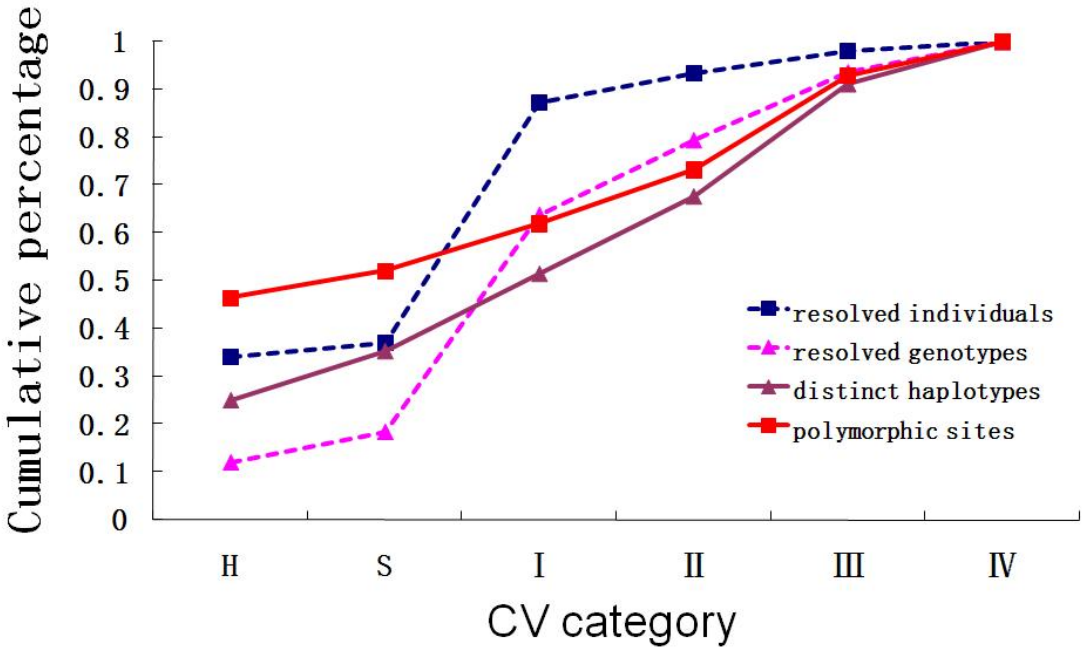
CVhaplot

(Version 2.01)

A Perl script package to implement the consensus vote (CV) approach

Zu-Shi Huang & De-Xing Zhang



August 24, 2009

Institute of Zoology

Chinese Academy of Sciences

Beijing

User Guide

- **CVhaplot version 2.01** -
---

1. Introduction

Haplotype inference from population genotypic data is a complex statistical problem, showing considerable internal algorithm variability and among-algorithm discordance (Huang *et al*., 2009). Recently, Huang *et al*. (2008) have explored the consensus vote (CV) approach to increase the confidence of statistical haplotyping results. This approach places its emphasis on examining discordance among independent algorithms and identifying uncertain inferrals in the solutions. Alternatively, halpotype inference uncertainty can also be reduced by controlling the internal variability of individual algorithms (e.g. Orzack *et al*., 2003). CVhaplot has been developed to combine these two complementary approaches and automate the analysis procedure.

CVhaplot is a small package of Perl scripts. It has the following features:

- reformats the data into the input file formats of several popular algorithms for haplotype reconstruction: PHASE (Stephens *et al.*, 2001; Stephens & Donnelly 2003), HAPLOTYPER (Niu *et al.*, 2002), HAPLOREC (Eronen *et al.*, 2004, 2006), ARLEQUIN-EM (Excoffier *et al.*, 2005), GCHAP (Thomas, 2003a, b), GERBIL (Kimmel and Shamir, 2005), and HAPINFERX (Clark, 1990);
- facilitates the realization of multiple iterations of these algorithms;
- extends the applicability of HAPLOTYPER, GCHAP and GERBIL to deal with triallelic sites with a coding switch technique (i.e. coding a triallelic site as two biallelic ones);
- provides various indices to evaluate the consistency of individual algorithms;
- identifies a HAPINFERX ensemble solution with high accuracy from multiple independent HAPINFERX iterations to overcome its deficiency of high inconsistency;
- generates CV solutions according to consensus rules;
- identifies uncertain haplotypes in the data;
- identifies samples that show any mismatch between the inferred and original genotypes after the CV analysis.

