Adkins 2004; Marchini *et al*. 2006), but based on the confidence probability of haplotype assignment on our data, its solution is not reliable (as noticed by Sabbagh & Darlu 2005. Here, we have summarized its performance over independent analyses from multiple runs). It should be noted that GCHAP performed well in our study, and yielded an exceptionally low value of NLCP (Table 3). However, it would be

Table 3 Running specifications of the six selected algorithms and basic statistics of their solutions on the scnpc76 genotype data of *Locusta migratoria*

Algorithms	Coding switch	Initial conditions	Diagnostic indices			
			NDH*	NLCP†	IRD(SD)‡	HID(SD)§
HAPINFERX	–	Variable input order	78	14	3.4% (2.0%)	13.7% (8.4%)
PHASE 2.1.1	–	MR, 2000 iterations	67	30	0.3% (0.3%)	2.1% (1.5%)
HAPLOTYPER 1.0	+	Default	67	31	0.7% (0.4%)	4.1% (2.0%)
HAPLOREC 1.0	–	VMM	76	47	0.7% (0.4%)	3.7% (2.3%)
ARLEQUIN 3.11-EM	–	Default	73	15	0.04% (0.08%)	0.3% (0.6%)
GCHAP	+	Default	70	2	NA	NA

NA, not applicable. The two statistics of consistency (IRD and HID) resulted from the pairwise comparisons of 10 independent runs. *Number of distinct haplotypes in a solution. †Number of individuals whose haplotype pair is of low confidence probability in a solution. ‡Individual-resolving discrepancy, the proportion of individuals whose genotypes were resolved differently between two solutions. §Haplotype-inferring discrepancy, i.e. the discrepancy of distinct haplotypes between two solutions. SD, standard deviation.

Table 4 Pairwise comparisons of the solutions of the CV approach and the six algorithms according to the haplotype-inferring discrepancy (above diagonal) and individual-resolving discrepancy (below diagonal). All values in %

	HIF	PHA	HTY	HAR	AEM	GCH	CV	Mean discrepancy* IRD†	Mean discrepancy* HID‡
HIF		24.1	17.2	33.8	6.0	16.2	12.3	5.1	19.5
PHA	6.1		16.4	30.1	21.4	24.1	15.6	4.4	23.2
HTY	5.3	2.3		27.3	14.3	18.2	5.2	3.7	18.7
HAR	*6.7*	4.8	4.2		32.9	*37.0*	26.4	*5.6*	*32.2*
AEM	2.3	4.2	3.4	5.9		16.1	9.2	4.0	18.1
GCH	5.3	4.2	3.6	6.3	4.2		14.5	4.8	22.3
CV	4.0	2.5	1.0	3.8	2.5	2.9		NA	NA

NA, not applicable. Values in italics indicate the largest discrepancy. Abbreviations used: HIF, HAPINFERX; PHA, PHASE 2.1.1; HTY, HAPLOTYPER 1.0; HAR, HAPLOREC 1.0; AEM, ARLEQUIN 3.11-EM; GCH, GCHAP; CV, consensus vote. *The mean values only consider pairwise comparison among the solutions of the six algorithms. †Individual-resolving discrepancy, i.e. the proportion of individuals whose genotypes were resolved differently in two solutions. ‡Haplotype-inferring discrepancy, i.e. the discrepancy of distinct haplotypes between two solutions.

imprudent to directly employ its solution in downstream analyses, because it gave high confidence probabilities to many incorrect inferrals in its solution. Therefore, behind its superficial least value of NLCP is rather frequent misassignment.

Statistical haplotyping using the CV approach

The CV approach partitions the ambiguous genotype data into different categories according to the number of votes a solution has received. It is expected that solutions that were supported by all algorithms should be mostly likely correct because they pass the test of all algorithms with different underlying statistical assumptions and computational techniques. This expectation has been confirmed in our study. The incorrect haplotypes inferred in the CV solution only occurred in categories II–IV and not in the category I (Table 1). Therefore, the CV approach allows us to focus on solutions with greater uncertainty in experimental verification, i.e. those in categories II–IV. This considerably reduces experimental work while acknowledging uncertainty in statistical haplotype inference. Our experimental work proved the effectiveness of such an approach.

