PIrANHA

PhylogenetIcs ANd pHylogeograPhy

Scripts for file processing and analysis in phylogenomics & phylogeography

LICENSE

All code within the PIrANHA v0.1.4 repository is available "AS IS" under a generous GNU license. See the LICENSE file for more information.

CITATION

If you use scripts from this repository as part of your published research, I require that you cite the repository as follows (also see DOI information below):

- Bagley, J.C. 2017. PIrANHA. GitHub repository, Available at: http://github.com/justincbagley/PIrANHA.

Alternatively, please provide the following link to this software repository in your manuscript:

- https://github.com/justincbagley/PIrANHA

DOI

The DOI for PIrANHA, via Zenodo, is as follows: DOI 10.5281/zenodo.166309. Here are some examples of citing PIrANHA using the DOI:

Bagley, J.C. 2017. PIrANHA. GitHub package, Available at: http://doi.org/10.5281/zenodo.166309.

Bagley, J.C. 2017. PIrANHA. Zenodo, Available at: http://doi.org/10.5281/zenodo.166309.

INTRODUCTION

Taking steps towards automating boring stuff during analyses of genetic data in phylogenomics & phylogeography...

PIrANHA v0.1.4 is a repository of shell scripts and R scripts written by the author, as well as additional code (R, Perl, and Python scripts) from other authors, that is designed to help automate processing and analysis of DNA sequence data in phylogenetics and phylogeography research projects (Avise 2000; Felsensetin 2004). PIrANHA is fully command line-based and, rather than being structured as a single pipeline, it contains a series of scripts, some of which form pipelines, for aiding or completing tasks during evolutionary analyses of genetic data. Currently, PIrANHA scripts facilitate running or linking the following software programs:

- **pyRAD** or **ipyrad** (Eaton 2014)
- **PartitionFinder** (Lanfear et al. 2012)
- **BEAST** (Drummond et al. 2012; Bouckaert et al. 2014)
- **starBEAST** (Heled & Drummond 2010)
- **ExaBayes** (Aberer et al. 2014)
- **dadi** (Gutenkunst et al. 2009)
- **fastSTRUCTURE** (Raj et al. 2014)
- **PhyloMapper** (Lemmon and Lemmon 2008)

The current code in PIrANHA has been written largely with a focus on 1) analyses of DNA sequence data and SNPs or SNP loci generated from massively parallel sequencing runs on ddRAD-seq genomic libraries (e.g. Peterson et al. 2012), and 2) automating running these software programs on the user's personal machine (e.g. MAGNET pipeline and pyRAD2PartitionFinder scripts) or a remote supercomputer machine, and then conducting post-processing of the results. In particular, a number of scripts have been written with sections allowing them to be run (or cause other software to be called) on a supercomputing cluster, using code suitable for SLURM or TORQUE/PBS resource management systems (in some cases, this functionality is noted by adding "Super" in the script filename, as in Super-pyRAD2PartitionFinder.sh).

Distribution Structure and Pipelines

What's new in this release?

The current build, v0.1.4, has not yet been released, but contains several goodies listed below, in addition to minor improvements in the code!! - **May 2017:** build now contains new 'BEASTRunner.sh' script and 'beast_runner.cfg' configuration file. BEASTRunner now has options to allow specifying 1) number of runs, 2) walltime, and 3) Java memory allocation per run, as well as calling reg or verbose help documentation from the command line. - **April 2017:** build now contains new 'pyRADLocusVarSites.sh' script (with example run folder) that calculates numbers of variable sites (i.e. segregating sites, S) and parsimony-informative sites (PIS; i.e. hence with utility for

phylogenetic analysis) in each SNP locus contained in .loci file from a pyRAD assembly run. - **April 2017:** I added new 'dadiRunner.sh' script that automates transferring and queuing multiple runs of dadi input files on a remote supercomputer (similar to BEASTRunner and RAxMLRunner scripts already in the repo). - **January 2017:** I added a new script called 'BEAST_PSPrepper.sh' that, while not quite polished, automates editing any existing BEAST v2+ (e.g. v2.4.4) input XML files for path sampling analysis, so that users don't have to do this by hand!

