Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btx733
Advance Access Publication Date: 28 November 2017
Applications Note

OXFORD

Genetics and population analysis

ThetaMater: Bayesian estimation of population size parameter θ from genomic data

Richard H. Adams, Drew R. Schield, Daren C. Card, Andrew Corbin and Todd A. Castoe*

Department of Biology, The University of Texas at Arlington, Arlington, TX 76019, USA

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on May 9, 2017; revised on October 9, 2017; editorial decision on October 26, 2017; accepted on November 27, 2017

Abstract

Summary: We describe *ThetaMater*, an open source R package comprising a suite of functions for efficient and scalable Bayesian estimation of the population size parameter θ from genomic data.
Availability and implementation: *ThetaMater* is available at GitHub (https://github.com/radamsRHA/ThetaMater).
Contact: todd.castoe@uta.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The population size parameter $\theta = 4N_e\mu$ ($2N_e\mu$ for haploid organisms) reflects the mutation–drift balance occurring within a population with an effective size of N_e individuals and a mutation rate of μ per site per generation. As a measure of genetic diversity, θ represents the expected number of segregating sites observed between a pair of homologous sequences sampled from a given population (Wakeley, 2008). Given an estimate of mutation rate, information about θ can be leveraged to obtain an estimate of the effective population size N_e. θ is a therefore a fundamental parameter of population genetics and is useful for understanding the degree to which neutral processes shape patterns of genetic variation in nature. Quantifying genetic diversity is also important to conservation biology, and thus estimates of θ provide critical insight into the genetic health of endangered species for informed conservation practices (Crandall *et al.*, 1999).

Numerous methods and genetic models have been developed to estimate θ from genetic data (see Wang, 2005 for examples). As any estimate obtained from a single locus or a small set of loci entails substantial uncertainty, large genome-scale datasets offer opportunity to estimate θ with high accuracy and precision. However, few likelihood-based methods are currently scalable to such massive datasets ($>10^6$ loci, >10 kb/locus), are often restricted to using a single or small set of diploid genomes, are restricted to a specific type of sequence data (i.e. whole genomes versus reduced representation) or require users to make assumptions about generation time and mutation rates. For example, most implementations of the popular pairwise-sequential Markov coalescent model require whole-genome data and that users provide a mutation rate assumed to be identical across all loci (Li and Durbin, 2011), while other methods are restricted to using individual diploid genomes (Haubold *et al.*, 2010). There are many genealogy-based methods for estimating demographic parameters (Felsenstein, 1992; Kuhner *et al.*, 1995), but these are intractable for genomic datasets that include many individuals. Furthermore, no current methods provide a statistical framework for leveraging estimates of θ to filter potentially spurious loci from datasets (i.e. paralogs). Accordingly, there is major need for efficient and scalable likelihood-based methods for estimating θ from diverse genome-scale data.

2 Implementation

The R package *ThetaMater* was written in R and C++ and requires the R package MCMCpack (Martin *et al.*, 2011) to simulate posterior probability distributions of θ. At the core of *ThetaMater* is the infinite-sites likelihood function (Watterson, 1975), which describes the probability distribution of observing k segregating sites in a sample size of n sequences obtained from a locus of size l. The likelihood of a genomic dataset under a given value of θ is then computed as a product of the individual locus likelihoods, each with an associated number of segregating sites k, sample size n and length l (see Supplementary Material). We have further expanded this approach to

incorporate a discretized gamma model of among-locus rate variation to accommodate rate variation and to characterize the genomic distribution of rate variation by estimating the gamma shape parameter α (Yang, 1997). Importantly, our method provides a user-friendly framework for efficient estimation of θ and genome-wide among-locus rate variation that is scalable to diverse genome-scale datasets ($>10^6$ loci) with larger samples sizes (>10 genomes), while accounting for uncertainty within a likelihood-based framework. Our method collapses datasets into sets of unique patterns, such that under many conditions, there is almost no limit to the number of loci that can be used to estimate θ within minutes on a desktop computer. Unlike other methods restricted to a particular format, *ThetaMater* includes scripts for converting a variety of widely used alignment formats into usable input, including whole-genome sequences, reduced representation data (i.e. RADseq, sequence capture) and single or multi-locus Sanger sequenced datasets. Finally, *ThetaMater* includes a posterior predictive simulator (PPS) that allows users to leverage estimates of θ to identify loci with evidence of model violations, such as selection (Adams *et al.*, 2016) or paralogy.

ThetaMater includes three Bayesian Markov Chain Monte Carlo (MCMC) simulation models for estimating posterior distributions of θ: M1 (*ThetaMater.M1*) assumes no among-locus rate variation, M2 (*ThetaMater.M2*) estimates θ using a fixed α parameter and M3 (*ThetaMater.M3*) estimates the joint posterior distribution of θ and α. We implement a gamma prior distribution for both θ and α with user-specified shape and scale parameters, and users can specify the number of rate classes used to approximate the distribution. The PPS function (*ThetaMater.PPS*) is directly integrated with the results from the three Bayesian models.

