Genetic and genomic analyses using RAD-seq and Stacks

Instructors:

Julian Catchen <jcatchen@illinois.edu>
Department of Animal Biology, University of Illinois at Urbana-Champaign
Josie Paris <J.R.Paris@exeter.ac.uk>
Molecular Ecology and Evolution Group, University of Exeter
Konrad Paszkiewicz <K.H.Paszkiewicz@exeter.ac.uk>
Exeter Sequencing Service, Biosciences, University of Exeter

Objectives:

The goal of this exercise is to familiarize students with the use of next generation sequence data produced from Reduced Representation Libraries (RRL) approaches such as Restriction site Associated DNA (RAD-tags). These libraries are often used for genotyping by sequencing, and can provide a dense set of single nucleotide polymorphism (SNP) markers that are spread evenly across a genome. These markers are useful for a variety of genetic and genomic analyses in model and non-model organisms. Students will gain experience with a computational pipeline called *Stacks* that was designed for the analysis of such data. Data will be analyzed *de novo* to perform a population analysis without the aid of a reference genome, and from an organism with a reference genome to identify signatures of selection. *Stacks* can be used for other analyses of RAD-seq data as well, such as constructing genetic maps and phylogeography, although those are beyond the scope of this exercise.

Students will learn how to:

1. Prepare raw RAD Illumina data for analysis by removing low quality reads and de-multiplexing a set of barcoded samples.
2. Use *Stacks* to assemble RAD-tags *de novo* from several populations of samples.
3. Call SNPs, genotypes, and haplotypes of these individuals within *Stacks*.
4. Export data from *Stacks* for analysis in Structure.
5. Align RAD sequences against a reference genome.
6. Use *Stacks* to perform genome scans using population genetics statistics like F_{ST}.
7. Use *Stacks* to generate RAD loci in order to build phylogenetic trees.

By the end of this workshop you will be expected to know how to:

8. Manipulate raw RAD Illumina data for analysis using a variety of different parameters.
9. Produce *de novo* assemblies using reads from an organism without a reference genome.
10. Align RAD tags against a reference genome to identify signatures of selection.
11. Extend what was learned to more complicated 'on your own' problems.

Introduction

The advent of short-read sequencing technologies has revolutionized the study of genomic variation from complete sequences in model organisms such as nematode worms, fruit flies, zebrafish, mice and humans. In addition, the long-standing dream of biologists of having complete genomic information from numerous individuals from different populations in the lab and wild, the field of **population genomics**, is becoming a possibility for a variety of ecological and evolutionary studies.

Until just a few years ago the goal of acquiring complete genomic information from numerous individuals in many populations was out of reach for all but a small number of model organisms. For example, producing a high density genetic map for an organism required an immense investment of resources to first produce and then type the large number of genetic markers needed to adequately cover the genome. Furthermore, identifying genomic regions associated with phenotypic variation, or involved in the adaptation of organisms to novel conditions, was restricted to organisms for which re-sequencing projects produced a dense battery of genetic markers at a significant cost.

The limits to population genomic studies will gradually fade as the costs of second generation sequencing continue to drop. However, many studies using complete genome re-sequencing will not be feasible for a while because costs are still significant, high quality read lengths are still too short, and analysis remains challenging. Luckily, many population genomic studies can now efficiently be performed by using an alternative approach called genotype-by-sequencing that occurs by the sequencing of reduced representation libraries (RRL), and subsequent identification and scoring of SNPs and inference of haplotypes. Although they do not provide complete genomic information, these approaches provide a sufficient picture via data on hundreds of thousands of SNPs and haplotypes spread across a genome at a fraction of the cost of complete re-sequencing.

We developed one such approach called **R**estriction-site **A**ssociated **D**NA sequencing (RAD-seq), which has been used to identify signatures of selection, produce high density genetic maps, help assemble genomes, and be useful for studies of allelic specific transcriptional profiling. Because these data are so new, and the sample sizes of sequences often so massive, a critical related breakthrough has been the development of algorithms and software pipelines for the analysis of such data. We have produced *Stacks* for the analysis of RAD-seq data. You will learn how to analyze RAD data with and without the use of a reference genome with the goal of identifying population structure in one case and identifying signatures of selection in a second case. Through the completion of these tasks you will learn how to process RAD-seq data and use the software programs *Stacks*, *GSnap and Structure*.

