Bioinformatics Advance Access published July 13, 2016

Bioinformatics, 2016, 1–2
doi: 10.1093/bioinformatics/btw352
Advance Access Publication Date: 16 June 2016
Applications Note

Phylogenetics

RADIS: analysis of *RAD*-seq data for interspecific phylogeny

Astrid Cruaud[1,*,†], Mathieu Gautier[1,2,†], Jean-Pierre Rossi[1], Jean-Yves Rasplus[1] and Jérôme Gouzy[3,4]

[1]INRA, UMR1062 CBGP, F-34988 Montferrier-sur-Lez, France, [2]IBC, F-34095 Montpellier, France, [3]INRA, UMR441 LIPM, F31326 Castanet Tolosan, France and [4]CNRS, UMR2594 LIPM, F31326 Castanet Tolosan, France

*To whom correspondence should be addressed.
†These authors contributed equally to this work.

Associate Editor: Alfonso Valencia

Received on February 3, 2016; revised on April 26, 2016; accepted on May 30, 2016

Abstract

In an attempt to make the processing of RAD-seq data easier and allow rapid and automated exploration of parameters/data for phylogenetic inference, we introduce the perl pipeline *RADIS*. Users of *RADIS* can let their raw Illumina data be processed up to phylogenetic tree inference, or stop (and restart) the process at some point. Different values for key parameters can be explored in a single analysis (e.g. loci building, sample/loci selection), making possible a thorough exploration of data. *RADIS* relies on *Stacks* for demultiplexing of data, removing PCR duplicates and building individual and catalog loci. Scripts have been specifically written for trimming of reads and loci/sample selection. Finally, *RAxML* is used for phylogenetic inferences, though other software may be utilized.

Availability and implementation: RADIS is written in perl, designed to run on Linux and Unix platforms. RADIS and its manual are freely available from http://www1.montpellier.inra.fr/CBGP/software/RADIS/.

Contact: astrid.cruaud@supagro.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Restriction site Associated DNA sequencing (RAD-seq, Baird *et al.*, 2008; Miller *et al.*, 2007) is a promising tool to confidently resolve phylogenetic relationships among eukaryote species and genera (e.g. Cruaud *et al.*, 2014; Eaton and Ree, 2013; Hipp *et al.*, 2014; McCluskey and Postlethwait, 2015). However, analyzing RAD-seq data to infer phylogenies remains challenging as it requires many steps and decisions to process raw data into a format ready for analysis. Some steps can be achieved using a collection of well-packaged software, but others require bioinformatic skills. An examination of data is required to better analyze their quality and impact on topology/branch lengths. Assessment of the robustness of the resulting trees to the parameters chosen for loci building and loci/sample selection is required (Leaché *et al.*, 2015) but represents a tedious and error-prone task when done manually. In an attempt to standardize processing of RAD-seq data for phylogenetic inference, allow fast and automated exploration of key options, and facilitate comparison among clustering methods to form sets of loci (e.g. Stacks Catchen *et al.*, 2011; Catchen *et al.*, 2013 *versus PyRAD*, Eaton 2014), we designed the user-friendly perl pipeline *RADIS*.

2 Description

Processing of raw data has been split up into two steps: data cleaning and data analysis (Supplementary Fig. S1). Example datasets are provided with the package (processing time <1 min on 8-cores of a 16-core Linux, 2.9 GHz, 64 GB RAM computer). Users can choose to (i) process their raw Illumina data up to phylogenetic tree inference (*RADIS.pl*); (ii) perform only data cleaning (*RADIS_step1_data_cleaning.pl*) or (iii) perform only data analysis (*RADIS_step2_data_analysis.pl*).

RADIS operates on the basis of a configuration file (RADIS.cfg) in which users provide parameters values to be tested and paths to external software. The software comes with a fully annotated configuration file to facilitate the initialization. A correspondence between barcodes and sample codes can be provided to allow file renaming (barcodes_lib_names.txt). Progress of the analysis can be followed (stdout/stderr files). Output files and necessary subdirectories are automatically created in a directory specified by the user.

Data cleaning—Reads that do not pass Illumina's filtering are discarded. *RADIS* relies on *process_radtags* from the software pipeline *Stacks* to demultiplex data. Users can choose to remove nucleotides from the 5′ and 3′ ends of forward and reverse reads (e.g. to remove enzyme cut sites or bad quality nucleotides). If barcodes of different sizes are used, reads are automatically trimmed to the same length. To remove PCR duplicates, *RADIS* then uses *clone_filter* (*Stacks*). Finally, sequence files are renamed after the sample codes. At each step of the process, files are created that provide summary statistics on the number of reads removed/kept

