BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 11 2006, pages 1399–1401
doi:10.1093/bioinformatics/btl136

Genetics and population analysis

SNAP: Combine and Map modules for multilocus population genetic analysis

David L. Aylor[1,2], Eric W. Price[2] and Ignazio Carbone[2,*]

[1]Bioinformatics Research Center and [2]Center for Integrated Fungal Research, Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA

Received on February 16, 2006; revised on March 31, 2006; accepted on April 4, 2006
Advance Access publication April 6, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: We have added two software tools to our Suite of Nucleotide Analysis Programs (SNAP) for working with DNA sequences sampled from populations. SNAP Map collapses DNA sequence data into unique haplotypes, extracts variable sites and manipulates output into multiple formats for input into existing software packages for evolutionary analyses. Map collapses DNA sequence data into unique haplotypes, extracts variable sites and manipulates output into multiple formats for input into existing software packages for evolutionary analyses. Map includes novel features such as recoding insertions or deletions, including or excluding variable sites that violate an infinite-sites model and the option of collapsing sequences with corresponding phenotypic information, important in testing for significant haplotype–phenotype associations. SNAP Combine merges multiple DNA sequence alignments into a single multiple alignment file. The resulting file can be the union or intersection of the input files. SNAP Combine currently reads from and writes to several sequence alignment file formats including both sequential and interleaved formats. Combine also keeps track of the start and end positions of each separate alignment file allowing the user to exclude variable sites or taxa, important in creating input files for multilocus analyses.

Availability: SNAP Combine and Map are freely available at http://snap.cifr.ncsu.edu/. These programs can be downloaded separately for Mac, Windows and Unix operating systems or bundled in SNAP Workbench. Each program includes online documentation and a sample dataset.

Contact: ignazio_carbone@ncsu.edu

Supplementary information: A description of system requirements and installation instructions can be found at http://snap.cifr.ncsu.edu

INTRODUCTION

Recent advances in theoretical approaches for exploring population processes from DNA sequence variation within populations have resulted in a surge of new software tools (Beerli, 2006; Coop and Griffiths, 2004; De Iorio and Griffiths, 2004a, b; De Iorio *et al*., 2005; Hey and Nielsen, 2004; Lyngsø *et al*., 2005; Song, *et al*., 2005). These tools are often designed and validated using simulated data with unique input file formats and rarely make provisions for converting data into those formats. As a result, biologists with real data of varying complexity must create input files manually. Large multilocus datasets make this increasingly complex and necessitate the development of software tools to make multiple DNA sequence alignments accessible to new evolutionary methods. We have developed two such tools that we report here. Suite of Nucleotide Analysis Programs (SNAP) Combine and Map will help researchers to visualize the distribution of DNA sequence variation within populations, extract and merge information from one or multiple sequence alignments and enable further analysis by creating input files for several population genetic analysis programs.

SYSTEMS AND METHODS

Previously we developed a workbench program that can manage and coordinate a series of programs (Price and Carbone, 2005). SNAP Map will manipulate raw DNA sequence data from population samples into a variety of useful formats. This utility was conceived both to extract and characterize sequence variation in multiple sequence alignments and to serve as a bridge between existing applications requiring dissimilar input file formats. Variation includes both single nucleotide polymorphisms (SNPs) and insertions or deletions (indels). An indel is defined as one or more contiguous sites in a multiple sequence alignment that contain gaps in at least one sequence. Beyond identifying variable sites, SNAP Map provides the option to include or exclude indels, missing data or infinite-sites violations. The infinite-sites model is based on the assumption that few polymorphic sites will have >2 nt present (Hartl and Clark, 1997), and recent softwares such as Beagle (Lyngsø *et al*., 2005) and Shrub (Song *et al*., 2005) require that sites violating this assumption be eliminated from input data. In addition, we have included the ability to merge biogeographic or other phenotypic information with genetic sequence data to enable association-based analyses (Dean *et al*., 2005). Our goal was to maximize flexibility of the program so that it may be used to catalog sequence variation or simply convert multiple sequence alignments into specific file formats (Table 1).

A key feature of SNAP Map is the ability to collapse individual sequences into unique haplotypes, and to keep track of the count of each haplotype in the population sample. This is a necessary step for analyses that assume an infinite-sites model, and is a requirement for several of the software implementations we support (Table 1). A haplotype is a specific sequence of alleles or SNPs. Haplotypes are a useful way of grouping individuals according to genotype and are part of a powerful framework for testing significant associations with phenotype (Carbone *et al*., 2004; Dean *et al*., 2005; Phillips *et al*., 2002).

