Bioinformatics, 32(10), 2016, 1580–1582
doi: 10.1093/bioinformatics/btv763
Advance Access Publication Date: 18 January 2016
Applications Note

Genetics and population analysis

Sampletrees and Rsampletrees: sampling gene genealogies conditional on SNP genotype data

Kelly M. Burkett[1,2,*], Brad McNeney[1] and Jinko Graham[1]

[1]Department of Statistics, Simon Fraser University, Burnaby V5A 1S6, Canada and [2]Department of Mathematics and Statistics, University of Ottawa, Ottawa K1N 6N5, Canada

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on 28 July 2015; revised on 29 November 2015; accepted on 23 December 2015

Abstract

Summary: The program `sampletrees` is a Markov chain Monte Carlo sampler of gene genealogies conditional on either phased or unphased SNP genotype data. The companion program `Rsampletrees` is for pre- and post-processing of `sampletrees` files, including setting up the files for `sampletrees` and storing and plotting the output of a `sampletrees` run.

Availability and implementation: `sampletrees` is implemented in C++. The source code, documentation and test files are available at http://stat.sfu.ca/statgen/research/sampletrees.html. The R package `Rsampletrees` is available on CRAN http://cran.r-project.org/web/packages/`Rsampletrees`/index.html.

Contact: kburkett@uottawa.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The gene genealogy is a tree that describes the ancestral relationships among chromosomal segments sampled from individuals in a population. Knowledge of this ancestral tree has application in population genetic inference and for gene-mapping (Burkett *et al.*, 2013a; Zöllner and Pritchard, 2005). As the true tree cannot be known, the uncertainty can be handled with Markov chain Monte Carlo (MCMC) by sampling trees that are consistent with the observed data.

The C++ program `sampletrees` is used to sample gene genealogies conditional on genetic data observed at present. The algorithm is based on the approach outlined in (Zöllner and Pritchard, 2005). Letting τ represent the genealogical tree, including branch times at a genomic location called the focal point, and G represent the SNP data, we wish to sample from the posterior distribution $\mathrm{PR}(\tau|G)$. This is achieved using an MCMC approach, where we sample from a distribution proportional to $\mathrm{PR}(G, \tau)$. This joint distribution is modelled by including additional latent variables representing the haplotype sequences of internal nodes of the tree, and the rates of mutation and recombination on the branches of the tree. To sample trees from the posterior distribution, at a step of the sampler, a proposal distribution that updates some components of the latent data (i.e. the mutation rate, recombination rate, internal node sequences, tree topology, or tip haplotypes) is selected and an update to the latent data is proposed. This update is either accepted or rejected according to the Metropolis-Hastings ratio. For further statistical details on the algorithm, please see Burkett *et al.* (2013b) and references therein.

In addition to the C++ program, the R package `Rsampletrees` is available to assist users with creating the input files needed by `sampletrees` and manipulate and display the results after sampletrees has been run. We expect the two programs to be used as follows:

1. With `Rsampletrees`, generate the necessary input files for `sampletrees`.
2. Run the `sampletrees` executable at the command line, passing the name of the settings file generated in Step 1 to the program.
3. Read in and summarize the results of Step 2 in R, using `Rsampletrees`.

In the next section, we briefly describe the input and output files for `sampletrees` and how `Rsampletrees` can be used to generate and analyze these files.

2 Implementation

2.1 sampletrees

sampletrees samples gene genealogies conditional on user-supplied genetic data. The source code, detailed documentation, test files and instructions for compiling the program are available at http://stat.sfu.ca/statgen/research/sampletrees.html. Mac and Linux users need to install from source; Windows users can download a pre-compiled version.

The input files for sampletrees consist of

- **settings file:** The settings file is used to specify the input and output data file names, the type of data, information about the MCMC run (chain length, burn-in, thinning), the location of the focal point, initial values for the latent variables and parameters for the prior distributions.
- **weights file:** The weights file indicates which proposal distributions to sample from and their sampling probabilities.
- **SNP data file:** Both unphased genotype and phased haplotype data are allowed. If phased SNP data are available, the haplotype option is used and the haplotype file should consist of one haplotype per row, with allele codes as 0 or 1. If the genetic data is not phased, the genotype option can be used and the genotype file should consist of one multi locus genotype per row with genotypes coded as 0, 1 or 2 (number of copies of one of the alleles).
- **SNP location file:** The relative genomic locations of the SNPs.
- **haplotype frequency file:** With the genotype option, the user must provide an additional data file giving 2-locus haplotype frequency estimates between adjacent SNPs.
- **(optional) initial haplotype phase configurations file and haplotype list file:** To improve convergence with the genotype option, we recommend that the user also provide a file with an initial haplotype configuration for each individual in the dataset to start the sampling and a file with a list of haplotypes with high probability to exist in the population.