Platform requirements

Windows XP/Vista, Linux and Mac OS X implemented with Perl version 5.10.0 or later versions. To see if the computer has Perl 5.10.0, type "perl -version" on the command line.

2. Installation

No special installation action is required. The program includes three Perl scripts. Once they have been copied to a local directory (e.g. C:/perl), these scripts can be run separately with the following commands (Table 1):

Table 1. The Perl scripts of CVhaplot and their launching commands.

Scripts	Functions	Commands
trans.pl	Reformat a PHYLIP file into the formats of seven algorithms	trans.pl inputfile integer1 integer2 integer3 integer4 integer5[1]
consistency.pl	Consistency test and identify a HAPINFERX ensemble solution with high accuracy	consistency.pl
CV.pl	Consensus vote	CV.pl genotypicfile[2]

Note:

[1]Inputfile refers to the name of the raw genotype data file in sequential PHYLIP format, and the five integers correspond to the number of independent iterations of the five programsms HAPINFERX, HAPLOREC, PHASE, HAPLOTYPER and ARLEQUIN-EM, respectively. The default values of the five integers are 100, 10, 10, 10 and 10, respectively.

[2]The name of the raw genotype data file in PHYLIP format.

The programs HAPINFERX and GCHAP are included in the current version of CVhaplot (We thank Dr. A. G. Clark and Dr. A. Thomas for their permission; users please make proper citations of their work). The other five programs need to be downloaded by users from their websites. Users should replace the empty files with the corresponding names in CVhaplot package with the downloaded programs HAPLOREC and GERBIL. With ARLEQUIN, PHASE and HAPLOTYPER, users should place their input files under the same directory as the program. Then the corresponding batch files ("batchsoftwarewin.bat" for HAPINFERX, GCHAP, HAPLOREC and GERBIL, "batchsoftwarelinux" for PHASE and HAPLOTYPER, and "batchArlequinEM.arb" for ARLEQUIN) can be run. Note that HAPLOREC is implemented with Java, JDK version 1.5.

3. File formats and the three-step implementation

3.1 Input file for trans.pl

A sequential PHYLIP format of the raw data is required as the input file for running the script trans.pl. The first line of the input file contains the sample number and the

sequence length, separated by space. The rest of the file contains the genotype sequences of individuals. Each sequence starts with a ten-character name followed with the genotype sequence (one line for each sequence, i.e. non-interleaved). The IUPAC codes are used to encode heterozygous sites (i.e. R=A+G; Y=C+T; K=G+T; M=A+C; S=C+G; W=A+T).

An example of genotypic data is as follows:

```
2 20
sample1   CYGATTRAGTCCTAGCTTRM
sample2   CTGATTGAGTCCTAGCTTRC
```

2 = number of samples (sequences);
20 = length of the sequence (20 nt);
sample1: name of the first sample (10 characters maximum).

Note that indels and missing data are not considered by the current version of CVhaplot (i.e. the script trans.pl will not work properly) and should be excluded from the input file. If indels are present, we suggest the user to try some other software (e.g. CHAMPURU, http://www.mnhn.fr/jfflot/champuru/) first to deduce more haplotypes. It is also a useful practice to decrease the complexity of the data if statistical methods fail to infer haplotypes satisfactorily, for example, divide the original genotype sequences into two partitions.

The script trans.pl first produces a text file containing the basic summary statistics of the raw genotype data (e.g. the frequency of each nucleotide state at segregating sites). Then it reformats the input file into the input file formats of seven algorithms (Table 2). These files will be saved into a folder named “transfile”.

