Genetic and genomic analyses using RAD-seq and Stacks

de novo assembly of RAD tags without a genome for phylogenetic analysis

Instructors:

Julian Catchen <jcatchen@illinois.edu>
Department of Animal Biology, University of Illinois at Urbana-Champaign
William Cresko <wcresko@uoregon.edu>
Institute of Ecology and Evolution, University of Oregon
Josie Paris <J.R.Paris@exeter.ac.uk>
Molecular Ecology and Evolution Group, University of Exeter

Datasets and Software

- **Data sets - All are produced using an Illumina HiSeq 2500 sequencer**
  - ***Dataset 6 (DS6)*** - This is a set of population genomic data from several *Danio* species. The dataset comprises 1 individual from each of 15 species, 13 *Danios* and two outgroups. These data are from McClusckey and Postlethwait, 2015.

- **Software - All are open source software**
  - ***Stacks*** **(http://catchenlab.life.illinois.edu/stacks/)** - A set of interconnected open source programs designed initially for the *de novo* assembly of RAD sequences into loci for genetic maps, and extended to be used more flexibly in studies of organisms with and without a reference genome.
  - ***RAxML*** **(http://sco.h-its.org/exelixis/web/software/raxml/index.html)** - A software program written by the Exelixis Lab for the construction of maximum likelihood phylogenetic trees.

Exercise IV. Building a RADseq-based Phylogenetic Tree

1. In this exercise we will be working on a set of RAD-seq data taken from a number of *Danio* species, and two outgroups to *Danio*. We want to understand how these species relate to one another phylogenetically by examining variable sites between the genomes of these 15 species. The data set includes one sample per species. Our goal is to process the raw RAD data and reconstruct loci shared across the species. We will then feed these loci to the phylogenetic software RAxML to build our tree. *For more information on the study this data originated with, see McCluskey, et al. 2015, listed at the end of this document.*

Species	
Danio rerio (AB)	*Danio feegradei*
Danio rerio (Tubigen)	*Danio kerri*
Danio rerio (Nadia)	*Danio margaritatus*
Danio rerio (WIK)	*Danio nigrofasciatus*
Danio albolineatus	*Danio tinwini*
Danio choprae	*Devario aequipinnatus*
Danio danglia	*Microdevario kubotai*
Danio erythromicron	

2. There are a number of ways that GBS data can be analyzed for the construction of trees:

- We can look at only fixed differences between species. These are sites that are fixed within each species, but variable among species. The most common way to do this is to export a set of concatenated, fixed differences in Phylip format (http://evolution.genetics.washington.edu/phylip/doc/sequence.html).
- We can look at fixed and variable differences — this is an export of *all* variable sites concatenated together in Phylip format (variable sites are encoded in IUPAC format, https://en.wikipedia.org/wiki/Nucleic_acid_notation).
- We can look at the fixed and variable differences and include the surrounding invariant sequence again, all concatenated together. This is equivalent to exporting the full sequence for each RAD locus that contains one or more variable sites in Phylip format.
- We can look at the fixed and variable differences including the surrounding invariant sequence, but we can maintain each locus independently in the Phylip output. This allows us to pass a *partition* file to the phylogenetic software so that each locus can be analyzed independently.

3. In this study, we have only one sample per species. While this setup is standard in phylogenetics, in population genetics our goal is almost always to gather many representatives from each population as possible. **What are the implications for building phylogenetic trees from one of the above types of data when we only have one sample per species?**

4. In your `./working` workspace, create a directory called `phylo` to contain all the data for this exercise.

   Unarchive data set 6 (DS6):

   ```
   ~/workshop_materials/stacks/phylo/danio_phylo.tar
   ```

   into the `samples` directory.

5. Run the Stacks' `denovo_map.pl` pipeline program. This program will run `ustacks`, `cstacks`, and `sstacks` on the individuals in our study.

   [Once you get `denovo_map.pl` running, it will take approximately 30 minutes.]

   - Information on `denovo_map.pl` and its parameters can be found online:
     - http://catchenlab.life.illnois.edu/stacks/comp/denovo_map.php

- We want Stacks to understand which individuals in our study belong to which population. To specify this, create a file in the working directory called *popmap*, using an editor. The file should be formatted like this:

  ```
  <sample file prefix><tab><population ID>
  ```

  Include all 15 samples in this file and specify and assign each individual a different population ID (an abbreviation of the species name works well). You must supply the population map to `denovo_map.pl` when you execute it.