Our results indicate that pooled data from multiple geographical populations with varying levels of genetic diversity do not affect much the overall performance of the CV approach, provided that populations do not have

extremely high genetic diversity. This considerably extends the conclusions from studies treating only two subsamples with conventional statistical approaches (Stephens & Scheet 2005; Scheet & Stephens 2006; Andres *et al*. 2007). Another feature worth pointing out is that the elevated number of heterozygous sites in scnpc76 marker has little influence on the inference quality in our study — both the resolution of the CV approach and the error rate of the six algorithms were independent of the number of heterozygous sites in ambiguous individuals (Fig. 2). Therefore, Adkins' warning that 'computationally assigned haplotypes for subjects heterozygous at more than four single nucleoutide polymorphisms (SNPs) should be viewed with extreme caution' (Adkins 2004) may not be generally applicable in scnp DNA studies.

Several qualities makes the CV approach a potentially effective strategy for haplotype reconstruction, viz. its considerable tolerance of pooled population samples with varying genetic diversity (even population substructure), insensitivity to the number of heterozygous sites borne in ambiguous genotypes, and the aforementioned unique power to partition the inferrals into a reliable group and an uncertain group. However, there are some imperfections of this approach as seen in our analysis. First, we observed that some statistical biases generated by different algorithms are not independent, which may somehow compromise this approach. A closer inspection reveals that several algorithms gave wrong solutions to a number of identical individuals in category II, producing an unexpected elevated error rate of individuals in this category (Table 1). Qin *et al*. (2002) also noticed that algorithms can make inference errors on the same individuals. This is a limitation of current statistical haplotype inference algorithms that may unexpectedly reduce the effectiveness of the CV approach. Second, because the CV approach is an algorithm-dependent strategy, its performance is inevitably affected by the algorithms included. Therefore, the CV approach retains some characteristics common to the individual algorithms it includes. For example, it shows some sensitivity to high genetic diversity in populations, because each of the six individual algorithms employed in this study (HAPINFERX, PHASE, HAPLOTYPER, ARLEQUIN-EM, HAPLOREC, and GCHAP) produced significantly more inference errors in populations with high genetic diversity (Fig. S3) (Pearson correlation coefficients: with haplotype diversity, 0.518/0.530/0.491/0.551/0.505/0.459, $P = 0.003/0.002/0.005/0.001/0.004/0.009$, respectively, d.f. = 30; with nucleotide diversity, 0.521/0.448/0.473/0.551/0.472/0.442, $P = 0.003/0.011/0.007/0.001/0.007/0.013$, respectively, d.f. = 30). Third, for some algorithms, data pooling may occasionally decrease the confidence probability of the correct solution of certain low frequency genotypes. In our case, this was observed for one genotype (frequency 1.7%) in HAPLOTYPER. Fourth, further examination is needed to see how robustly the CV approach can deal with pooled population data in a much wider context.

Rare variations and rare haplotypes

In applications such as linkage analysis or trait mapping, rare variations can be ignored without losing much information. However, rare variations and rare haplotypes can be very important, for example in the genetic study of complex human diseases (Pritchard 2001; Cohen *et al*. 2004), or even essential, for example in evolutionary analysis of natural populations (Crandall & Templeton 1993; Brumfield *et al*. 2003; Templeton 2004). Thus, it is important to understand the behaviours and effects of rare variations and rare haplotypes in statistical haplotype reconstruction for scnp markers. As an important feature, our data contain a large proportion (about 90%) of rare ambiguous genotypes (i.e. those with frequency < 1%) which harbour a large portion (> 70%) of rare variations (i.e. those with frequency ≤ 5%) and a large number (47) of rare haplotypes (i.e. those with frequency < 1%).