What is possible with PIrANHA? Who cares?

How PIrANHA scripts work together

PIrANHA facilitates analysis pipelines that could be of interest to nearly anyone conducting evolutionary analyses of DNA sequence data using maximum-likelihood and Bayesian methods. **Figure 1** and **Figure 2** below demonstrate flow and interactions of the current partition scheme, population structure, and phylogenetics pipelines with **software** and **"file types"** used to generate input for PIrANHA in the left column, and the way these are processed within/using PIrANHA illustrated in the right column. External software programs are shown in balloons with names in black italic font, while PIrANHA scripts are given in blue. Arrows show the flow of files through different pipelines, which terminate in results (shown right of final arrows at far right of each diagram).

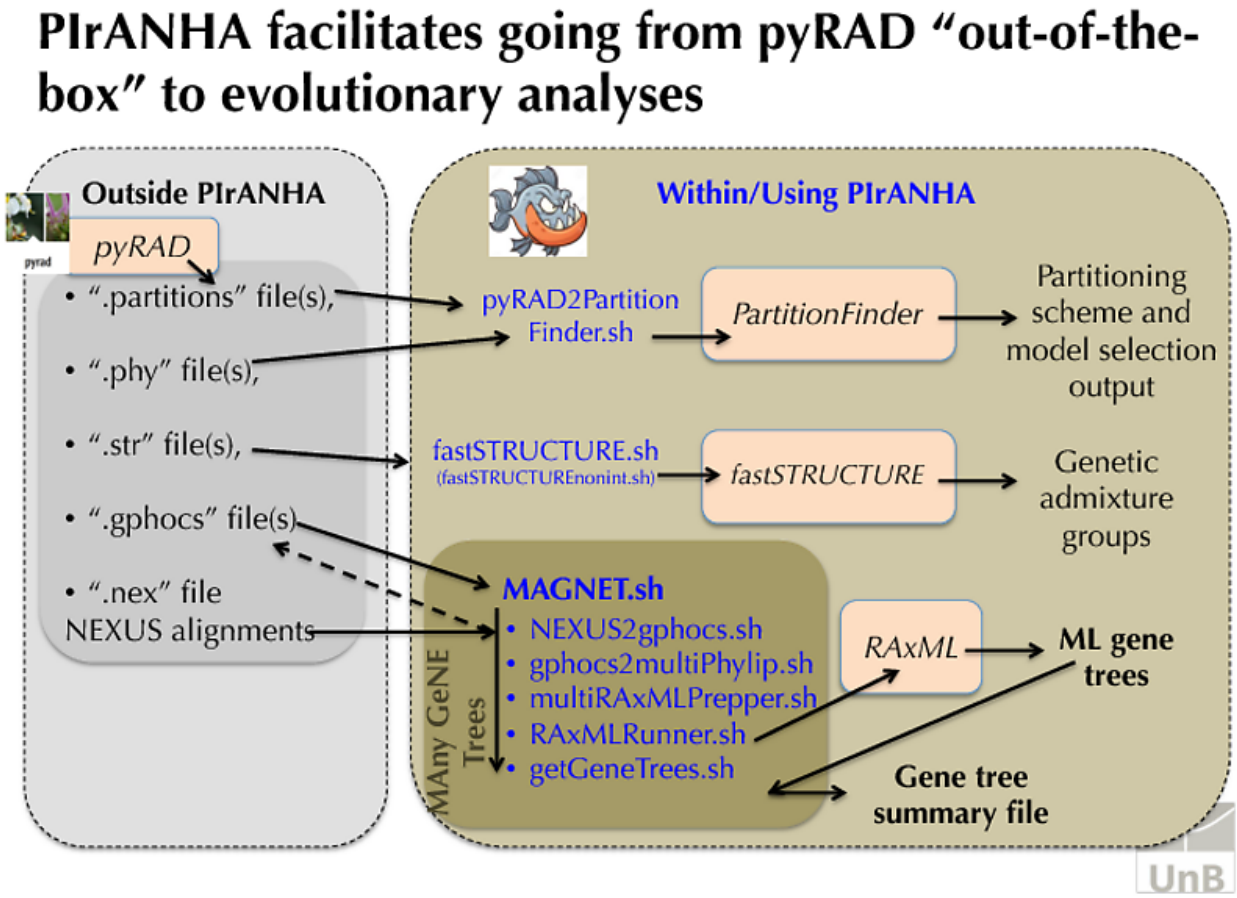

Figure 1

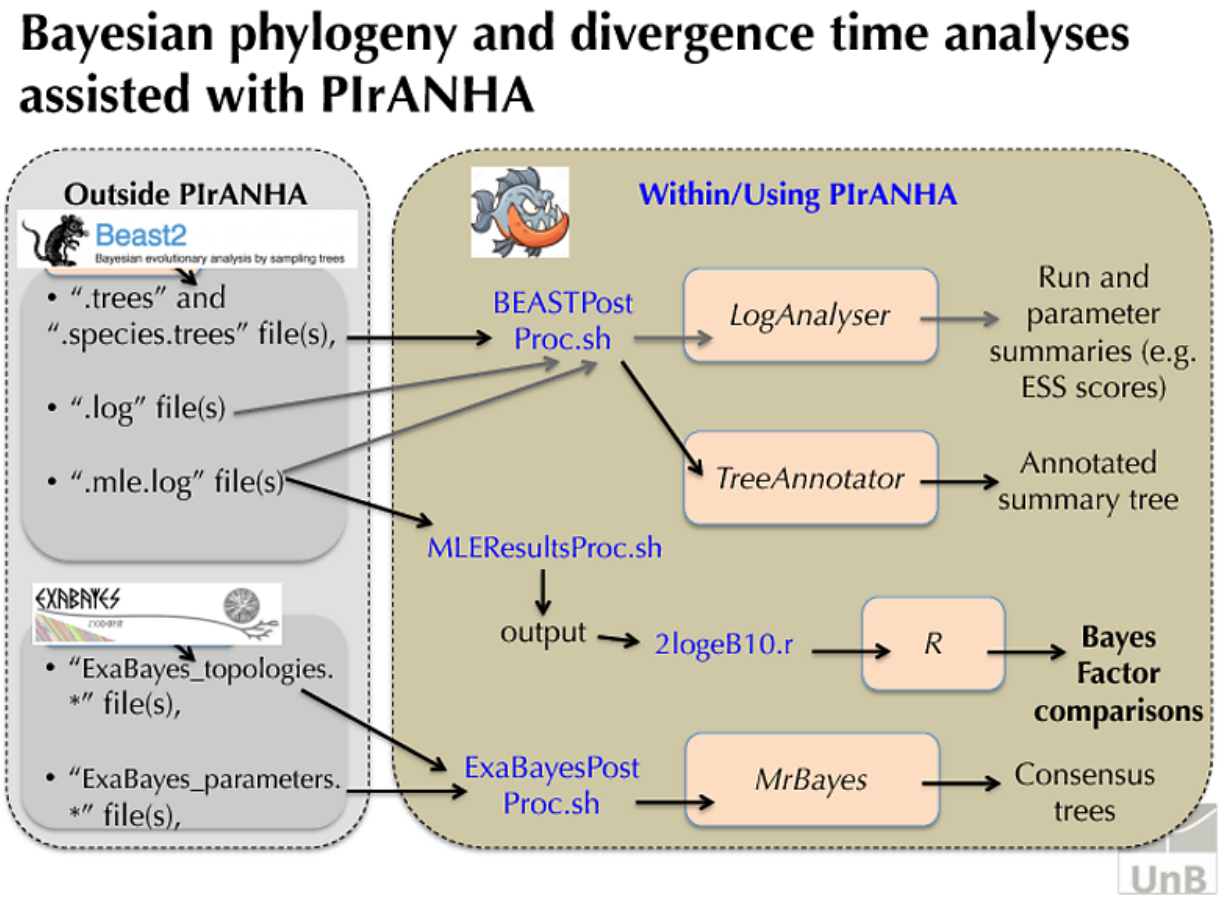

Figure 2

The following **Figure 3** illustrates new capacities of running and processing dadi (Gutenkunst et al. 2009) files in PIrANHA (note: the post-processing script is still under development).

