BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 24 2011, pages 3435–3436
doi:10.1093/bioinformatics/btr589

Genetics and population analysis Advance Access publication October 23, 2011

PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)

Robert Kofler, Ram Vinay Pandey and Christian Schlötterer*
Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Sequencing pooled DNA samples (Pool-Seq) is the most cost-effective approach for the genome-wide comparison of population samples. Here, we introduce PoPoolation2, the first software tool specifically designed for the comparison of populations with Pool-Seq data. PoPoolation2 implements a range of commonly used measures of differentiation (F_{ST}, Fisher's exact test and Cochran-Mantel-Haenszel test) that can be applied on different scales (windows, genes, exons, SNPs). The result may be visualized with the widely used Integrated Genomics Viewer.

Availability and Implementation: PoPoolation2 is implemented in Perl and R. It is freely available on http://code.google.com/p/popoolation2/

Contact: christian.schloetterer@vetmeduni.ac.at

Supplementary Information: Manual: http://code.google.com/p/popoolation2/wiki/Manual

Test data and tutorial: http://code.google.com/p/popoolation2/wiki/Tutorial

Validation: http://code.google.com/p/popoolation2/wiki/Validation

Received on August 19, 2011; revised on September 28, 2011; accepted on October 18, 2011

1 INTRODUCTION

Next-generation sequencing of pooled DNA samples (Pool-Seq) allows the comparison of population samples on a genomic scale, thus facilitating the transition from single marker studies to population genomics. Due to its cost-effectiveness (Futschik and Schlötterer, 2010), Pool-Seq can be used for a range of applications. The most intuitive application is the comparison of natural populations to perform standard population genetic analyses on a genomic scale (e.g. Begun *et al.*, 2007). The comparison of natural *Arabidopsis lyrata* populations from different habitats allowed the characterization of genes involved in heavy metal tolerance (Turner *et al.*, 2010). Also in experimental evolution studies, Pool-Seq has been used to identify genomic regions that show high differentiation between different selective treatments (Burke *et al.*, 2010; Parts *et al.*, 2011; Turner *et al.*, 2011). Finally, Pool-Seq offers an enormous potential for selective genotyping (Darvasi and Soller, 1994; Hillel *et al.*, 1990; Lander and Botstein, 1989).

While several tools for analyzing Pool-Seq data of single populations are already available (Bansal, 2010; Kofler *et al.*, 2011; Pandey *et al.*, 2011), to our knowledge no standalone software tool is available for the comparison of Pool-Seq data for multiple populations. PoPoolation2 is a software tool dedicated to the comparison of allele frequencies between populations.

2 IMPLEMENTATION

As input PoPoolation2 requires a 'pileup' file for every population (sample) of interest or alternatively a single multi 'pileup' file (mpileup) may be used. These files can be obtained by mapping the reads of a Pool-Seq experiment to a reference genome and subsequently converting the mapping results into the 'pileup/mpileup' format with samtools (Li *et al.*, 2009) (For Manual see http://code.google.com/p/popoolation2/wiki/Manual; Test data and tutorial http://code.google.com/p/popoolation2/wiki/Tutorial). PoPoolation2 requires Pool-Seq data from at least two populations, but may be used with an unlimited number of populations.

To assess allele frequency differences between population samples PoPoolation2 implements a wide variety of statistics.

- As the most intuitive measure of population differentiation, the allele frequency differences are reported.
- The fixation index (F_{ST}) can be calculated to measure differentiation between populations. F_{ST} values may either be calculated with the classical approach (Hartl and Clark, 2007) or with an approach adapted to digital data (Karlsson *et al.*, 2007)
- The statistical significance of allele frequency differences is determined with Fisher's exact test (Fisher, 1922).
- Since in experimental evolution experiments and selective genotyping studies often biological replicates are available, we implemented the Cochran–Mantel–Haenszel (CMH) test (Landis *et al.*, 1978) to test for the statistical significance between groups.

When data from more than two populations are available, PoPoolation2 automatically computes all pairwise comparisons for these tests (except for the CMH test).

All these analyses can be performed on different levels. We have implemented a sliding window analysis, which permits a genome-wide scan for differentiation using a specified window size. For the analysis of single SNPs, a window size of 1 may be used. Finally, with a user-provided GTF file the analysis of genes, coding sequence, introns, etc. is possible. To visualize the population differentiation across the genome, PoPoolation2 converts the results into file formats that are compatible with the Integrative Genomics Viewer (Robinson *et al.*, 2011).

*To whom correspondence should be addressed.