Rare variations may have an important impact on statistical inference if the reliability of haplotype phases rather than the haplotype frequencies in a population is the central concern (Fallin & Schork 2000; Stephens *et al*. 2001; Lin *et al*. 2002). With the CV approach, the high proportion of rare variations in our data leads to increased genotype-resolving uncertainty and error (Fig. 3). Statistically, genotypes with heterozygous sites bearing extremely rare variation tend not to be identically resolved by all algorithms (Fisher's exact test, $P < 0.001$), thus leading to more inference uncertainties and errors. The test is still significant even when the variations were extended to those with a frequency of ≤ 5% ($P = 0.012$). We observed that 72.7% inference errors of distinct genotypes directly resulted from heterozygous sites with rare variation, and all incorrectly resolved genotypes contained heterozygous sites with rare variation. For the six statistical haplotype reconstruction methods, both the incorrect haplotypes and the lost haplotypes have an extremely low population frequency (< 0.7%). In the CV solution, 11 of the 44 inferred haplotypes (all in categories II–IV and with a frequency < 0.5%) are inference artefacts. This indicates that the haplotype-inferring errors resulted entirely from extremely rare haplotypes in our study. Such a tendency was also observed in other studies (e.g. Tishkoff *et al*. 2000 and Adkins 2004). Dramatic changes of haplotype frequency estimates between the values of the solution of a given approach and the true solution occurred only with the low frequency haplotypes, as noticed by Tishkoff *et al*. (2000) and Sabbagh & Darlu (2005). The above observations call for extra surveillance of rare variations and haplotypes in the inferred solutions. This also means that empirical verification should focus more on such haplotypes.

Application of the CV approach: some practical considerations

The potential value of nuclear DNA analysis in genetic studies of populations has greatly stimulated the development of haplotype reconstruction algorithms. In consequence, the recent rapid development of statistical haplotying algorithms (see review, Salem *et al*. 2005) provides the basis to develop the CV approach. This has the potential to become a cost-effective strategy for haplotyping in large-scale investigations such as population genetic studies with scnp markers. Thus, an interesting question is which (and also how many) algorithms should be included in the CV approach. A large number of algorithms will in general increase the reliability but at the expense of more effort and time, whereas fewer algorithms or a poor combination of algorithms will impair the power of the CV approach. It is a challenging task to find the best combination of algorithms, since the actual phase information of the raw genotype data is unclear.

We recommend the following operational procedures when the CV approach is applied. First, the characteristics of the genotype data should be reviewed to see how close they fit to the theoretical expectations (see above). If the data fit well, less difficulty is expected for haplotype inference. Second, a pilot test of algorithms should be performed to examine the relevant diagnostic indices on initial conditions and consistency (see Materials and methods). Most algorithms developed so far are orientated, either explicitly or implicitly, to particular applications. The theories and computational techniques embedded in an algorithm (for reviews, see Halldorsson *et al*. 2004; Niu 2004) reflect the authors' opinions on the specific data in question (e.g. type of data, haplotype structure, etc.). Therefore, the performance of these methods is likely to be different when they are applied to other data. For example, the human genome contains millions of SNPs (Kruglyak & Nickerson 2001; International HapMap Consortium 2005), most of which are biallelic. Thus, many algorithms were especially designed for biallelic SNP data, and hence unable to satisfy genealogical research requirements with scnp marker containing triallelic sites. These are rather common in large-scale population genetic analysis; for example, 10% polymorphic sites are triallelic in our case. In this regard, PHASE assumes a special model of parent-independent mutation that is quite appropriate for triallelic sites, and performs well on our genotype data (as well as its stepwise mutation model). Some biallelic data-orientated algorithms (e.g. HAPLOTYPER, GCHAP) can use the coding switch technique to accommodate triallelic sites. But the coding switch technique occasionally can lead to inference errors; for example, it has falsely chosen the fictitious haplotype created by some redundant coding in GCHAP in our study.