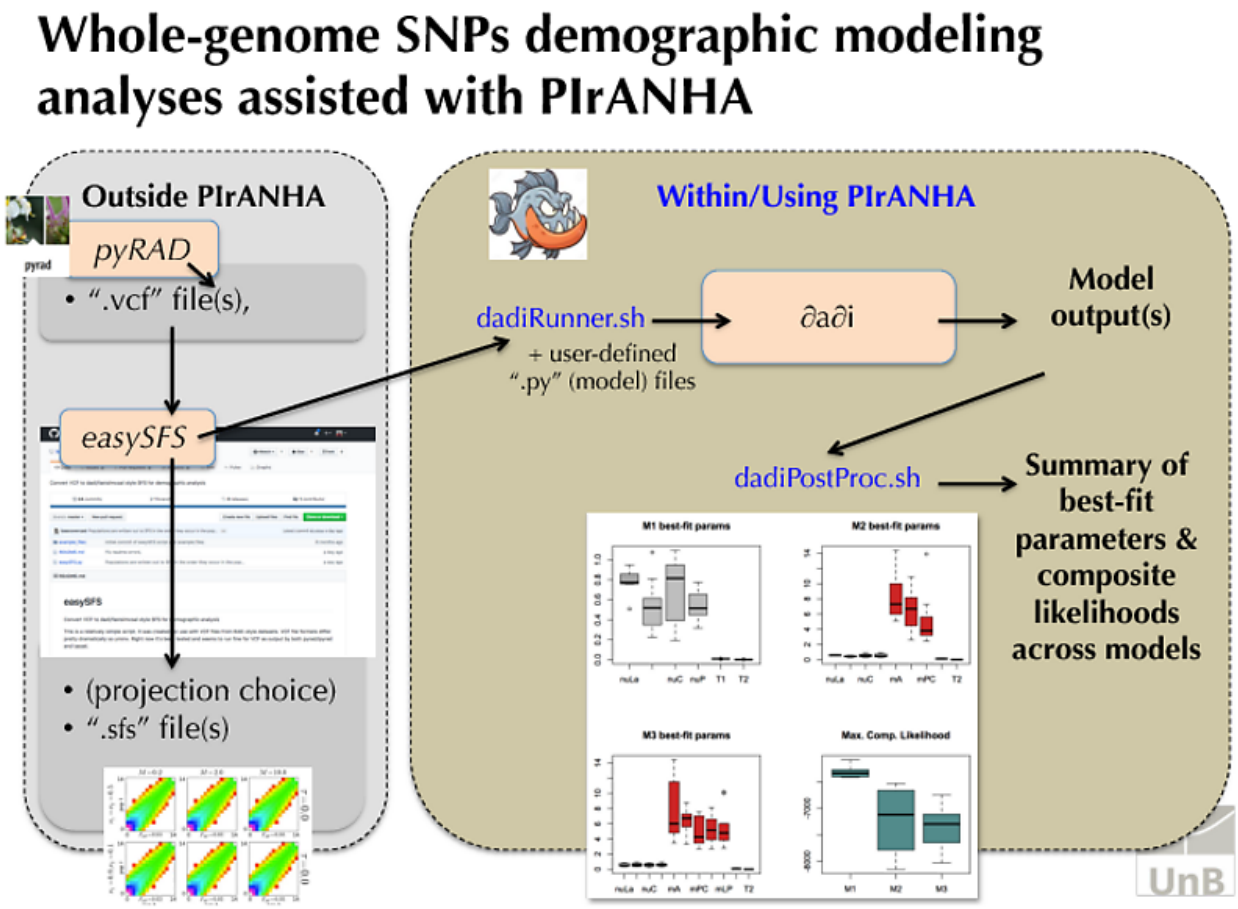

Figure 3

GETTING STARTED

Dependencies

PIrANHA, and especially the MAGNET package (here or here) within PIrANHA, relies on several software dependencies. These dependencies are described in some detail in README files for different scripts or packages; however, I provide a full list of them below, with asterisk marks preceding those already included in the MAGNET subdirectory of the current release. Of course, you can usually get away with not installing dependencies or software that are unrelated to the analysis you are conducting with PIrANHA, but it is recommended that you install all dependencies to take full advantage of PIrANHA's capabilities, or be prepared for any analysis!

