MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2010) **10**, 162–166 doi: 10.1111/j.1755-0998.2009.02732.x

COMPUTER PROGRAM NOTE

SEQPHASE: a web tool for interconverting PHASE input/output files and FASTA sequence alignments

J-F. FLOT*†

*UMR UPMC-CNRS-MNHN-IRD 7138, Département Systématique et Évolution, Muséum National d'Histoire Naturelle, Case Postale 26, 57 rue Cuvier, 75231 Paris Cedex 05, France

Abstract

The program PHASE is widely used for Bayesian inference of haplotypes from diploid genotypes; however, manually creating PHASE input files from sequence alignments is an error-prone and time-consuming process, especially when dealing with numerous variable sites and/or individuals. Here, a web tool called SEQPHASE is presented that generates PHASE input files from FASTA sequence alignments and converts PHASE output files back into FASTA. During the production of the PHASE input file, several consistency checks are performed on the dataset and suitable command line options to be used for the actual PHASE data analysis are suggested. SEQPHASE was written in PERL and is freely accessible over the Internet at the address http://www.mnhn.fr/jfflot/seqphase.

Keywords: FASTA, haplotyping, PHASE, statistical haplotyping

Received 5 March 2009; revision accepted 17 April 2009

Introduction

Using nuclear sequence markers for phylogenetic or population genetic studies requires obtaining haplotype sequences from a large number of diploid individuals, a difficult task for which a number of molecular (reviewed in Kwok & Xiao 2004), combinatorial (reviewed in Lancia 2008) and statistical (e.g. Excoffier & Slatkin 1995; Stephens *et al.* 2001) methods can be used. Among these methods, statistical haplotyping using PHASE (Stephens *et al.* 2001) is probably the single most popular with 1982 citations in ISI Web of Science and 2359 citations in Google Scholar as of March 1st, 2009, and has been shown to represent a reliable alternative to cloning (Xu *et al.* 2002; Bos *et al.* 2007; Harrigan *et al.* 2008).

However, PHASE is not particularly geared towards sequence data and requires complex input files that are difficult to generate from sequence alignments: first, PHASE does not accept letter codes at positions for which more than two possible nucleotides are found, only numbers; second, to shorten computing times, constant positions should be omitted from the main input file and positions for which only two possible nucleotides are found should be singled out and treated differently (S mode vs. M mode); and third, if known phases are available, a secondary input file (.known) has to be generated specifying, for each character of each individual, whether the phase is already known or needs to be inferred. As a result, manually creating PHASE input files is often an error-prone and time-consuming process, especially when dealing with numerous variable sites and/or individuals; moreover, converting back PHASE output files into sequence alignments while trying to account for the posterior probabilities of each nucleotide or haplotype in a synthetic way can also be a daunting task. In spite of these drawbacks, the PHASE input/output format has become a standard adopted by other haplotyping programs (e.g. Delaneau *et al.* 2007) to ensure compatibility; hence the need for a simple, user-friendly web platform performing automated generation of PHASE input files from FASTA alignments and automatic conversion of PHASE output files back into FASTA.

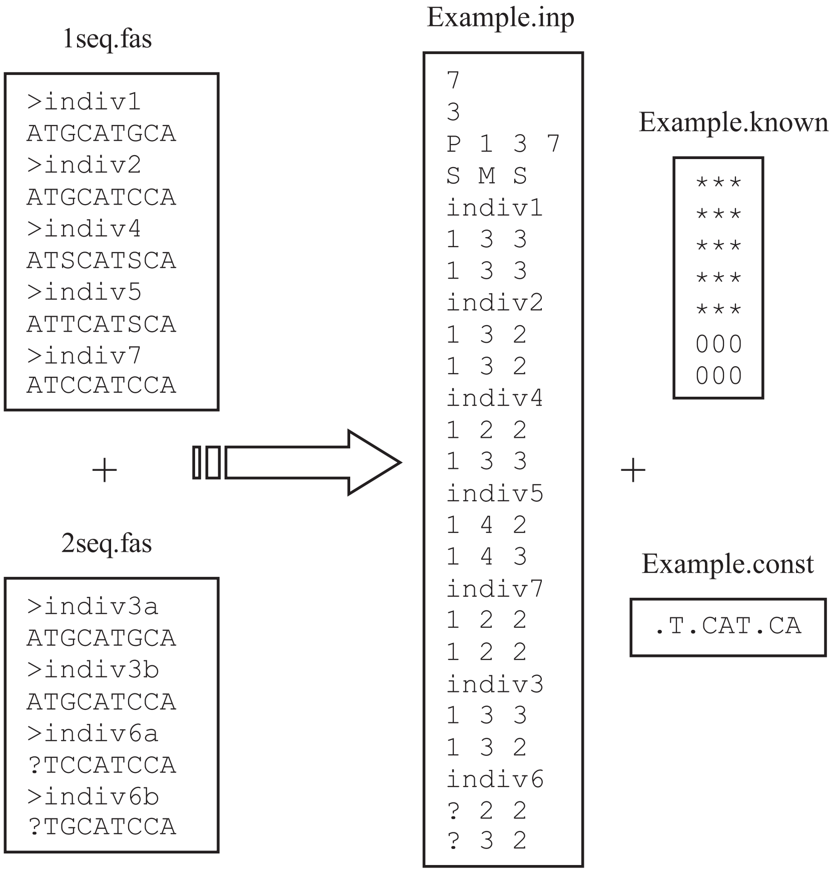
First step: FASTA to PHASE (Fig. 1)