Third, a careful selection of representative algorithms should be undertaken. Basing on the type of technical framework with which a program handles segregating sites in long sequence data (i.e. data bearing a large number of segregating sites), statistical haplotype reconstruction methods can be divided into four groups: PL-based (e.g. HAPLOTYPER), sequential-based (e.g. GCHAP), block partitioning-based [e.g. GERBIL (Kimmel & Shamir 2005)] and others (e.g. HAPINFERX). Therefore, there are three different strategies for selecting the statistical methods in a CV approach. The first is to select methods basing on different theoretical frameworks (parsimony, EM or Bayesian, see Introduction); the second is to select methods basing on different technical frameworks; and the third is a combination of the above two strategies. A retrospective examination showed that each of these strategies can produce an optimal combination of algorithms for our data (here, optimal combination refers to a combination that has the least number of algorithms and the best performance power). It is encouraging to see that there are multiple optimal combinations, and the minimum number of algorithms required can be just three for our data (GCHAP, HAPLOREC and GERBIL). This further suggests the power and effectiveness of the CV approach in population genetic study. A good rule of thumb is that a fair number of disparate algorithms should be included in the CV approach to safeguard the inference results, since the optimal combination can not be known ahead of analysis. As a good starting point, we recommend that the following algorithms, HAPINFERX, PHASE, HAPLOTYPER, ARLEQUIN-EM, GCHAP, HAPLOREC and GERBIL, should be first tested in haplotype reconstruction with scnp DNA markers.

Fourth, we strongly suggest considering carefully the confidence probabilities proposed by individual algorithms. When ambiguous individuals have two solutions with equal votes with some algorithm, a lower weight should be accorded by the particular rules mentioned above. In such cases, the accuracy of the CV approach can be improved by giving a lower weight to HAPINFERX, or to the solution with the largest NDH value if HAPINFERX is not involved.

Finally, some parsimonious criteria assist in statistical haplotype reconstruction. First, the more distinct haplotypes a solution of a given algorithm contains, the more uncertainties and/or errors it includes. A similar observation was made in the study of the human apolipoprotein E locus using some rule-based algorithms (Orzack *et al*. 2003). Moreover, there is a significant correlation between the number of distinct haplotypes resolved by statistical methods and inference errors (Spearman rank correlation coefficient, 0.674 for haplotype-inferring error, d.f. = 11, $P = 0.016$; 0.857 for the individual error, d.f. = 11, $P < 0.001$). In general, solutions that contain fewer distinct haplotypes are more accurate. This is true both for comparison among algorithms and among independent runs of a given

algorithm. Second, the true solution of our data contains 63 haplotypes, which is smaller than the number of polymorphic sites ($N = 70$). This is consistent with the parsimonious prediction that if there is no recombination and no recurrent mutation, the number of observed haplotypes should be in general no greater than $(N + 1)$ despite of the fact that N biallelic sites can generate up to 2^N different haplotypes. Thus, the parsimonious expectation defines the upper limit of the number of distinct haplotypes contained in the data. Third, haplotypes in a short nuclear genomic region should typically have an extremely skewed frequency distribution. At the scnpc76 locus of the migratory locust, the top four common haplotypes (frequency > 5%) and the top nine common haplotypes (frequency > 2%) account for 65.0% and 78.8% of haplotypes of the entire sample (1052 chromosomes), respectively. This is analogous with the observation that a few (two to five) common haplotypes constituted some 70–90% of the haplotypes in population in the analysis of human chromosome 21 (Patil *et al*. 2001). These features provide a practical tool for identifying anomaly in statistical haplotype reconstruction.

