Inferring the evolutionary history of populations

Renaud Vitalis

Centre de Biologie pour la Gestion des Populations
INRA, Montpellier
A long standing question

“One fairly obvious attack [to this problem] is to investigate [...] the expected consequences of drift by examining the variation of gene frequencies in time, or space. [...] The likelihood that [selection] simulate exactly [the amount of variation due to drift] will become smaller the more independent gene systems we examine, as the expectation of drift, unlike selective variation, will be the same for all genes”

Neutral vs. locally adapted genes

Characterizing the expected variation due to drift:

• $T_{LK} = (n - 1) \frac{F_{ST}}{\bar{F}_{ST}}$ (Lewontin and Krakauer 1973; Bonhomme et al. 2010)
Characterizing the expected variation due to drift:

- $T_{LK} = (n - 1) F_{ST}/\bar{F}_{ST}$ (Lewontin and Krakauer 1973; Bonhomme et al. 2010)
- Coalescent simulations (Beaumont and Nichols 1996; Vitalis et al. 2001; Excoffier et al. 2009)
Coalescent-based simulations

**FDIST** – Beaumont & Nichols 1996

Symmetrical population differentiation ($F_{ST}$), as a function of heterozygosity

**DETSel** – Vitalis et al. 2001

Joint distribution of $F_1$ and $F_2$ (which measure the divergence of populations 1 and 2 from their ancestor)

Ignoring hierarchical structure

Ignoring higher levels of structure increases the rate of false-positives...

Credits: Excoffier et al. (2009) Heredity 103: 285-298
Neutral vs. locally adapted genes

Characterizing the expected variation due to drift:

- $T_{LK} = (n - 1) F_{ST}/\bar{F}_{ST}$ (Lewontin and Krakauer 1973; Bonhomme et al. 2010)
- Coalescent simulations (Beaumont and Nichols 1996; Vitalis et al. 2001; Excoffier et al. 2009)
- Using empirical distributions (Akey et al. 2002; Weir et al. 2005)
Empirical distributions

Credits: Weir et al. (2005) Genome Research 15: 1468-1476
Characterizing the distribution of allele frequencies, conditionally on some model and parameters

• Island model: Beaumont and Balding (2004); Riebler et al. (2008); Foll and Gaggiotti (2008)
• Hierarchical island model: Gompert and Buerkle (2011); Foll et al. (2014)
• **Explicit modelling of selection**: SELESTIM (Vitalis et al. 2014)
The data

SNPs at many loci, in several populations (allele counts)
Population allele frequencies

Binomial likelihood that depends upon (unknown) population frequencies
\( \pi_j \): frequency at the jth locus in the total population (migrant pool)

\[ M_i = 4N_i m_i \]

Population 1
(frequency \( p_{1j} \))

... 

Population i
(frequency \( p_{ij} \))

... 

The population model

Infinite island model: the population frequencies depend on $M_i = 4N_i m_i$ and the frequencies in the migrant pool.
Relation to previous models

- Logistic regression model
- $F_{ST}$ is decomposed into a “locus-specific” effect and a “population-specific” effect

$$\log \left( \frac{F_{ST}^{ij}}{1 - F_{ST}^{ij}} \right) = \log \left( \frac{1}{\theta_{ij}} \right) = \alpha_i + \beta_j.$$  

- Beaumont and Balding (2004); Riebler et al. (2008); Foll and Gaggiotti (2008)
Allele frequencies in the migrant pool

Shape parameters of the beta distribution of (migrant) allele frequencies
• Allele frequencies = stationary density of the diffusion process (Wright 1949)
• All marker loci are targeted by selection, to some extent
• Sampling from the joint posterior distribution of the parameters using MCMC
Markov chain Monte Carlo (MCMC)

We use the Metropolis – Hastings algorithm to sample from the joint posterior distribution of the model parameters:

$$f(M, \pi, \kappa, \sigma, \delta, \lambda | n) \propto \prod_{i=1}^{n_d} \prod_{j=1}^{L} \mathcal{L}(p_{ij}; n_{ij}) \psi(p_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma_{ij}) \times$$

$$f(M) f(\pi) f(\kappa) f(\sigma | \delta) f(\delta | \lambda) f(\lambda)$$

Image courtesy of Peter Beerli, Florida State University, USA.

Decision criterion

- We compare the posterior distribution of $\delta_{ij}$ to a “centering distribution” that integrates over the overall departure from neutrality.

- We use the Kullback-Leibler divergence (KLD) as a distance between these distributions, calibrated using pseudo-observed datasets.
An example of application on simulated data ($F_{ST} = 0.10$): the distribution of the KLD for positively selected markers departs from that of neutral markers (and correlates with $F_{ST}$).
Comparison with BayeScan

- Positive selection
- Balancing selection
- SelEstim
- BayeScan

Comparison with BayeScan

- False positive rate
- True positive rate

Comparison with BayeScan

- False positive rate
- True positive rate

Comparison with BayeScan

- False positive rate
- True positive rate
Application on human data (CEPH)

- Strong signature of selection in the vicinity of the lactase gene *LCT*
- Strongest KLD at at 2 SNPs reported to be tightly associated with lactase persistence (13910T and 22018A; see Bersaglieri *et al.* 2004)
Application on human data (CEPH)