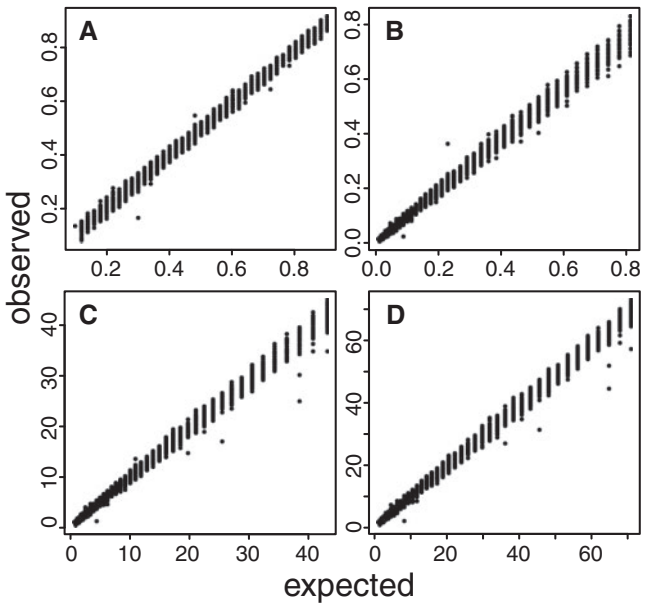

Fig. 1. Expected versus observed values for the tests implemented in PoPoolation2 using 10 000 simulated SNPs. (**A**) allele frequency difference; (**B**) F_{ST}; (**C**) Fisher's exact test $[-\log 10(P\text{-value})]$; (**D**) CMH test $[-\log 10(P\text{-value})]$.

Finally, PoPoolation2 also implements the functionality to randomly subsample the data to achieve a uniform coverage. The subsampling is based on a user-defined quality threshold. For analyzing the data with standard software, such as Mega5 (Tamura *et al.*, 2011) and Arlequin (Excoffier and Lischer, 2010), PoPoolation2 allows exporting the data as artificial chromosomes as 'multi-fasta' files and as 'GenePop' files (Raymond and Rousset, 1995).

3 VALIDATION

To test PoPoolation2, we placed 10 000 SNPs for two populations on chromosome 2R of *Drosophila melanogaster* (v5.38). For these SNPs, we simulated 75 bp reads such that the coverage was 100× and the allele frequency differences between the two populations ranged from 0.1 to 0.9. Subsequently, the simulated reads were mapped to the reference genome (*D.melanogaster*, chromosome 2R, v5.38) with BWA (0.5.8) (Li and Durbin, 2009) and a 'mpileup' file was created using samtools (0.1.13) (Li *et al.*, 2009). Finally, we compared the expected values with the observed ones and found an almost perfect correlation between the simulated data and the estimates based on PoPoolation2 for all implemented tests (allele frequency differences: $R^2 = 0.9979$, $P < 2.2e\text{-}16$; F_{ST}: $R^2 = 0.9967$, $P < 2.2e\text{-}16$; Fisher's exact test: $R^2 = 0.9974$, $P < 2.2e\text{-}16$; CMH test: $R^2 = 0.9978$, $P < 2.2e\text{-}16$; Fig. 1). These high correlations confirm that PoPoolation2 yields highly reliable results (for details, see http://code.google.com/p/popoolation2/wiki/Validation).

To ensure that all scripts continue to work properly, we implemented Unit-tests for the main scripts (which may be run by providing the parameter '–test').

ACKNOWLEDGEMENTS

We are grateful to V. Nolte, M. Kapun and P. Orozco-ter Wengel for helpful comments and discussions. We thank all members of the 'Institut für Populationsgenetik' for early testing and feedback.

Funding: Austrian Science Fund (FWF): P19467-B11, P22725-B11.

Conflict of Interest: none declared.

REFERENCES

Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Begun,D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.*, **5**, e310.

Burke,M.K. *et al.* (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–590.

Darvasi,A. and Soller,M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*, **138**, 1365–1373.

Excoffier,L. and Lischer,H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, **10**, 564–567.

Fisher,R.A. (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87–94.

Futschik,A. and Schlötterer,C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Hartl,D.L. and Clark,A.G. (2007) *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts.

Hillel,J. *et al.* (1990) DNA fingerprints from blood mixes in chickens and in turkeys. *Animal Biotechnol.*, **1**, 201–204.

Karlsson,E.K. *et al.* (2007) Efficient mapping of Mendelian traits in dogs through genome-wide association. *Nat. Genet.*, **39**, 1321–1328.

Kofler,R. *et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, **6**, e15925.

Lander,E.S. and Botstein,D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

Landis,J.R. *et al.* (1978) Average partial association in 3-way contingency-tables - review and discussion of alternative tests. *Int. Stat. Rev.*, **46**, 237–254.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Pandey,R.V. *et al.* (2011) PoPoolation DB: a user-friendly web-based database for the retrieval of natural polymorphisms in *Drosophila*. *BMC Genet.*, **12**, 27.

Parts,L. *et al.* (2011) Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res.*, **21**, 1131–1138.

Raymond,M. and Rousset,F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Heredity*, **86**, 248–249.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Tamura,K. *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.

Turner,T.L. *et al.* (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat. Genet.*, **42**, 260–263.

Turner,T.L. *et al.* (2011) Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.*, **7**, e1001336.