In population genetic studies, often it is more effective to sample more populations than to increase the number of samples per population if the total number of samples to be analysed is constrained, provided that a minimum sample size per population (e.g. 15–20) can be guaranteed. It is not uncommon in practice that the number of samples that can be collected from natural populations is limited for various reasons. The performance of statistical algorithms in such situations can be problematic, due to higher error rate when individual populations are used as the analysis unit in haplotype reconstruction (see Stephens *et al*. 2001; Wang & Xu 2003; Zhang *et al*. 2006). In addition, the reliability of the inferrals cannot always be guaranteed by the confidence probability of haplotype assignment suggested by individual algorithms, although a number of statistical methods can somehow accommodate population pooling. In such cases, the CV approach can be an effective strategy for statistical haplotype reconstruction by pooling data from all sampling sites.

Acknowledgements

We're grateful to Q.C. Long, L. Kang, Y.L. Chen, S. Tanaka, M. Lecoq and K. Ibrahim for providing some of the locust samples used in this study, Q. Li and Y.J. Wu for their help in cloning experiments. We thank A.G. Clark, S. Lin, M. Stephens, Z.S. Qin, J.S. Liu, A. Thomas, L. Eronen, D. Clayton, L. Richard, Y. Xu, L.N. Chen, L. Cardon and K.M. Chao for kindly providing their haplotype reconstruction software and/or valuable advices and suggestions. We thank L. Excoffier and four anonymous referees for their valuable comments and suggestions on earlier versions of the manuscript. We're in debt to Godfrey Hewitt for his thorough linguistic corrections on the manuscript. This research was supported by the CAS Knowledge Innovation Program (grant no. KZCX2-YW-428), CAS Innovative Research International Partnership Project (CXTDS2005-4), the Natural Science Foundation of China (NSFC grant no. 30730016) and the 'Bai Ren Ji Hua' Professorship.

Websites

The Websites of haplotype reconstruction software used in this article are as follows:

ARLEQUIN, http://cmpg.unibe.ch/software/arlequin3/
GCHAP, http://www-genepi.med.utah.edu/~alun/software/
GERBIL, http://acgt.cs.tau.ac.il/gevalt/
HAP, http://research.calit2.net/hap/
HAPLOREC, http://www.cs.helsinki.fi/group/genetics/haplotyping.html
HAPLOTYPER, http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm
HPLUS, http://qge.fhcrc.org/hplus/
PHASE, http://www.stat.washington.edu/stephens/software.html
PL-EM, http://www.people.fas.harvard.edu/~junliu/plem/
PTG, http://zhangroup.aporc.org/bioinfo/ptg/
SNPHAP, http://www-gene.cimr.cam.ac.uk/clayton/software/

References

Adkins RM (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genetics*, **5**, 22.

Andres AM, Clark AG, Shimmin L *et al*. (2007) Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic Epidemiology*, **31**, 659–671.

Antunes A, Templeton AR, Guyomard R, Alexandrino P (2002) The role of nuclear genes in intraspecific evolutionary inference: genealogy of the transferrin gene in the brown trout. *Molecular Biology and Evolution*, **19**, 1272–1287.

Ayers KL, Sabatti C, Lange K (2007) A dictionary model for haplotyping, genotype calling, and association testing. *Genetic Epidemiology*, **31**, 672–683.

Bartish IV, Kadereit JW, Comes HP (2006) Late Quaternary history of *Hippophae rhamnoides* L. (Elaeagnaceae) inferred from chalcone synthase intron (*Chsi*) sequences and chloroplast DNA variation. *Molecular Ecology*, **15**, 4065–4083.

Becker T, Knapp M (2002) Efficiency of haplotype frequency estimation when nuclear family information is included. *Human Heredity*, **54**, 45–53.

Brown DG, Harrower IM (2006) Integer programming approaches to haplotype inference by pure parsimony. *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, **3**, 141–154.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, **18**, 249–256.

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

Clark AG, Weiss KM, Nickerson DA *et al*. (1998) Haplotype

structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, **63**, 595–612.

Clayton D (2001) SNPHAP: a program for estimating frequencies of large haplotypes of single nucleotide polymorphisms. http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt.