Rsampletrees can be used to generate and manipulate the settings file, the weights file and the haplotype frequency file (unphased genotype data). The optional initial haplotype phase configurations and optional haplotype list can be estimated using programs like PHASE (Stephens *et al.*, 2001) or with Rsampletrees functions.

Sampletrees is run by typing

```
>./sampletrees settingsfile
```

in a shell (linux and Mac) or

```
sampletrees.exe settingsfile
```

at the command prompt (Windows), where 'settingsfile' gives the name of the settings file. The program will print the run settings to the screen so that the user can verify that these were read in correctly. In addition, the iteration number of the sampling is also periodically printed out in order to monitor progress.

sampletrees generates multiple output files that give: (i) sampled mutation and recombination rates, (ii) sampled tree topology and branch lengths in Newick format, (iii) initial latent data, (iv) final latent data, which is useful for increasing the chain length and (v) a table of acceptance proportions for each of the updates.

MCMC sampling of genealogic trees is known to be computationally intensive and to therefore require long run times (Kuhner, 2009). For example, it takes approximately 8 h for one million iterations on a dataset of 100 individuals and 25 SNPs; increasing the dataset to 150 individuals or doubling to 50 SNPs increases the run time to almost 14 h (Burkett *et al.*, 2013b). We have found that run time of the algorithm increases linearly with both number of SNPs and number of individuals in the sample. As with all MCMC algorithms, convergence of the chain should be assessed by the user. Larger datasets and genotype data will require greater numbers of iterations and so the run time is increased by both the larger dataset size and the longer Markov chain. To estimate run time, we recommend timing a preliminary run with a short chain first, and then extrapolating.

2.2 Rsampletrees

The R package Rsampletrees was developed to help users create the input files for sampletrees and process the output of a sampletrees run. The functions are designed around two main tasks:

- Pre-sampletrees: create the settings file with the arguments for a sampletrees run, including setting up initial haplotype files if the data is of type genotype.
- Post-sampletrees: read in the output files, compute summary statistics on the trees, plot the results.

We have also developed functions to help re-start a sampletrees run to allow for increasing the number of MCMC iterations.

The pre-sampletrees functions take and/or return a list object that consists of the settings for a sampletrees run. The post-sampletrees functions take and/or return a list object consisting of three components:

1. runinfo: The settings used in the sampletrees run.
2. rawdata: Raw data from the run. This is a list consisting of the iteration number for the MCMC samples, the sampled mutation and recombination rates, and either the sampled trees or the name of the file containing the sampled trees (because the tree files are very large).
3. procdata: Processed data and tree statistics from the run. This is initially a list with one element: a matrix of acceptance proportions of each of the MCMC updates for the run. If the user computes summary statistics on the sampled trees, Rsampletrees functions can be used to apply these tree statistics to the trees and store these results in procdata. The summary statistics can also be written to file for later analysis.

Additional details about Rsampletrees functions can be found in the R help files and in the documentation provided at http://stat.sfu.ca/statgen/research/sampletrees.html.

Users are expected to create R functions that take a tree as input and return a summary. Rsampletrees uses the R package ape (Paradis *et al.*, 2004) and its classes 'phylo' or 'multiPhylo' (a list of trees of class 'phylo') to represent the trees. For more information about these classes, please see http://ape-package.ird.fr/misc/FormatTreeR_24Oct2012.pdf.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada. Portions of this work were undertaken while K.B. held a Canadian Institutes of Health Research Doctoral Research Award and a Michael Smith Foundation for Health Research (MSFHR) Senior Graduate Trainee Award, and while J.G. was a MSFHR Scholar.

Conflict of Interest: none declared.

References

Burkett,K.M. *et al*. (2013a) Gene genealogies for genetic association mapping, with application to Crohn's disease. *Front. Stat. Genet. Methodol*., **4**, 260.

Burkett,K.M. *et al*. (2013b) Markov chain Monte Carlo sampling of gene genealogies conditional on unphased SNP genotype data. *Stat. Appl. Genet. Mol. Biol*., **12**, 559–581.

Kuhner,M.K. (2009) Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol*., **24**, 86–93.

Paradis,E. *et al*. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet*., **68**, 978–989.

Zöllner,S. and Pritchard,J.K. (2005) Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, **169**, 1071–1092.