Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btw795
Advance Access Publication Date: 30 December 2016
Applications Note

OXFORD

Genetics and population analysis

GppFst: genomic posterior predictive simulations of F_{ST} and d_{XY} for identifying outlier loci from population genomic data

Richard H. Adams[1], Drew R Schield[1], Daren C. Card[1], Heath Blackmon[2] and Todd A. Castoe[1,*]

[1]Department of Biology, The University of Texas at Arlington, Arlington, TX 76019, USA and [2]Department of Ecology, Evolution & Behavior, University of Minnesota, Saint Paul, MN 55108, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on May 31, 2016; revised on December 7, 2016; editorial decision on December 11, 2016; accepted on December 13, 2016

Abstract

Summary: We introduce *GppFst*, an open source R package that generates posterior predictive distributions of F_{ST} and d_x under a neutral coalescent model to identify putative targets of selection from genomic data.
Availability and Implementation: *GppFst* is available at (https://github.com/radamsRHA/GppFst).
Contact: todd.castoe@uta.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic distributions of genetic differentiation provide a powerful framework for inferring evolutionary processes that have impacted regions of the genome. Two commonly used measures of genetic differentiation are F_{ST} (Wright, 1949), and d_{XY} (Takahata and Nei, 1985). These metrics have been applied extensively to characterize genome-wide patterns of genetic variation and differentiation across a wide range of populations and species (Jensen *et al.*, 2016).

In nature, most genomic variation is thought to derive from genetic drift occurring within structured populations. This expectation serves as a null model for identifying loci with patterns of genetic variation that differ significantly from the rest of the genome (so-called 'outlier' loci). Numerous studies have applied this principle to identify loci with extreme patterns of genetic differentiation that are poorly explained by neutral processes alone, and thus may indicate selection (Jensen *et al.*, 2016). Genetic differentiation can, however, be influenced by multiple factors; for example, small population sizes and deep divergence may shift neutral genomic distributions towards larger values of F_{ST} and d_{XY}, which can confound inferences of selection. Furthermore, most F_{ST}-based models assume equal rates of drift within the populations under study (Weir and Cockerham, 1984). Currently, no methods use an explicit probabilistic population model that incorporates demographic parameters to predict the distribution of neutral variation in F_{ST} and d_{XY}. For example, *pFst* employs a likelihood ratio test of allele frequency differences between populations (Shapiro *et al.*, 2013).

Here we describe a posterior predictive simulation (PPS) framework to generate theoretical distributions of F_{ST} and d_{XY} under the neutral coalescent model for two populations that accounts for demographic parameters in a probabilistic framework. Importantly, our method allows users to explicitly test the null hypothesis of genetic drift when conducting genomic scans. PPS is a popular method for evaluating model fit within a Bayesian framework that has been used to test a variety of evolutionary models (Gelman *et al.*, 2004; Reid *et al.*, 2014). Unlike other F_{ST} outlier tests, our PPS approach explicitly accounts for the demographic history of two genetically isolated species, including multiple demographic and experimental parameters (and uncertainty in those parameters), such as sample sizes, demographic parameters ($\theta = 4N_e\mu$), unequal rates of genetic drift within populations (unequal θs), and divergence time (τ). Additionally, other genomic F_{ST} outlier tests assume free recombination among SNPs. Our method allows users to simulate theoretical distributions that are conditioned on sampling multiple linked SNPs per locus—allowing users to take full advantage of large genomic datasets. We provide our PPS model in the package *GppFst*

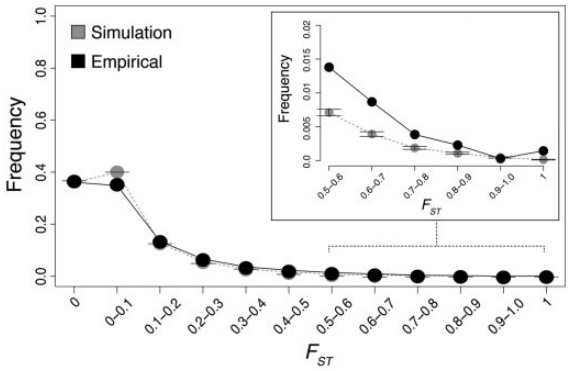

Fig. 1. Empirical and posterior predictive simulated (PPS) distributions of F_{ST} for example data, with standard deviations. The mean proportion of loci from the 100 replicate PPS runs (gray) and proportion of loci in the empirical data (black) are shown. Inset (top-right) shows upper limit of the F_{ST} distribution

(Genomic Posterior Predictive distributions of F_{ST}), which offers a user-friendly, open-source framework to generate theoretical distributions of F_{ST} and d_{XY} under the neutral coalescent model.