A novel extension of the collapse functionality is the option to collapse indels to unique integers. Indels are often removed from multiple sequence alignments because of the difficulty in modeling the mutation process at these sites. Our software provides the user with the option to extract indels and recode each unique indel with an one-digit integer. The appropriate integer is reinserted into each individual sequence, yielding alignments in

[*]To whom correspondence should be addressed.

Table 1. File formats currently generated using SNAP combine and map

File format	Includes phenotypic information	Reference
NEXUS	No	Maddison *et al*., 1997
CLUSTAL	No	Thompson *et al*., 1994
FASTA	No	Pearson and Lipman, 1988
PHYLIP	No	Felsenstein, 2004
MDIV[a]	Yes	Nielsen and Wakeley, 2001
GENETREE	Optional	Griffiths and Tavaré, 1994
RECOM58	No	Griffiths and Marjoram, 1996
RECMIN	No	Myers and Griffiths, 2003
RECPARS	No	Hein, 1993
HUDSON[b]	Yes	Hudson, 2000; Hudson *et al*., 1992
MIGRATE[c]	Yes	Beerli, 2006; Beerli and Felsenstein, 1999
SHRUB and HAPBOUND	No	Song *et al*., 2005
BEAGLE	No	Lyngsø *et al*., 2005

[a] IM (Hey and Nielsen, 2004) file format is also supported.
[b] Refers to the file format for the programs Seqtomatrix, Permtest, Permchi, and Snn developed by R. Hudson.
[c] Can also combine multiple single locus MIGRATE files into one multilocus file.

which gaps are recoded as a single polymorphic site. By recoding indels, we can take full advantage of variation at these sites in parsimony analyses and identify those sites that are compatible with an infinite-sites model. For example, the recoding of multilocus microsatellite and fingerprint data has important applications in phylogenetics and allows us to combine rapidly evolving markers with more slowly evolving base substitutions when reconstructing patterns of descent (Carbone *et al*., 1999; Dettman and Taylor, 2004).

SNAP Combine is designed to facilitate multilocus analyses. Since most existing software has no provisions for multilocus sequences, we developed a tool that could seamlessly merge sequence data for each individual/locus within the population. The merging operation performs a union of the input loci by default but intersection is also supported. The intersection is important to accommodate loci with missing sequence data for some taxa, thereby allowing researchers to begin data analysis while continuing the work to fill-in missing data. SNAP Combine merges multiple, potentially heterogeneously formatted, input files into an output file of specified format. Combine supports the following interleaved sequence formats for input and output: PHYLIP, NEXUS, FASTA and CLUSTAL. PHYLIP, NEXUS and FASTA are also supported as sequential sequence formats. SNAP Combine can be used to extract sequence subregions or taxon subsets and create new alignment files with specific combinations of loci. This is a powerful functionality and has been useful for examining patterns of sequence variation within and among loci (Charles *et al*., 2005).

The lack of standard file formats for population analysis software is a ubiquitous problem and the process of manually converting between different file formats can be tedious and problematic. Both SNAP Map and Combine will simplify this process. The input file for SNAP Map is a sequential PHYLIP-formatted sequence alignment, a standard output file option in sequence alignment programs, such as CLUSTAL W (Thompson *et al*., 1994), Sequencher Version 4.5 (Gene Codes Corporation, Ann Arbor, MI) and phylogeny inference packages, such as PHYLIP (Felsenstein, 2004) and PAUP (Swofford, 1998). To facilitate the conversion, SNAP Combine has the added functionality of converting CLUSTAL W and NEXUS-formatted alignment files into the sequential PHYLIP-formatted files for SNAP Map and vice versa. The conversion of combined PHYLIP files to CLUSTAL format is especially important when excluding strains from multiple alignments; this may result in suboptimal alignments with unnecessary alignment gaps that can easily be removed by realigning with CLUSTAL W.

IMPLEMENTATION

Our current implementation of SNAP Map generates more than 10 distinct output formats (Table 1). These particular formats were included because of necessity as they are the required input file formats for the various analysis tools we use frequently in our laboratory and center (Carbone *et al*., 2004; Dean *et al*., 2005). Each tool requires a specific input format that is not generated by other programs; this can be a source of frustration and error if generating these files manually. These are primarily applications developed for analyzing population structure and history within a non-parametric, coalescent or Bayesian framework but are also useful in multilocus macro-evolutionary analyses (Charles *et al*., 2005; Geml *et al*., 2006). Several of these tools can include geographical location or other phenotypic data in their analyses of population processes, and SNAP Map can merge specific individuals or haplotypes with corresponding phenotypic data. This ability allows us to take full advantage of both non-parametric and parameter-rich sequence-based models in testing for significant genotype–phenotype associations.