Table 2. Commands to start the relevant haplotype reconstruction algorithms (Users should consult the instruction of the individual algorithms for any questions about the commands).

Algorithms	Commands	Batch file*
PHASE2.11	./phase –Sinteger inputfile outputfile (Linux)	batchsoftwarelinux
HAPLOTYPER1.0	./htyperv2 inputfile outputfile snp individual round (Linux)	batchsoftwarelinux
HAPINFERX	hap.pl input output (Perl)	batchsoftwarewin.bat
HaploRec2.3	Java –Xmx1024m -jar HaploRec.jar -od -s -cp -p S -n iteration-number inputfile (Java)	batchsoftwarewin.bat
ARLEQUIN-EM3.11	use the zipper option(Windows)	batchArlequinEM.arb
Gerbil	gerbil.exe inputfile outputfile(Windows)	batchsoftwarewin.bat
GCHap	Java ApproxGCHap inputfile.par inputfile.ped >outputfile(Windows)	batchsoftwarewin.bat

* Batch files are generated by trans.pl for simplifying the launch of the relevant programs.

It is highly recommended that multiple independent iterations are run for the first five algorithms (i.e. except Gerbil and GCHap) to examine their internal variability on the genotype data (see below). The batch files generated by trans.pl (batchsoftwarewin.bat, batchsoftwarelinux and batchArlequinEM.arb) can help to perform the computation. As a general rule, the relevant files and the programs should be located in the same folder. Double-clicking the batch file batchsoftwarewin.bat (or typing its name in command line) allows to automatically launch HAPINFERX, HAPLOREC, GCHAP and GERBIL in Windows and read their input files from the folder "transfile". Their output files are automatically saved in the folder "softwareresults" created by trans.pl. Similarly, the batch file batchsoftwarelinux allows the launch of PHASE and HAPLOTYPER in Linux (by simply typing the name of the batch file. There is no need to type the input file name). The batch file batchArlequinEM.arb allows multiple independent iterations of ARLEQUIN-EM to be performed using the zipper option in Windows (Users should select the zipper option themselves for large data set). Note that the output files of PHASE, HAPLOTYPER and ARLEQUIN-EM (see example) will be saved in the folder named "softwareresults".

3.2 consistency.pl

The script consistency.pl automatically analyzes the output files of multiple iterations of the relevant algorithms in the folder "softwareresults" to perform consistency test. The script uses several indices that measure the performance of each algorithm. In addition, this script will identify a HAPINFERX ensemble solution with high accuracy from those independent iterations whose NDH (number of distinct haplotypes) values are among the smallest**.** This is because HAPINFERX often dispalys largest internal variability among the algorithms employed here (Orzack *et al*., 2003; Huang *et al*., 2008, 2009), which is generally associated with higher error rate (Huang *et al*., 2009). Note that the abnormal solutions of some HAPINFERX iterations (i.e. those that contain unresolved genotypes and/or resolved genotypes with one extra mutation being introduced) are filtered out. The consistency results and the ensemble files are saved into the folder "consistencyresult".

3.3 Input files for CV.pl

An option (y/n) is available with this script to allow the users to decide whether they need to create a folder for saving the solutions of individual programs. If the answer is "n", it would automatically read the solutions of the six algorithms in the folder "softwareresults" and "consistencyresult" (HAPINFERX, PHASE, HAPLOTYPER, HAPLOREC, ARLEQUIN-EM and GCHAP) and perform the consensus vote analysis. Note that there exists another option allowing users to decide whether the solution of GERBIL is included in the consensus vote analysis. The solution of HAPINFERX is the ensemble (consensus) solution of independent iterations whose NDH (see below) values are among the smallest. If the answer is "y", users are allowed to choose for each algorithm, according to the consistency of the algorithm, the iteration whose solutions are

to be included in the subsequent consensus vote analysis. In this case, users should create a new folder for each program under the same directory as CV.pl, and save there the selected solutions using the input file name(s) listed in Table 3. Note that if HAPLOTYPER or GCHAP are included in the CV analysis, users must also copy the file character.txt produced by trans.pl to the folder newly created by users.