- Population-specific selection coefficient at 13910T (left) correlates with lactase persistence phenotype, particularly in Europe and the Indus valley.
A software package

SelEstim
Detecting and measuring selection from gene frequency data

Overview

The software package SelEstim is aimed at distinguishing neutral from selected polymorphisms and estimate the intensity of selection at the latter. The SelEstim model accounts explicitly for positive selection, and it is assumed that all marker loci in the dataset are responding to selection, to some extent. SelEstim is written in C. The source code as well as executables for various platforms (currently OS X, Windows, Linux) are available. The C executable reads a data file supplied by the user, and a number of options can be passed through the command line. The manual provides information about how to format the data file, how to specify the user-defined parameters, and how to interpret the results.

Citation


A command-line, parallelized (OpenMP), interface:

http://www1.montpellier.inra.fr/CBGP/software/selestim/index.html
How to use the information brought by haplotype structure?
$F_{ST}$-based tests using haplotype data

Bonhomme et al. (2010): a generalization of the Lewontin-Krakauer test that accounts for a tree-like history

Fariello et al. (2013): application on haplotypes obtained by local clustering (fastPHASE)

Credits: Fariello et al. (2013) *Genetics* **193**:929-941
$F_{ST}$-based tests using haplotype data

Credits: Fariello et al. (2013) *Genetics* 193:929-941
SELESTIM with haplotypes

SNPs

10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000

haplotypes (multiallelic markers)

10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000

10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000
SELESTIM with haplotypes

• Assuming a single haplotype is selected for

\[ \lambda \]

\[ \delta_j \]

\[ \kappa_{ij} \]

\[ \sigma_{ij} \]

\[ \pi_j \]

\[ M_i \]

\[ p_{ij} \]

\[ n_{ij} \]

Categorical prior distribution

Dirichlet prior distribution

Multinomial likelihood that depends upon (unknown) allele frequencies
Performance in the island model

- Improved statistical power with haplotype-based analyses (vs. SNPs)
Performance in the island model

• Improved statistical power with haplotype-based analyses (vs. SNPs)
• Better performance than FLK (Bonhomme et al. 2010) and hapFLK (Fariello et al. 2013)
Performance in divergence models

- Improved statistical power with haplotype-based analyses (vs. SNPs)
- Poorer performance than FLK (Bonhomme *et al*. 2010) and hapFLK (Fariello *et al*. 2013)
A software package

KimTree
Inferring population histories using genome-wide allele frequency data

Overview

The software package KimTree implements a hierarchical Bayesian model to estimate divergence times (in a diffusion time scale) in a population tree, from large single nucleotide polymorphism (SNP) data. The joint analysis of autosomal and X-linked polymorphisms further allows KimTree to infer the effective sex ratios or ESR (defined as the female proportion of the effective population), along each branch. The manual provides information about how to format the data file, how to specify the user-defined parameters, and how to interpret the results.

Citations


A command-line, parallelized (OpenMP), interface:
http://www1.montpellier.inra.fr/CBGP/software/kimtree/index.html
A digression on Pool-seq...

Credits: Schlötterer et al. (2014) Nature Reviews Genetics 15: 749-763

Valentin Hivert’s PhD (2015 – 2018)
Straightforward!

Binomial likelihood that depends upon (unknown) allele counts and coverage reads from pooled samples
More tricky: $F_{ST}$ from pooled data...

- Naive approaches may fail...
  - ✓ Considering reads as allele counts
  - ✓ Imputing allele counts using a maximum likelihood argument
A new estimator of $F_{ST}$ for pooled data

- Method-of-moments estimator, based on an analysis-of-variance framework
  - No bias
  - Performs better than any other estimator available (PoPoolation2, etc.)
A new estimator of $F_{ST}$ for pooled data

Measuring Genetic Differentiation from Pool-seq Data

Valentin Hivert,*† Raphaël Leblois,*† Eric J. Petit,* Mathieu Gautier,*,†† and Renaud Vitalis*,†‡
*CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, 34988 Montferrier-sur-Lez Cedex, France; †Institut de Biologie Computationnelle, Univ Montpellier, 34095 Montpellier Cedex, France; and ‡ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, 35042 Rennes, Cedex, France

poolfstat: Computing F-Statistics from Pool-seq Data

Functions for the computation of F-statistics from Pool-Seq data in population genomics studies. The package also includes several utilities to manipulate Pool-Seq data stored in standard format (vcf and recf files as obtained from the popular software ‘VarScan’ and ‘PoPoolation’ respectively) and perform conversion to alternative format (as used in the ‘BayPass’ and ‘SelEstim’ software).

Take home messages

• All these methods are designed to identify overly differentiated marker loci: local adaptation or intrinsic genetic incompatibilities?

• Be aware of the underlying population models and assumptions (e.g., island model vs. divergence models) and the robustness of the methods to model misspecifications

• Pool-seq experiments: random sampling of reads from allele counts must be (properly) accounted for!

Acknowledgements and credits

• Mark A. Beaumont, Florian Clemente (former postdoc), Kevin J. Dawson, Mathieu Gautier, Valentin Hivert (former PhD student)