3 Biological application

As a demonstration, we applied *ThetaMater* on a previously published RADseq dataset (2051 loci; Schield *et al.*, 2017). We conducted Bayesian estimation of θ using *ThetaMater.M1* for the empirical dataset before and after filtering loci with *ThetaMater.PPS* (Fig. 1A). We also simulated a large genomic dataset comprised of 10^6 loci (2 kb each), sampling 20 genomes from a population with $\theta = 0.002$ and among-locus rate variation $\alpha = 0.5$ (Fig. 1B). We specified the shape and scale parameters of the prior distribution at 10 and 0.0001 for the empirical example and set prior parameters to 20 and 0.0001 for θ and 5 and 0.01 for α in the simulated analysis. We ran the MCMC chain for a total of 10^6 generations and discarded 10% as burn in. PPSs were run using the unfiltered posterior distribution, simulating a single locus for all 10^4 generations present in the post-burn in MCMC samples using *ThetaMater.PPS*.

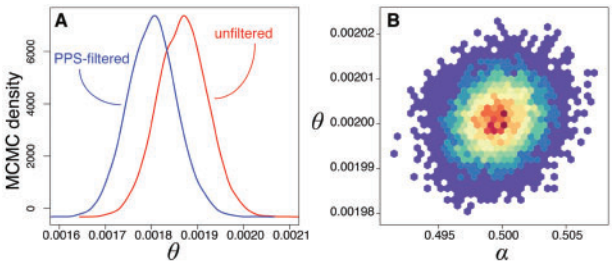

Fig. 1. (**A**) Empirical posterior estimates of θ before (red) and after (blue) filtering with *ThetaMater.PPS* and (**B**) The joint posterior distribution of θ and α for the simulated dataset showing highest densities (warm colors) at the true simulated values ($\theta = 0.002$, $\alpha = 0.5$) (Color version of this figure is available at *Bioinformatics* online.)

ThetaMater analysis of the unfiltered RADseq dataset suggested a mean θ estimate of 0.0019, corresponding to $N_e = 47\,500$ assuming a mutation rate of 10^{-8} (Fig. 1A, red). PPS based on this posterior distribution identified three loci with a significant excess of mutations, and these loci were filtered prior to reanalysis with *ThetaMater*. The posterior distribution of N_e inferred was centered around 45 000 individuals after removing these potentially spurious loci (Fig. 1A, blue). *ThetaMater* analysis of the simulated data returned the simulated parameter values with high probability (Fig. 1B).

ThetaMater is optimized for diverse datasets, including single diploid genome analyses, multi-genome data, reduced representation data and single or multi-locus alignments. *ThetaMater* assumes free recombination between loci, no recombination within loci, error-free SNP calls and neutral evolution. We encourage all users to carefully consider these assumptions prior to analysis with *ThetaMater* (see manual). Given the user-friendly framework and tractability of *ThetaMater*, we expect *ThetaMater* to be useful for a variety of applications, including population biology, comparative genomics and conservation biology.

Acknowledgement

We thank the Texas Advanced Computer Center for computational resources.

Funding

This work has been supported by the University of Texas at Arlington Phi Sigma Society Grant (to R.H.A.) and the National Science Foundation [DEB-1501747 to D.C.C. and T.A.C.; DEB-1501886 to D.R.S. and T.A.C.; DEB-1655571 to T.A.C].

Conflict of Interest: none declared.

References

Adams,R. *et al.* (2016) GppFst: genomic posterior predictive simulations of F_{ST} and d_{XY} for identifying outlier loci from population genomic data. *Bioinformatics*, **3**, 1414–1415.

Crandall,K. *et al.* (1999) Effective population sizes: missing measures and missing concepts. *Anim. Conserv.*, **2**, 317–319.

Felsenstein,J. (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.*, **59**, 139–147.

Haubold,B. *et al.* (2010) mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.*, **19**, 277–284.

Kuhner,M.K. *et al.* (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Martin,A.D. *et al.* (2011) MCMCpack: Markov Chain Monte Carlo in R. *J. Stat. Softw.*, **42**, 1–21.

Schield,D.R. *et al.* (2017) Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes. *Ecol. Evol.*, **7**, 3951–3966.

Wakeley,J. (2008) Coalescent Theory: An Introduction. Roberts and Company, Greenwood Village, CO.

Wang,J. (2005) Estimation of effective population sizes from data on genetic markers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **360**, 1395–1409.

Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.

Yang,Z.H. (1997) On the estimation of ancestral population sizes of modern humans. *Genet. Res.*, **69**, 111–116.