Data analysis—'Purified read 1' outputs from the first step are processed individually and a set of loci is produced for each sample using *ustacks* (*Stacks*). Users can provide a list of values to be tested for M, the maximum number of nucleotides that may be different between stacks (assembly of exactly matching reads) to be merged into a single locus (http://creskolab.uoregon.edu/stacks/param_tut.php). Individual loci are then merged into a catalog of loci with *cstacks* (*Stacks*). Users can provide a list of values for the parameter *n*, the number of mismatches allowed between individual loci when generating the catalog. Exploring alternate parameterization allows the user to find a good compromise for merging orthologous loci from distant species, whilst ensuring that paralogs and non-homologous loci are not merged. Users can then perform loci and sample selection to build datasets that fit with their prior knowledge of the studied species. They can fix a minimum number of loci required for a sample to be kept in the analysis, or retain only loci for which at least a given number of samples have sequences (a list of values can be provided). Users can also choose to remove loci in which paralogs and non-homologous sequences are probably merged together. Phylip-formatted files that meet the selection criteria are produced by *RADIS*. Finally, combined datasets (concatenation of the full sequence of each locus) are analyzed using *RAxML* (Stamatakis, 2006a,b) to produce phylogenetic trees. Users can delay implementation of *RAxML* analyses in order to increase the number of cpus to be used. Explicit names are used for output directories and files making the results obtained with different sets of parameters easily distinguished and compared (e.g. stacks_M2n4S12L10000.sel.phy is the phylip-formatted combined dataset obtained when individual loci are built using $M = 2$ (M2, *ustacks*), the catalog of loci is built using $n = 4$ (n4, *cstacks*), only sample with at least 10 000 loci (L10000) and loci for which at least 12 samples have sequences (S12) are selected (.sel). It is noteworthy that *RADIS* can process data from as many RAD libraries as needed.

3 Evaluation using empirical data

To test the program, we reanalyzed the raw data from Cruaud *et al.* (2014). Experimental design was as follows: DNA from 31 samples was first digested with *PstI* and P1 adaptors containing 5 or 6 bp barcodes were then ligated. Paired-end sequencing of the library (2 * 100 nt) was performed on a single lane of a HiSeq 2000. Raw Illumina data were processed with *RADIS.pl* using *Stacks-1.32* and *RAxML 8.2.0-ALPHA* on 8-cores of a 16-core Linux, 2.9 GHz, 64 GB RAM computer. In the R ADIS.cfg file, radis_nttrim_read1_5p was set to 5 (to remove the overhang of the restriction site), radis_nttrim_read2_3p was set to 5 (to remove low quality bases) while radis_nttrim_read1_3p and radis_nttrim_read2_5p were set to 0. *M* was set to 2 and 4 values of *n* were tested (4, 6, 8, 10). We only retained samples with more than 10 000 loci, and loci for which at least 12 or all samples were represented. Phylogenetic analyses were performed without partitioning. A $\mathrm{GTR} + \Gamma$ model with a discrete gamma approximation (4 categories, Yang, 1994) was used for the ML analysis and a GTRCAT approximation of models was used for bootstrapping (1000 replicates). Results (identical to published ones) were obtained within 4 h (Supplementary Fig. S2).

4 Conclusion

By facilitating testing the impact of different parameter combinations, the pipeline *RADIS* automates and standardizes the analyses of RAD-seq data for phylogenetic inference. The program may prove useful to evaluate the robustness of the results to the options chosen to process real RAD-seq data, or to carry out simulation studies. Most importantly, *RADIS* could also help to assess how different clustering methods may impact tree topology (e.g. Stacks versus *PyRAD*).

Acknowledgements

We thank C. Kerdelhué and A. Dehne Garcia for discussions, S. McCairns for proofreading and three anonymous reviewers for their comments.

Funding

This work was funded by the division SPE of the INRA, the ANR project '*TriPTIC*' (ANR-14-CE18-0002).

Conflict of Interest: none declared.

References

Baird,N.A. *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Catchen,J. *et al.* (2013) Stacks: an analysis tool set for population genomics. *Mol. Ecol.*, **22**, 3124–3140.

Catchen,J.M. *et al.* (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3-Genes Genomes Genet.*, **1**, 171–182.

Cruaud,A. *et al.* (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.*, **31**, 1272–1274.

Eaton,D.A.R. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, btu121.

Eaton,D.A.R. and Ree,R.H. (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Syst. Biol.*, **62**, 689–706.

Hipp,A.L. *et al.* (2014) A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS ONE*, **9**, e93975.

Leaché,A.D. *et al.* (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.*, **7**, 706–719.

McCluskey,B.M. and Postlethwait,J.H. (2015) Phylogeny of zebrafish, a "model species," within *Danio*, a "model genus". *Mol. Biol. Evol.*, **32**, 635–652.

Miller,M.R. *et al.* (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.*, **17**, 240–248.

Stamatakis,A. (2006a) Phylogenetic models of rate heterogeneity: A High Performance Computing Perspective. International Parallel and Distributed Processing Symposium (IPDPS 2006), Rhodes Island, Greece, pp. 8.

Stamatakis,A. (2006b) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.