Table 3. The corresponding names of input files for CV.pl.

Algorithm	Result file(s)	Input file name(s)[1]
PHASE2.11	summary output, *_pairs	**phase**, phase**_pairs**
HAPLOTYPER1.0	output	**haplotyper**
HAPINFERX	output	**hapinferx** or **hapinfCV**.txt
HaploRec2.3	*.reconstructed	**HaploRec**.reconstructed
ARLEQUIN-EM3.11	input.htm	**Arlequin**.htm
GCHap	output	**gchap**
Gerbil	output.01	**gerbil**.01

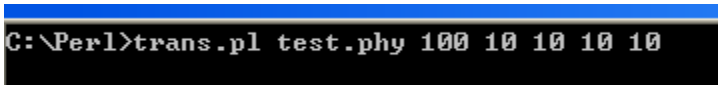
Note:

[1] The bold letters must be included in the input file names for their identification by CV.pl.

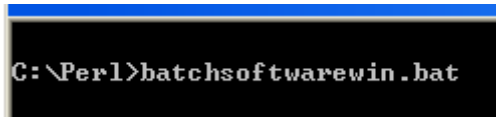
3.4 Implementation of CVhaplot

An example under windows:

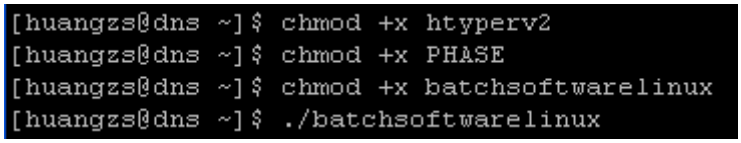
a. Type the command of trans.pl in the command line in Windows;



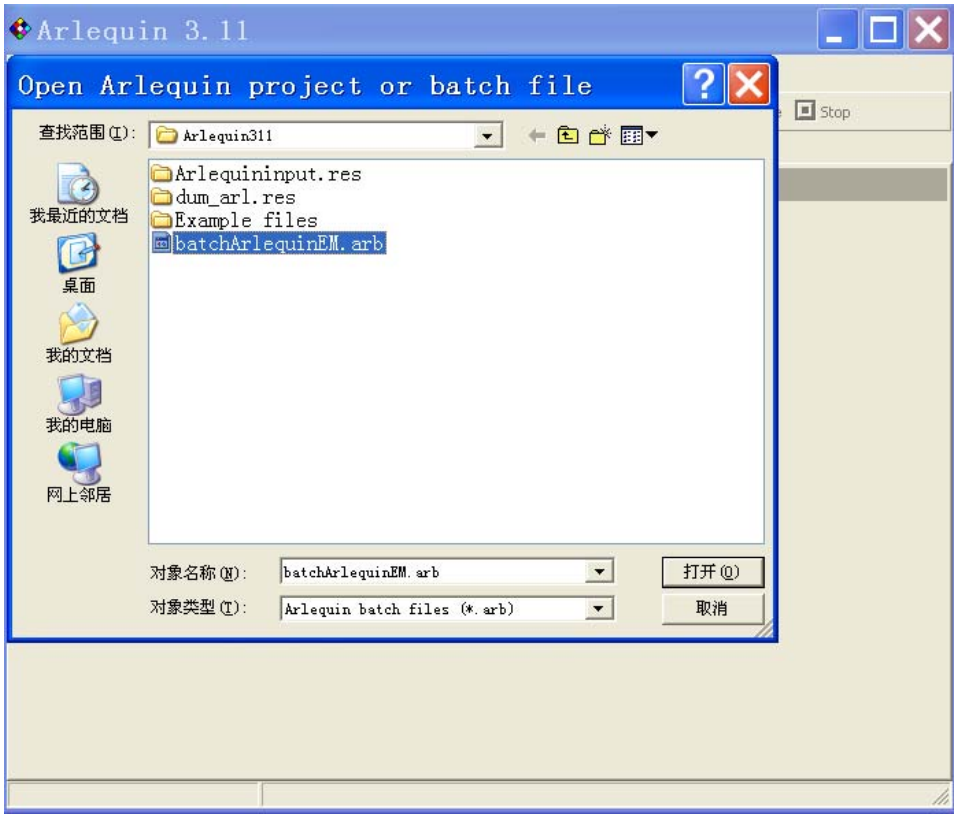
b. Perform independent iterations of HAPINFERX, HAPLOREC, GCHAP, and GERBIL in windows;