Instead of generating PHASE input files from a single FASTA alignment, which would require users to specify manu-

Correspondence: Jean-François Flot, Fax: +49 551 39 7918; E-mail: jflot@uni-goettingen.de

†Present address: Georg-August-Universität Göttingen, Geowissenschaftliches Zentrum, Courant Research Centre Geobiology, Goldschmidtstr. 3, 37077 Göttingen, Germany

Fig. 1 From two FASTA alignments, SEQPHASE generates up to three files, two of which (*.inp and *.known) are used to run PHASE.



```
>indiv1
ATGCATGCA
>indiv2
ATGCATCCA
>indiv4
ATSCATSCA
>indiv5
ATTCATSCA
>indiv7
ATCCATCCA
```

```
>indiv3a
ATGCATGCA
>indiv3b
ATGCATCCA
>indiv6a
?TCCATCCA
>indiv6b
?TGCATCCA
```

```
7
3
P 1 3 7
S M S
indiv1
1 3 3
1 3 3
indiv2
1 3 2
1 3 2
indiv4
1 2 2
1 3 3
indiv5
1 4 2
1 4 3
indiv7
1 2 2
1 2 2
indiv3
1 3 3
1 3 2
indiv6
? 2 2
? 3 2
```

```
***
***
***
***
***
000
000
```