Cohen JC, Kiss RS, Pertsemlidis A *et al*. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.

Crandall KA, Templeton AR (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, **134**, 959–969.

Eronen L, Geerts F, Toivonen H (2004) A Markov chain approach to reconstruction of long haplotypes. In: *Proceedings of the 9th Pacific Symposium on Biocomputing* (eds Altman RB, Dunker AK, Hunter L, Jung TA, Klein TE), pp. 104–115. World Scientific, Singapore.

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

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

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.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–927.

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, **67**, 947–959.

Greenspan G, Geiger D (2004) Model-based inference of haplotype block variation. *Journal of Computational Biology*, **11**, 493–504.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

Gusfield D (2001) Inference of haplotypes from samples of diploid populations: complexity and algorithms. *Journal of Computational Biology*, **8**, 305–323.

Gusfield D (2002) Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In: *Proceedings of the 6th International Conference on Computational Molecular Biology* (eds Myers G, Hannenhalli S, Istrail S, Pevzner P, Waterman M), pp. 166–175. ACM Press, New York.

Gusfield D (2003) Haplotype inference by pure parsimony. In: *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching* (eds Baeza-Yates R, Chavez E, Crochemore M), Lecture Notes in Computer Science, 2676, pp. 144–155. Springer-Verlag, Berlin.

Halldorsson BV, Bafna V, Edwards N *et al*. (2004) A survey of computational methods for determining haplotypes. In: *Computational Methods for SNPs and Haplotype Inference* (eds Istrail S, Waterman M, Clark A), Lecture Notes in Computer Science, 2983, pp. 26–47. Springer-Verlag, Berlin.

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. *Journal of Heredity*, **86**, 409–411.

Hoehe MR, Kopke K, Wendel B *et al*. (2000) Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. *Human Molecular Genetics*, **9**, 2895–2908.

Huang YT, Chao KM, Chen T (2005) An approximation algorithm for haplotype inference by maximum parsimony. *Journal of Computational Biology*, **12**, 1261–1274.

Hurley JD, Engle LJ, Davis JT, Welsh AM, Landers JE (2005) A simple, bead-based approach for multi-SNP molecular haplotyping. *Nucleic Acids Research*, **32**, e186.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

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

Kirk KM, Cardon LR (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, **10**, 616–622.

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**, 234–236.

Lancia G, Bafna V, Istrail S, Lippert R, Schwartz R (2001) SNPs problems, complexity, and algorithms. In: *Proceedings of the Ninth Annual European Symposium on Algorithms* (ed. Meyer auf der Heide F), Lecture Notes in Computer Science, 2161, pp. 182–193. Springer-Verlag, Berlin.

Lancia G, Pinotti MC, Rizzi R (2004) Haplotyping populations by pure parsimony: complexity of exact and approximation algorithms. *Informs Journal on Computing*, **16**, 348–359.

Landwehr N, Mielikainen T, Eronen L, Toivonen H, Mannila H (2007) Constrained hidden Markov models for population-based haplotyping. *BMC Bioinformatics*, **8**, S9.

Li SS, Khalid N, Carlson C, Zhao LP (2003) Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, **4**, 513–522.

Li Z, Zhou W, Zhang XS, Chen L (2005) A parsimonious tree-grow method for haplotype inference. *Bioinformatics*, **21**, 3475–3481.

Lin S, Chakravarti A, Cutler DJ (2004) Haplotype and missing data inference in nuclear families. *Genome Research*, **14**, 1624–1632.

Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. *American Journal of Human Genetics*, **71**, 1129–1137.

Lippert R, Schwartz R, Lancia G, Istrail S (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, **3**, 23–31.

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, **56**, 799–810.

Marchini J, Cutler D, Patterson N *et al*. (2006) A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics*, **78**, 437–450.

Mountain JL, Knight A, Jobin M *et al*. (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Research*, **12**, 1766–1772.

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.