- PartitionFinder
- BEAST v1.8.3 and v2.4.2 (or newer; available at: http://beast.bio.ed.ac.uk/downloads and http://beast2.org, respectively)
  - Updated Java, appropriate Java virtual machine / jdk required
  - beagle-lib recommended
  - default BEAST packages required
- ExaBayes (available at: http://sco.h-its.org/exelixis/web/software/exabayes/)

- RAxML (available at: http://sco.h-its.org/exelixis/web/software/raxml/index.html)
- Perl (available at: https://www.perl.org/get.html).
- *Nayoki Takebayashi's file conversion Perl scripts (available at: http://raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/perl-scripts.html; note: some, but not all of these, come packaged within MAGNET)
- Python v2.7 and/or 3+ (available at: https://www.python.org/downloads/)
- bioscripts.convert v0.4 Python package (available at: https://pypi.python.org/pypi/bioscripts.convert/0.4; also see README for "NEXUS2gphocs.sh")
- fastSTRUCTURE v1.0 (available at: https://rajanil.github.io/fastStructure/)
    - Numpy (available at: http://www.numpy.org/)
    - Scipy (available at: http://www.scipy.org/)
    - Cython (available at: http://cython.org/)
    - GNU Scientific Library (available at: http://www.gnu.org/software/gsl/)
- dadi v1.7.0 (or v1.6.3 as modified by Tine et al. 2014; available at: https://bitbucket.org/gutenkunstlab/dadi/overview)
- R v3+ (available at: https://cran.r-project.org/)

Users must install all software not included in PIrANHA, and ensure that it is available via the command line on their supercomputer and/or local machine (best practice is to simply install all software in both places). For more details, see the MAGNET README.

Installation

:computer: As PIrANHA is primarily composed of UNIX shell scripts and customized R scripts, it is well suited for running on a variety of types of machines, especially UNIX/LINUX-like systems that are now commonplace in personal computing and dedicated supercomputer cluster facilities. The UNIX shell is common to all Linux systems and mac OS X. There is no installation protocol for PIrANHA, because these systems come with the shell preinstalled; thus PIrANHA should run "out-of-the-box" from most any folder on your machine.

IMPORTANT! - Passwordless SSH Access

PIrANHA largely focuses on allowing users with access to a remote supercomputing cluster to take advantage of that resource in an automated fashion. Thus, it is implicitly assumed in most scripts and documentation that the user has set up passowordless ssh access to a supercomputer account.

:hand: If you have not done this, or are unsure about this, then you should set up passwordless acces by creating and organizing appropriate and secure public and private ssh keys on your machine and the remote supercomputer prior to using PIrANHA. By "secure," I mean that, during this process, you should have closed write privledges to authorized keys by typing "chmod u-w authorized keys" after setting things up using ssh-keygen.

:exclamation: Setting up passwordless SSH access is **VERY IMPORTANT** as PIrANHA scripts and pipelines will not work without setting this up first. The following links provide useful tutorials/discussions that can help users set up passwordless SSH access:

- http://www.linuxproblem.org/art_9.html
- http://www.macworld.co.uk/how-to/mac-software/how-generate-ssh-keys-3521606/
- https://coolestguidesontheplanet.com/make-passwordless-ssh-connection-osx-10-9-mavericks-linux/ (preferred tutorial)
- https://coolestguidesontheplanet.com/make-an-alias-in-bash-shell-in-os-x-terminal/ (needed to complete preceding tutorial)
- http://unix.stackexchange.com/questions/187339/spawn-command-not-found

Input and Output File Formats

:page_facing_up: PIrANHA scripts accept a number of different input file types, which are listed in Table 1 below. These can be generated by hand or are output by specific upstream software programs. As far as *output file types* go, PIrANHA outputs various text, PDF, and other kinds of graphical output from software that are linked through PIrANHA pipelines.