2 Implementation

The R package *GppFst* was written in R 3.2.2 and requires two other R packages, phybase (Liu and Yu, 2010) and Geneland (Guillot *et al.*, 2005) for simulating genealogies and computing Weir and Cockerham's F_{ST} (Weir and Cockerham, 1984). The functions *GppFst* and *GppDxy* require a posterior distribution of coalescent parameters ($\theta 0$, $\theta 1$, $\theta 01$, $\tau 01$) for a two-population model inferred via Markov Chain Monte Carlo (MCMC) sampling. This posterior distribution can be obtained using any program that implements a two-population coalescent model (see tutorial for examples). For each step in the MCMC, *GppFst* simulates coalescent genealogies and sequence alignments using a modified version of the function *simSeq from Sp* provided from phybase. F_{ST} and d_{XY} values are then computed for each simulated alignment, with the number of alignments to simulate per step specified by the user. Users can account for several experimental parameters, including variation in missing data per population and locus, locus lengths, and particular SNP-subsampling schemes. Both locus length and number of individuals per population are sampled from their empirical distributions, and users specify the number of SNPs to retain per simulated locus, which can be fixed at empirical values. After generating a theoretical distribution of F_{ST} or d_{XY}, users can compare empirical and simulated distributions to assign significance to outlier loci poorly explained by the neutral coalescent model.

3 Biological application

As a demonstration, we applied our *GppFst* model to a published RADseq SNP dataset (NCBI SRP051070) from two rattlesnake populations (Schield *et al.*, 2015). We inferred demographic parameters from 7031 unlinked nuclear SNPs with SNAPP (Bryant *et al.*, 2012). Using *GppFst*, we generated a PPS distribution of F_{ST} to identify loci that are poorly explained by neutral processes alone. Comparisons of the relative frequencies of simulated and empirical loci within F_{ST} intervals highlight extreme F_{ST} intervals that exhibit an excess of empirical loci when compared to the PPS distribution (Fig. 1). To calculate the empirical *P*-value, we use the PPS distribution to determine the probability of observing a given proportion of empirical loci within a specified F_{ST} interval. For example, the proportion of loci with $F_{ST}=1$ in the empirical distribution (0.0014) is more than ~10-fold greater than the proportion observed in the PPS distribution (0.00012). Thus, observing 10 loci with $F_{ST}=1$ is extremely unlikely under the neutral model ($P<0.0001$). Comparisons between our method and others that do not incorporate probabilistic model-based approaches suggest that *GppFst* provides more conservative estimates of outlier F_{ST} loci. For example, *GppFst* incorrectly identified a significant excess of SNPs with $F_{ST}=1$ in 4 of 100 simulated datasets (1000 neutral SNPs each), while the program *Arlequin* (Excoffier *et al.*, 2005) incorrectly assigned significance to every locus with an $F_{ST}=1$ in all 100 datasets (see tutorial). Our PPS framework employs the coalescent model of allopatric divergence between populations, which assumes free recombination between loci, no recombination within loci, and no gene flow. Because gene flow, recombination, and other factors may influence genomic variation, we recommend that users test all assumptions prior to using *GppFst*.

Acknowledgements

Texas Advanced Computer Center provided computational resources.

Funding

This work was supported by a Phi Sigma Grant to R.H.A, startup funds to T.A.C., and NSF DDIG grants to D.R.S & T.A.C (NSF DEB-1501886) and D.C.C. & T.A.C (NSF DEB-1501747).

Conflict of Interest: none declared.

References

Bryant,D. *et al.* (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29**, 1917–1932.

Excoffier,L. *et al.* (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online*, **1**, 47.

Gelman,A. *et al.* (2004) *Bayesian Data Analysis*. Vol. **2**. Chapman & Hall/CRC, Boca Raton, FL, USA.

Guillot,G. *et al.* (2005) GENELAND: a computer package for landscape genetics. *Mol. Ecol. Notes*, **5**, 712–715.

Jensen,J.D. *et al.* (2016) The past, present and future of genomic scans for selection. *Mol. Ecol.*, **25**, 1–4.

Liu,L. and Yu,L. (2010) Phybase: an R package for species tree analysis. *Bioinformatics*, **26**, 962–963.

Reid,N.M. *et al.* (2014) Poor fit to the multispecies coalescent is widely detectable in empirical data. *Syst. Biol.*, **63**, 322–333.

Schield,D.R. *et al.* (2015) Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*). *Mol. Phylogenet. Evol.*, **83**, 213–223.

Shapiro,M.D. *et al.* (2013) Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, **339**, 1063–1067.

Takahata,N. and Nei,M. (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325–344.

Weir,B.S. and Cockerham,C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wright,S. (1949) The genetical structure of populations. *Ann. Eugen*, **15**, 323–354.