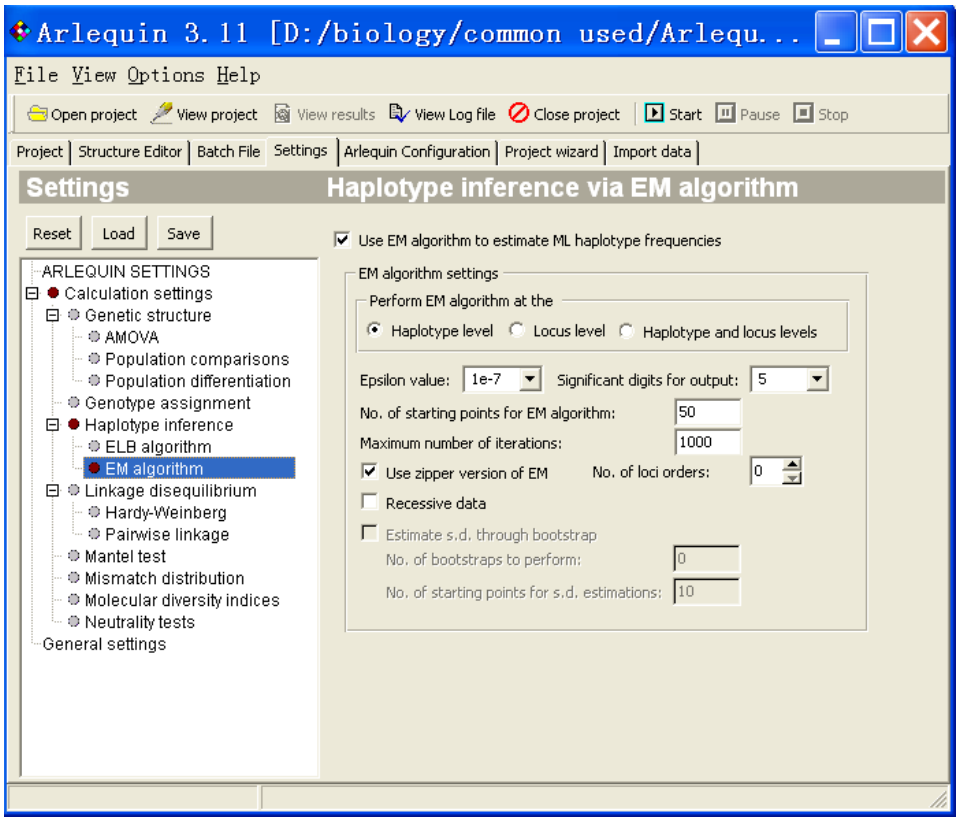
c. Perform independent iterations of HAPLOTYPER and PHASE in linux (the input files, programs and the batchfile should be under the same directory);



d. Perform independent iterations of ARLEQUIN-EM in Windows;



(1) Choose the batch file



(2) EM algorithm settings

e. Type the command of consistency.pl in the command line in Windows;

f. Type the commands of CV.pl in the command line in Windows;

```
C:\Perl\cvhaplot>CV.pl test.phy

**********CVhaplot2.0**********
*    Coded by Zu-Shi Huang    *
*    Institute of Zoology     *
* Chinese Academy of Sciences *
*     huangzs@ioz.ac.cn       *
*********************************

Have you saved the outputfiles of various programs into a new folder to control
which solution of an algorithm is used in the CV analysis?(y/n):n

Do you want to include the solution of GERBIL in the consensus vote analysis?(y/
n):n

Fine, no samples show any mismatch between the inferred and original genotypes!

The consensus vote analysis has finished!
```