SNAP Map generates a summary table output that provides a visual overview of sequence variation in the population sample. This serves as a convenient reference and as a tool for exploring evolutionary processes at specific variable sites. The summary table numbers each site and gives the position of the variable site in the original sequence alignment. Each site is further labeled as a transition/transversion and informative/uninformative polymorphism. If the sequences have been collapsed to haplotypes, the summary table includes the frequency of each haplotype and provides a haplotype consensus sequence.

FRAMEWORK

SNAP Combine is written in Java and Map in ANSI C. They were developed on Apple's OS X operating system, but can be compiled on any platform. Both are part of the SNAP suite of software tools developed in the Carbone laboratory at North Carolina State University (http://snap.cifr.ncsu.edu). SNAP WorkBench provides an event-driven graphical user interface for integrating SNAP Map, Combine and other command-line tools. Several program calls to SNAP Map, each including different options, can be included in the SNAP Workbench menus; we recommend using Map with the Workbench for maximum ease of use. These and other SNAP tools can be downloaded from our website.

ACKNOWLEDGEMENTS

D.L.A was supported by the NSF IGERT and NIEHS Training Programs in Bioinformatics. E.W.P was supported from a North Carolina Biotechnology Center, Education Enhancement Grant. Additional support was from a North Carolina Cooperative State Research, Education, and Extension Service Grant to I.C.

Conflict of Interest: none declared.

REFERENCES

Beerli,P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–345.

Beerli,P. and Felsenstein,J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Carbone,I. *et al*. (1999) Patterns of descent in clonal lineages and their multilocus fingerprints are resolved with combined gene genealogies. *Evolution*, **53**, 11–21.

Carbone,I. *et al*. (2004) Recombination and migration of *Cryphonectria hypovirus 1* as inferred from gene genealogies and the coalescent. *Genetics*, **166**, 1611–1629.

Charles,L. *et al*. (2005) Phylogenetic analysis of *Pasteuria penetrans* by use of multiple genetic loci. *J Bacteriol*., **187**, 5700–5708.

Coop,G. and Griffiths,R.C. (2004) Ancestral inference on gene trees under selection. *Theor. Popul. Biol*., **66**, 219–232.

De Iorio,M. and Griffiths,R.C. (2004a) Importance sampling on coalescent histories. *I, Adv. Appl. Prob*., **36**, 417–433.

De Iorio,M. and Griffiths,R.C. (2004b) Importance sampling on coalescent histories. II: Subdivided population models. *Adv. Appl. Prob*., **36**, 434–454.

De Iorio,M. *et al*. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol*., **68**, 41–53.

Dean,R.A. *et al*. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.

Dettman,J.R. and Taylor,J.W. (2004) Mutation and evolution of microsatellite loci in *Neurospora*. *Genetics*, **168**, 1231–1248.

Felsenstein,J. (2004) *PHYLIP (Phylogeny Inference Package)*. Department of Genomic Sciences, University of Washington, Seattle.

Geml,J. *et al*. (2006) Beringian origins and cryptic speciation events in the fly agaric (*Amanita muscaria*). *Mol. Ecol*., **15**, 225–239.

Griffiths,R.C. and Marjoram,P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol*., **3**, 479–502.

Griffiths,R.C. and Tavaré,S. (1994) Ancestral inference in population genetics. *Statist. Sci*., **9**, 307–319.

Hartl,D. and Clark,A. (1997) *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, MA.

Hein,J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol*., **36**, 396–406.

Hey,J. and Nielsen,R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *Drosophila persimilis*. *Genetics*, **167**, 747–760.

Hudson,R.R. (2000) A new statistic for detecting genetic differentiation. *Genetics*, **155**, 2011–2014.

Hudson,R.R. *et al*. (1992) A statistical test for detecting geographic subdivision. *Mol. Biol. Evol*., **9**, 138–151.

Lyngsø,R.B., Song,Y.S. and Hein,J. (2005) Minimum recombination histories by branch and bound. In *Proceedings of the 5th International Workshop on Algorithms in Bioinformatics (Lecture Notes in Bioinformatics 3692)*, pp. 239–250.

Maddison,D.R. *et al*. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol*., **46**, 590–621.

Myers,S.R. and Griffiths,R.C. (2003) Bounds on the minimum number of recombination events in a sample history. *Genetics*, **163**, 375–394.

Nielsen,R. and Wakeley,J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Phillips,D.V. *et al*. (2002) Phylogeography and genotype–symptom associations in early and late season infections of canola by *Sclerotinia sclerotiorum*. *Phytopathology*, **92**, 785–793.

Price,E.W. and Carbone,I. (2005) SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics*, **21**, 402–404.

Song,Y.S. *et al*. (2005) Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. *Bioinformatics*, **21** (Suppl. 1), i413–i422.

Swofford,D.L. (1998) *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.0*. Sinauer Associates, Sunderland, MA.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.