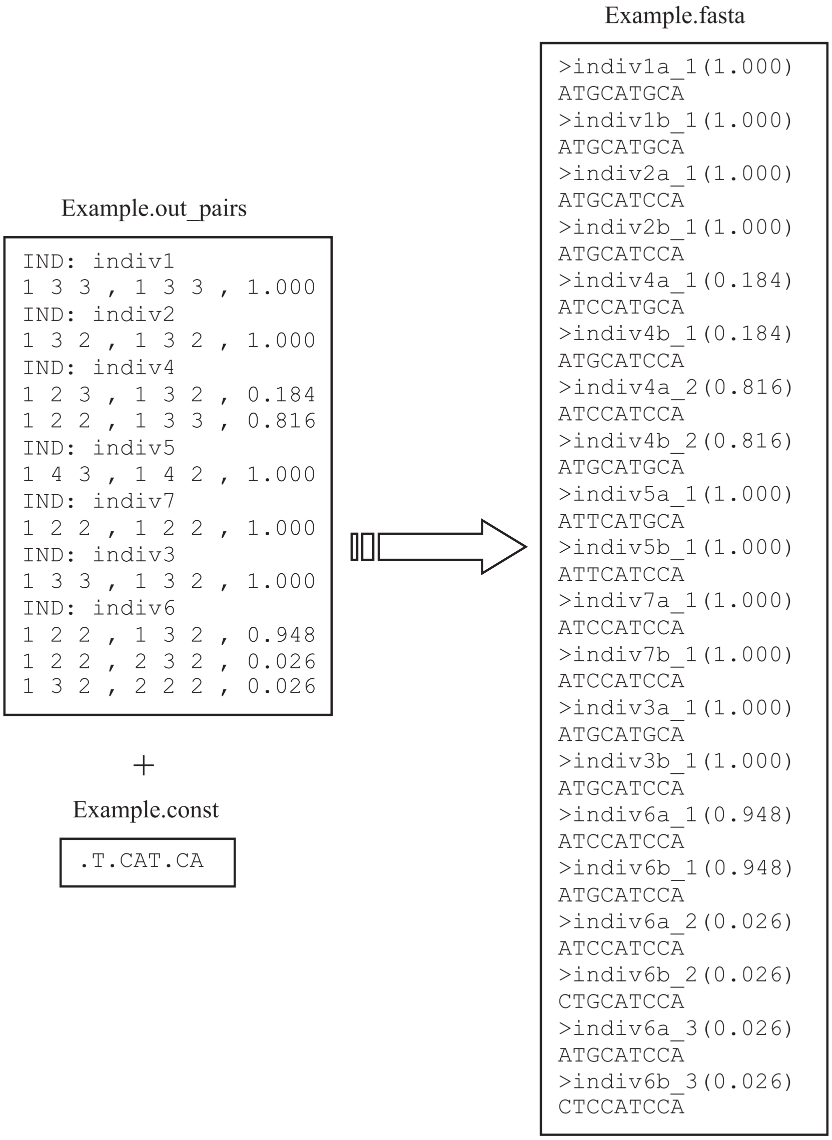
```
.T.CAT.CA
```

ally what phases are already known (for instance from cloning) and what individuals need to be phased, SEQPHASE takes as input two separate FASTA files: one for homozygous individuals and heterozygotes to be phased (with one sequence per individual), and a second one for heterozygotes whose phases are already known (with two sequences per individual). In the alignment of phased heterozygotes, the names of the two sequences of each individual should differ only by their last character (e.g. 'indiv3a' and 'indiv3b'). Heterozygous individuals whose two haplotypes differ only by one substitution or insertion/deletion can be indifferently entered in the first or the second alignment since haplotyping is trivial in such case.

After users hit the Submit button on the web form, SEQPHASE starts by verifying that all sequences in the alignments have the same length, that they contain only authorized symbols (namely, A, T, G, C, R, W, M, Y, S, K, N, ? and —) and that all sequence names are different. It then removes constant positions, inventories variable positions for which more than two possible nucleotides are found and creates up to three files: one *.inp necessary to run PHASE, one *.known detailing which phases are known (if any) and one *.const recording the constant positions that were removed from the alignment, if any. PHASE does not accept letters for multistate characters, which is why nucleotides are written into the *.inp file as numbers based on alphabetical order as a mnemonic: ? or −1 for missing information (depending on whether the position displays two or more than two different nucleotides), 0 for indel, 1 for A, 2 for C, 3 for G and 4 for T.

Second step: PHASE to FASTA (Fig. 2)

As statistical haplotyping can require long running times depending on the characteristics of the dataset and the analysis options used, offering a fully online PHASE data pipeline is not an option at the present time; hence, the analysis itself has to be performed locally by the user on the input files downloaded from the SEQPHASE website. After this analysis is completed, the *.out or *.out_pairs output file produced by PHASE and the *.const file created during the first step can be uploaded using a second web interface to generate a FASTA alignment of haplotype sequences containing both variable and constant positions in letter codes instead of numbers (if the *.const file box is left empty, a FASTA file containing only variable positions is returned).



Example.out_pairs

```
IND: indiv1
1 3 3 , 1 3 3 , 1.000
IND: indiv2
1 3 2 , 1 3 2 , 1.000
IND: indiv4
1 2 3 , 1 3 2 , 0.184
1 2 2 , 1 3 3 , 0.816
IND: indiv5
1 4 3 , 1 4 2 , 1.000
IND: indiv7
1 2 2 , 1 2 2 , 1.000
IND: indiv3
1 3 3 , 1 3 2 , 1.000
IND: indiv6
1 2 2 , 1 3 2 , 0.948
1 2 2 , 2 3 2 , 0.026
1 3 2 , 2 2 2 , 0.026
```