Input file types	**Software (from)**
.partitions	pyRAD
.phy	pyRAD (or by hand)
.str	pyRAD
.gphocs	pyRAD (or MAGNET/NEXUS2gphocs.sh)
.loci	pyRAD
.nex	pyRAD (or by hand)
.trees	BEAST
.species.trees	BEAST
.log	BEAST
.mle.log	BEAST
.xml	BEAUti
.sfs	easySFS
Exabayes_topologies.*	ExaBayes
Exabayes_parameters.*	ExaBayes

:construction: *NOTE: The following 'Getting Started' content is Under Construction!* :construction:

Phylogenetic Partitioning Scheme/Model Selection

pyRAD2PartitionFinder

Shell script for going directly from pyRAD output (de novo-assembled loci) to inference of the optimal partitioning scheme and models of DNA sequence evolution for pyRAD-defined loci. See current release of pyRAD2PartitionFinder scripts for more info (e.g. detailed comments located within the code itself; a README is coming soon).

Estimating Gene Trees for Species Tree Inference

MAGNET (MAny GeNE Trees) Package

Shell script (and others) for inferring gene trees for many loci (e.g. SNP loci from Next-Generation Sequencing) to aid downstream summary-statistics species tree inference. Please see the README for the MAGNET Package, which is available as its own stand-alone repository so that it can be tracked and continually given its own updated doi and citation by Zenodo.

Automating Bayesian evolutionary analyses in BEAST

BEASTRunner

BEASTRunner automates conducting multiple runs of BEAST1 or BEAST2 (Drummond et al. 2012; Bouckaert et al. 2014) XML input files on a remote supercomputing cluster that uses SLURM resource management with PBS wrappers, or a TORQUE/PBS resource management system. See the BEASTRunner README for more information.

BEAST_PathSampling

The BEAST_PathSampling directory is a new area of development within PIrANHA in which I am actively coding scripts to (1) edit BEAST v2+ XML files for path sampling and (2) automate moving/running the new path sampling XML files on a supercomputing cluster. This is very new stuff, as of January 2017, so stay tuned for more updates in the coming days/weeks.

ACKNOWLEDGEMENTS

I gratefully acknowledge *Nayoki Takebayashi*, who wrote and freely provided some Perl scripts I have used in PIrANHA. I also thank the Brigham Young University Fulton Supercomputing Lab (FSL) for providing computational resources used during the development of this software. J.C.B. received stipend support from a Ciência Sem Fronteiras (Science Without Borders) postdoctoral fellowship from the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; Processo 314724/2014-1). Lab and computer space was also supplied by The University of Alabama, during an internship in the Lozier Lab in the UA Department of Biological Sciences.

REFERENCES

- Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Molecular Biology and Evolution, 31, 2553-2556.
- Avise JC (2000) Phylogeography: the history and formation of species. Cambridge, MA: Harvard University Press.
- Bouckaert R, Heled J, Künert D, Vaughan TG, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology, 10, e1003537.
- Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, 30, 1844-1849.
- Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29, 1969-1973.
- Felsenstein J (2004) Inferring phylogenies. Sunderland, MA: Sinauer Associates.

- Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution, 27, 570–580.
- Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29,1695-1701.
- Lemmon AR, Lemmon E (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape. Systematic Biology, 57, 544–561.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One, 7, e37135.
- Raj A, Stephens M, and Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics, 197, 573-589.
- Tine et al. 2014. Nature comm.

RECOMMENDED READING

- Unix shell background info here, here, here, and here.
- GNU Bash Reference Manual

TODO

- ** Give supercomputer scripts options (header w/flags) that will work for both a) TORQUE/PBS and b) SLURM Workload Manager cluster management and job scheduling systems (need meticulous work on this in Super-pyRAD2PartitionFinder.sh, BEASTRunner.sh, BEASTPostProc.sh, and RAxMLRunner.sh) **
- Make pyrad and ipyrad batch run scripts available
- Consider separate scripts to work with ipyrad
- Add capacity of adding or not adding path sampling/stepping-stone sampling to BEAST runs (BEASTRunner.sh)

May 3, 2017 Justin C. Bagley, Richmond, VA, USA