Olsen KM, Schaal BA (1999) Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proceedings of the National Academy of Sciences, USA*, **96**, 5586–5591.

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.

Patil N, Berno AJ, Hinds DA *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, **69**, 124–137.

Qin ZS, Niu T, Liu JS (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, **71**, 1242–1247.

Rohde K, Fuerst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Human Mutation*, **17**, 289–295.

Ryynanen HJ, Primmer CR (2004) Distribution of genetic variation in the growth hormone 1 gene in Atlantic salmon (*Salmo salar*) populations from Europe and North America. *Molecular Ecology*, **13**, 3857–3869.

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

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.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schaid DJ (2002) Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genetic Epidemiology*, **23**, 426–443.

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.

Sharan R, Halldorsson BV, Istrail S (2006) Islands of tractability for parsimony haplotyping. *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, **3**, 303–311.

Sotka EE, Wares JP, Barth JA, Grosberg RK, Palumbi SR (2004) Strong genetic clines and geographical variation in gene flow in the rocky intertidal barnacle *Balanus glandula*. *Molecular Ecology*, **13**, 2143–2156.

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76**, 449–462.

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.

Taylor MS, Hellberg ME (2006) Comparative phylogeography in a genus of coral reef fishes: biogeographic and genetic concordance in the Caribbean. *Molecular Ecology*, **15**, 695–707.

Templeton AR (2004) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.

Thomas A (2003a) Accelerated gene counting for haplotype frequency estimation. *Annals of Human Genetics*, **67**, 608–612.

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

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *American Journal of Human Genetics*, **67**, 518–522.

Tost J, Brandt O, Boussicault F *et al*. (2002) Molecular haplotyping at high throughput. *Nucleic Acids Research*, **30**, e96.

Wang L, Xu Y (2003) Haplotype inference by maximum parsimony. *Bioinformatics*, **19**, 1773–1780.

Xing EP, Sohn KA, Jordan MI, Teh YW (2006) Bayesian mutipopulation haplotype inference via a hierarchical Dirichlet process mixture. In: *Proceedings of the 23rd International Conference on Machine Learning* (eds Cohen W, Moore A), pp. 1049–1056. ACM Press, New York.

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 (1998) Isolation of animal cellular total DNA. In: *Molecular Tools for Screening Biodiversity: Plants and Animals* (eds Karp A, Isaac PG, Ingram DS), pp. 5–9. Chapman & Hall, London.

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

Zhang Y, Niu T, Liu JS (2006) A coalescence-guided hierarchical Bayesian method for haplotype inference. *American Journal of Human Genetics*, **79**, 313–322.

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.

Zhang DX, Yan LN, Ji YJ *et al*. (2003) Isolation, characterization and cross-species amplification of eight microsatellite DNA loci in the migratory locust (*Locusta migratoria*). *Molecular Ecology Notes*, **3**, 483–486.

Zu-Shi Huang is a PhD student working on phylogeography and population genetics of the migratory locust with nuclear DNA markers. Ya-Jie Ji is an associate research professor focusing on molecular evolution and application of DNA markers. De-Xing Zhang's research group studies the consequence of past global changes in China applying evolutionary and phylogeographical approaches to small vertebrates, insects and other invertebrates as model organisms. Population dynamics and geographical interaction of agricultural and forestry pest insects are another important research focus of this laboratory.

Supplementary material

The following supplementary material is available for this article:

Fig. S1 Frequency distribution of the 141 distinct genotypes in the 526 migratory locusts and the 68 haplotypes of the CV solution of the total sample.

Fig. S2 Effects of haplotype diversity of populations on statistical haplotype inference.

Fig. S3 Inference errors of the six haplotype reconstruction methods and the CV approach in the 31 populations investigated.

Table S1 Information on population samples of *Locusta migratoria* in this study

Table S2 Information on the polymorphic sites observed at the scnpc76 locus of *Locusta migratoria*

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-294X.2008.03730.x
(This link will take you to the article abstract).

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