+

Example.const

```
.T.CAT.CA
```

Example.fasta

```
>indiv1a_1(1.000)
ATGCATGCA
>indiv1b_1(1.000)
ATGCATGCA
>indiv2a_1(1.000)
ATGCATCCA
>indiv2b_1(1.000)
ATGCATCCA
>indiv4a_1(0.184)
ATCCATGCA
>indiv4b_1(0.184)
ATGCATCCA
>indiv4a_2(0.816)
ATCCATCCA
>indiv4b_2(0.816)
ATGCATGCA
>indiv5a_1(1.000)
ATTCATGCA
>indiv5b_1(1.000)
ATTCATCCA
>indiv7a_1(1.000)
ATCCATCCA
>indiv7b_1(1.000)
ATCCATCCA
>indiv3a_1(1.000)
ATGCATGCA
>indiv3b_1(1.000)
ATGCATCCA
>indiv6a_1(0.948)
ATCCATCCA
>indiv6b_1(0.948)
ATGCATCCA
>indiv6a_2(0.026)
ATCCATCCA
>indiv6b_2(0.026)
CTGCATCCA
>indiv6a_3(0.026)
ATGCATCCA
>indiv6b_3(0.026)
CTCCATCCA
```

Fig. 2 The third file (*.const) is later used by SEQPHASE to convert PHASE output files *.out or *.out_pairs back into FASTA.

If a *.out file is uploaded, phased haplotypes are shown with 1-letter indetermination code letters (R, W, M, Y, S, or K) at positions where phase certainty is below a certain threshold (90% using PHASE default running options; this probability threshold can be modified by running PHASE using the -p and -q options, see PHASE documentation), whereas if a *.out_pairs file is uploaded, a list of all possible haplotype pairs for each individual is returned as FASTA with their respective probabilities indicated between parentheses (since FASTA alignments normally cannot accommodate two sequences bearing exactly the same name, the two haplotypes of each newly phased individual receive this individual's name with 'a' or 'b' appended). The resulting alignment comprises all sequences in the dataset and can be used directly as input for standard phylogenetic or population genetic analyses.

Implementation and availability

SEQPHASE is a web application composed of two web forms, two PERL scripts, a documentation page and a few example input files. Thanks to its minimalist interface, it can be used over a slow Internet connection and is thus accessible from anywhere around the globe. SEQPHASE automatically adjusts its procedure to the characteristics of the dataset submitted and generates the minimal number of files required to run PHASE; for instance, if no alignment of phased haplotypes is uploaded, SEQPHASE generates no *.known file and suggests a suitable, simplified syntax for running PHASE in the most efficient way. SEQPHASE outputs detailed information regarding the data processing it performs and what should be done with each file it generates, making haplotyping using PHASE easily accessible to nonspecialists and first-time users.

SEQPHASE is freely available over the Internet at the address http://www.mnhn.fr/jfflot/seqphase; advanced users may obtain compiled command-line executables for a variety of platforms (Linux, Win32, Mac OS X, Solaris, etc.) upon email request to the author.

Current limitations

SEQPHASE was created with in mind the phasing of sequences obtained from direct sequencing of nuclear markers. As a result, it assumes that all nucleotides in the input alignment are actually contiguous and considers the locus position for each variable site to be its actual position in the alignment. Although this may not always be the case (for instance when dealing with SNPs very distant from one another), there is presently no way to input a different list of loci positions other than by manually editing the PHASE input file produced by SEQPHASE (the positions of the variable sites are listed in the third line of the input file following the letter P); however, such feature may be added later on if requested by users.

Another issue concerns length-variant heterozygotes, i.e. individuals whose haplotypes are of different lengths because of the presence of one or several indels (Creer *et al.* 2005; Flot *et al.* 2006). Since there is presently no set of IUB codes defined to represent 'A or indel', 'C or indel', 'G or indel' and 'T or indel' (Nomenclature Committee of the International Union of Biochemistry 1985), the genotypes of such length-variant heterozygotes are impossible to code in SEQPHASE. A possible workaround is to input the data for each length-variant heterozygote as two fictitious haplotypes at the end of the second alignment, then to manually edit the *.known file and replace 0 (meaning 'known phase') in the corresponding lines by * (meaning 'unknown phase'). However, a better way that does not require any tampering with the data files is to phase all length-variant heterozygotes *prior* to running PHASE, using programs such as CHAMPURU (Flot 2007), TRACEHAPLOTYPER (Seroussi & Seroussi 2007) or INDELLIGENT (Dmitriev & Rakitov 2008) that analyse the patterns of double peaks in the chromatograms obtained from direct sequencing.

