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 \( d_{XY} \) 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 \( \theta_b \)), and divergence time (\( t \)). 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**.
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 ($0$, $0.01$, $0.001$, $0.001$) 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 and sequence alignments using a modified version of the function simSeq from Sp

3 Biological application

As a demonstration, we applied our GppFst model to a published RADseq SNP dataset (NCBI SRP051070) from two rattlesnake populations (Schild 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 $\sim$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