4. Output of CVhaplot

The script consistency.pl produces two or four files (Consistency.txt, pairwise.txt, and hapinfCV.txt and hapinfCVtable.txt if HAPINFERX is included). The file Consistency.txt contains four indices (HID, IRD, NLCP, NDH) that evaluate the consistency of each algorithm. Pairwise.txt details the individual-resolving discrepancy between different iterations of each algorithm. If users examine the consistency of HAPINFERX, the file hapinfCVtable.txt will include information of individual-resolving discrepancy among the independent iterations, and hapinfCV.txt includes the ensemble of multiple independent iterations. We recommend users to use the file “hapinfCV.txt”, which was produced by consistency.pl as the default input file of HAPINFERX for CV.pl in consensus vote analysis (Table 3). The meanings of the four indices (HID, IRD, NLCP, NDH) that measure the performance of algorithms are as follows (Huang *et al*., 2008):

HID: Haplotype-inferring discrepancy, i.e. the discrepancy of distinct haplotypes between two solutions.

IRD: Individual-resolving discrepancy, i.e. the proportion (number) of individuals whose genotypes were resolved differently between two solutions.

NLCP: Number of individuals whose haplotype pair is of low confidence probability in a solution based on a simple threshold (e.g. <0.95 for PHASE).
NDH: Number of distinct haplotypes in a solution.

The script CV.pl produces three files (CV.txt, CVsolution.txt and softwaresolutions.txt). CV.txt lists the consensus vote of each individual and the confidence probabilities of solutions from different algorithms. It also summarizes the comparison results. Users can thus design verification experiments for uncertain individuals according to this information. CV.txt also reports samples that show any mismatch between the inferred and original genotypes. The file CVsolution.txt contains the consensus solution calculated following some consensus rules. The consensus rules adequately consider the discordance among algorithms. The file softwaresolutions.txt lists the haplotype solutions (in the same order as in genotypic data) inferred by all the algorithms included in CV analysis.

5. Citation

Please cite the following work:
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.

6. Support

Please send all questions, suggestions, comments and bug reports to Zu-Shi Huang (huangzs@ioz.ac.cn) at the Institute of Zoology, Chinese Academy of Sciences, Beijing.

7. Disclaimer

No guarantees are given as to the suitability of these Perl scripts to a user's data. License of use is granted free for non-commercial use.

8. Version history

2.01 (August 24, 2009)
The new version can now run under Windows XP/vista, Linux and Mac OS X. A warning report bug was fixed: the previous version does not report a warning message when all HAPINFERX runs are abnormal. A bug of the haplotype output of

HAPINFERX has also been fixed: the present version now outputs the inferred haplotypes in the same order as in the original genotypic data.

2.0 (May 15, 2009)
CVhaplot 2.0 introduces more functions and greater flexibility. It generates three batch files to help to perform the independent iterations, and automatically analyzes the output files of various algorithms for consistency test and consensus vote analysis. The revised consensus vote rules further improve the performance of the CV approach. The new version can also identify a HAPINFERX ensemble solution with high accuracy from multiple independent iterations whose NDH (numbers of distinct haplotypes) values are among the smallest. This strategy help to overcome the deficiency of high inconsistency of some algorithms (i.e. HAPINFERX), as demonstrated in Orzack *et al*. (2003). The script CV.pl integrates the inspection function of inspect.pl in the previous version. Two programs (HAPINFERX and GCHAP) were included in the current version of CVhaplot package, and also the new version HAPLOREC2.3 was used in this updated version of CVhaplot.

1.0 (March 2008)
Original implementation of the consensus vote approach was introduced in Huang *et al*., 2008. It reformats the genotypic data into the input file formats of several popular algorithms for haplotype reconstruction, performs the consistency test for inspecting the internal variability of five algorithms (except for GERBIL and GCHAP), and performs the consensus vote analysis. This version only displays the discordance among algorithms and generates haplotype data for individual algorithms. It does not list out the consensus vote solution. It allows user to use an ensemble of random iterations of HAPINFERX as its input solution to CV.pl. It also includes a Perl script (inspect.pl) to identify samples that show any mismatch between the inferred and original genotypes after the CV analysis.