- We need to choose a name for the MySQL database that will hold the Stacks data (and allow us eventually to view it in the web interface). In this case, we will use the name, `danio_radtags`. A few notes:

  - We aren't required to use a database, and can tell Stacks to disable database access if we don't have that component of Stacks installed. In this course, we will use the database as it is installed and ready to go.
  - As a convention we append "`_radtags`" as a suffix onto all our RAD-tag databases. This is not necessary from a technical point of view, but the Stacks web interface will only show databases with this suffix (in case you have other, unrelated MySQL databases on your server).

- There are three important parameters that must be specified to `denovo_map.pl`, the *minimum stack depth*, the *distance allowed between stacks*, and the *distance allowed between catalog loci*.

  - These data are from a number of different species, therefore they may be much less related in terms of nucleotide divergence than a typical population genomics study. How might this influence how you set the distance allowed between stacks and the fixed distance allowed between individuals? You may also consider whether to use

gapped or non-gapped alignments: **how are gapped alignments related to the evolutionary distance between samples?**

- You must set the `stacks` directory as the output, specify the name of the database as well as a description of the data (which will be stored in the database), and use all the threads available on your instance. We need to a batch number for this run, use 1.
    - We can run the pipeline multiple times using the same database, specifying a different batch number each time. This allows us to store and look at a number of different batches, say with different parameters set for each one.
- Finally, specify the path to the directory containing your sample files. The `denovo_map.pl` program will read the sample names out of the population map, and look for them in the samples directory you specify.
- **Execute the Stacks pipeline.**

6. Examine the *Stacks* log and output files when execution is complete.

- How many reads are used in each `ustacks` execution and what is the final depth of coverage?
- After processing all the individual samples, denovo_map.pl will print a table containing the depth of coverage of each sample. Find this table in the log, what were the depths of coverage?
- Examine the output of the populations program in the log.
    - How many loci were identified?
    - How many were filtered and for what reasons?
- Familiarize yourself with the output of each Stacks' component:
    - `ustacks`: *.tags.tsv, *.snps.tsv, *.alleles.tsv
    - `cstacks`: batch_1.catalog.tags.tsv, batch_1.catalog.snps.tsv, batch_1.catalog.alleles.tsv
    - `sstacks`: *.matches.tsv
    - `populations`: *.sumstats.tsv, *.sumstats_summary.tsv

7. View the result of the Stacks analysis through the web interface:

- http://<Amazon Instance>/stacks/
- Explore the web interface
    - Why are some markers found in more samples?
    - Set the filters so that there are no fewer than 2 SNPs and no more than 5 SNPs per locus and so that there are at least 10 matching individuals per locus.
    - Select a locus (click on the locus ID in the left column) that has a reasonable ratio of genotypes (depending on your parameter choices you may have slightly different loci compared with another run of the pipeline). Click on `Allele Depths` to view additional information.

- Select a polymorphic sample, click on the alleles to see the actual stack that corresponds to the catalog locus.
  - Why do some nucleotides in the stack have a yellow background?
  - What are the different roles played by primary and secondary reads?
  - What is the `model` line in the output telling you, what does 'O' and 'E' stand for?
- Pick a locus from the `batch_1.sumstats.tsv` file and look it up in the web interface. Do the data match?

8. Now you need to run `populations` again with your population map and this time export the data for phylogenetic analysis.

You should export two Phylip data sets:

1. The first should maximize the number of taxa in the analysis — that is, you should select loci that are present in all loci in your tree.
2. The second should maximize the number of loci in the analysis — that is, you should require a minimum number of taxa that provides as many loci as possible for the analysis.

Think carefully about what is the best combination of taxa and loci and adjust the number of required populations parameter for the `populations` program (do not set this parameter too low or you might run out of memory on the Amazon Instance).

9. Run the RAxML software on each of your two data sets (the program is called `raxmlHPC` on the cluster) and choose a GTRCAT model and disable rate heterogeneity.

10. Download the resulting *best* tree in Newick format (https://en.wikipedia.org/wiki/Newick_format) from each run and visualize your trees on your local laptop.

- A good program for visualizing trees on your laptop in Newick format is FigTree (http://tree.bio.ed.ac.uk/software/figtree/).