Comparison with DnaSP

Since its version 4.50, the versatile Windows program DnaSP (Rozas & Rozas 1995, 1997, 1999; Rozas *et al.* 2003) includes an implementation of PHASE that takes as input a FASTA alignment (of one sequence per individual, with heterozygous positions represented by IUPAC ambiguity codes), and can write the phased haplotypes resulting from the analysis into a FASTA file (with two haplotypes per individual). Since this function is poorly documented in DNASP's help and has not been described yet in a publication, the implementation of PHASE in versions 4.50.3 (of 1 May 2008) and 4.90.1 (of 10 February 2009, the latest version available) was examined to determine whether DNASP represents a valuable 'all-in-one' alternative to SEQPHASE.

No difference in the implementation of PHASE was found between versions 4.50.3 and 4.90.1, and the menus and options available were nearly identical. Detailed examination revealed one major issue and two shortcomings with the present implementation of PHASE in DNASP. First and foremost, the input format for this implementation (a single FASTA file with one sequence per individual) makes it is impossible to input correctly the haplotypes of individuals whose phases are already known; each known phase included in the input FASTA is interpreted by DNASP as being found in an homozygous state, and PHASE gets run on an erroneous dataset comprising more individuals than in reality since each phased heterozygote is counted twice. As a result, the haplotype frequencies computed by PHASE are incorrect, which may also lead to errors in phase reconstruction (Stephens *et al.* 2001). In addition, DNASP does not generate automatically the *.known file necessary for PHASE to differentiate between known phases and individuals to be phased; instead, this file composed of one line per individual and one character per allele needs to be created 'by hand' by the user. And finally, although DNASP allows the user to decide the threshold confidence level for haplotype reconstruction, it does not flag haplotypes or positions that fall below this confidence threshold (neither in data view mode nor in the exported FASTA file): to find out the actual significance of the result of the PHASE analysis, DNASP users therefore need to examine carefully the posterior probabilities for each allele and haplotype in the PHASE output files.

In its present state, the implementation of PHASE in DNASP is thus far from satisfying, and one can but hope that these problems will be solved in future versions of this program. Even in such case, SEQPHASE will remain a useful alternative since it is platform-independent, is available both as a web-based graphical interface and as a command line executable, and can be used in conjunction with all programs that have the same input/output format as PHASE.

Acknowledgement

Thanks to Michael Hellberg and Ron Eytan for useful discussion and for field-testing earlier versions of this program.

References

Bos DH, Turner SM, DeWoody JA (2007) Haplotype inference from diploid sequence data: evaluating performance using non-neutral MHC sequences. *Hereditas*, **144**, 228–234.

Creer S, Malhotra A, Thorpe RS, Pook CE (2005) Targeting optimal introns for phylogenetic analyses in non-model taxa: experimental results in Asian pitvipers. *Cladistics*, **21**, 390–395.

Delaneau O, Coulonges C, Boelle P-Y, Nelson G, Spadoni J-L, Zagury J-F (2007) ISHAPE: new rapid and accurate software for haplotyping. *BMC Bioinformatics*, **8**, 205.

Dmitriev DA, Rakitov RA (2008) Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Computational Biology*, **4**, e1000113.

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

Flot J-F (2007) Champuru 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. *Molecular Ecology Notes*, **7**, 974–977.

Flot J-F, Tillier A, Samadi S, Tillier S (2006) Phase determination from direct sequencing of length-variable DNA regions. *Molecular Ecology Notes*, **6**, 627–630.

Harrigan RJ, Mazza ME, Sorenson MD (2008) Computation vs. cloning: evaluation of two methods for haplotype determination. *Molecular Ecology Resources*, **8**, 1239–1248.

Kwok P-Y, Xiao M (2004) Single-molecule analysis for molecular haplotyping. *Human Mutation*, **23**, 442–446.

Lancia G (2008) The phasing of heterozygous traits: Algorithms and complexity. *Computers and Mathematics with Applications*, **55**, 960–969.

Nomenclature Committee of the International Union of Biochemistry (1985) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Biochemical Journal*, **229**, 281–286.

Rozas J, Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Computer Applications in the Biosciences*, **11**, 621–625.

Rozas J, Rozas R (1997) DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. *Computer Applications in the Biosciences*, **13**, 307–311.

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174–175.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Seroussi Y, Seroussi E (2007) TraceHaplotyper: using direct sequencing to determine the phase of an indel followed by biallelic SNPs. *BioTechniques*, **43**, 452–456.

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

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