9. Key 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 (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, 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.

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

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.

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.

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.

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.

10. Websites of individual algorithms

Arlequin, http://cmpg.unibe.ch/software/arlequin3/

GCHap, http://www-genepi.med.utah.edu/~alun/software/

GERBIL, http://acgt.cs.tau.ac.il/gevalt/

HaploRec, http://www.cs.helsinki.fi/group/genetics/haplotyping.html

HAPLOTYPER, http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm

PHASE, http://www.stat.washington.edu/stephens/software.html

Appendix 1

Table A1. List of the output files from the three Perl scripts.

Perl scripts	Output files	Descriptions
trans.pl	character.txt	Lists the frequency of nucleotide states at each site of the raw data
	phaseinput	Input file of PHASE
	HAPLOTYPERinput	Input file of HAPLOTYPER
	hapinferx_input_all_sites.txt	Input file of HAPINFERX including all sites
	hapinferx_input_variable_sites.txt	Input file of HAPINFERX including polymorphic sites only
	HaploRecinput	Input file of HaploRec
	Arlequininput.arp	Input file of Arlequin
	Gerbilinput	Input file of Gerbil
	GCHapinput.par	Input files of GCHap
	GCHapinput.ped	
	batchsoftwarewin.bat	batch file for launching HAPINFERX, GCHAP, GERBIL and HAPLOREC in Windows
	batchsoftwarelinux	batch file for launching PHASE and HAPLOTYPER in Linux
	batchArlequinEM.arb	batch file for launching ARLEQUIN-EM in Windows
consistency.pl	Consistency.txt	Various indices measuring the consistency of algorithms
	pairwise.txt	Gives the discrepancy between solutions from different iterations
	hapinfCV.txt	Produces the HAPINFERX ensemble solution with high accuracy from multiple independent iterations
	hapinfCVtable.txt	Details the discordance of iterations of HAPINFERX
CV.pl	CV.txt	Summaries of the consensus vote information, including samples that show any mismatch between the inferred and original genotypes
	CVsolution.txt	Lists the consensus solutions of the CV analysis
	softwaresolutions.txt	Lists the solution of haplotypes of individual algorithms

Appendix 2

Table A2. List of files in the package CVhaplot

Contents	Files in sub-directory	Descriptions
User Guide	CVhaplotHelp.pdf	Program manual
trans.pl		Summarize the characterization of the raw data , and convert it to the input formats of seven algorithms
consistency.pl		Perform the consistency test for the algorithms HAPINFERX, PHASE, HAPLOTYPER, HAPLOREC, ARLEQUIN-EM
CV.pl		Perform the consensus vote approach
example	test.phy	Example data file in PHYLIP non-interleaved format
	softwareresults	Subdirectory containing the output files of individual algorithms
	transfile	Subdirectory containing the output files of trans.pl (refer to Table A1 for the inside files)
	consistencyresult	Subdirectory containing the output files of consistency.pl (refer to Table A1 for the inside files)
	Consensusvote	Subdirectory containing the output files of CV.pl (refer to Table A1 for the inside files)
	user_created_CV_folder	Subdirectory containing the input and output files of CV.pl (refer to Table A1 for the inside files)
	batchsoftwarewin.bat	batch file for launching HAPINFERX, GCHAP, GERBIL and HAPLOREC in Windows
	batchsoftwarelinux	batch file for launching PHASE and HAPLOTYPER in Linux
	batchArlequinEM.arb	batch file for launching ARLEQUIN-EM in Windows
alun, ApproxGCHap.class ApproxGCHap.java GCHap.calss GCHap.java		the program GCHAP (Thomas 2003a,b)
hap.pl		the program HAPINFERX (Clark 1990)

Appendix 3

The flow chart of CVhaplot