For more information on RAD genotyping and related methods, in particular conceptual and statistical issues, see the papers listed at the end of this document. These papers will help you better understand both the molecular biology, computational analyses, and conceptual framework for the analysis of RRL data such as RAD.

Datasets and Software

- **Data sets - All are produced using a Illumina sequencers**
  - ***Dataset 1 (DS1)*** - This data set comprises just a small proportion of a lane of single-end standard RAD data.
  - ***Dataset 2 (DS2)*** - A fragment of a lane of paired-end RAD sequences that have been double-digested with two restriction enzymes.

    You will use these first two data sets to become familiar with the structure of RAD sequences, as well as to become proficient with the pre-processing (i.e. cleaning and de-multiplexing) of data before alignment or assembly.
  - ***Dataset 3 (DS3)*** - A fragment of a lane of paired-end, double-digest RAD sequences that contain a random oligo sequence to identify PCR duplicates.
  - ***Dataset 4 (DS4)*** - This dataset consists of two samples of single-end RAD data from the same set of samples, but constructed in two different libraries and sequenced independently.
  - ***Dataset 5 (DS5)*** - This dataset comprises a subset of RAD-seq data generated from 30 individuals from three populations of threespine stickleback, each of which has been individually barcoded. The RAD-seq data were prepared using the restriction enzyme *SbfI*, and sequenced using an Illumina sequencer. These data are a component of the data originally published in Catchen, et al. 2013.
  - ***Dataset 6 (DS6)*** - This is a set of population genomic data from the threespine stickleback. The dataset comprises 8 individuals from each of two differentiated populations, for a total of 16 barcoded individuals. The RAD data were prepared using the restriction enzyme *SbfI*, and sequenced using an Illumina sequencer. These data are published in Lescak, et al. 2015.
  - ***Dataset 7 (DS7)*** - This is a set of population genomic data from several *Danio* species, including *Danio rerio*, the zebrafish. 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. The pipeline has a Perl wrapper allowing sets of programs to be run. However, the software is modular, allowing it to be applied to many scenarios. You will use the Perl wrapper in class and the modules on your own.
  - ***GSnap*** **(http://research-pub.gene.com/gmap/)** - *GSnap* is a very fast and efficient software package used for aligning sequences against a reference genome. We will use *GSnap* to align RAD reads against the stickleback reference genome, and then analyze these reads within the *Stacks* pipeline. Although we will use *GSnap* for this exercise, many other algorithms and software exist for aligning against a reference genome, and these could be used in conjunction with *Stacks* as well.
  - ***Samtools*** **(http://samtools.sourceforge.net)** - A suite of software tools designed to perform a variety of common tasks with next generation sequencing data tools. The SAM and BAM were developed associated.
  - ***Structure*** **(http://pritch.bsd.uchicago.edu/structure.html)** - A software program originally written by Jonathan Pritchard and colleagues that uses Bayesian stochastic models of multi-locus genotype data. The package was written to estimate the distribution and abundance of genetic variation within and among populations, patterns that are now commonly called the *genetic structure* of populations.
  - ***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.

Demultiplexing, Cleaning, and de-cloning RAD tags

Exercise 1. Data preparation, part 1

1. 10 minute mini-lecture on Phred scores and the `process_radtags` cleaning algorithm.
2. The first step in the analysis of all short-read sequencing data, including RAD-seq data, is removing low quality sequences and separating out reads from different samples that were individually barcoded. This 'de-multiplexing' serves to associate reads with the different individuals or population samples from which they were derived.
3. In each exercise you will set up a directory structure on the remote server (in this case our TGAC Instance) that will hold your data and the different steps of your analysis. We will use the directory ~/`working` on the server to hold these analyses. Be careful that your are reading and writing files to the appropriate directories within your hierarchy. You'll be making many directories, so stay organized!

   - Each step of your analysis goes into the hierarchy of the workspace, and each step of the analysis takes its input from one directory and places it into another directory, this is known as a '**waterfall workspace**'. We will name the directories in a way that correspond to each stage and that allow us to remember where they are. A well organized workspace makes analyses easier and prevents data from being overwritten.
   - In `working`, create a directory called `clean` to contain all the data for this exercise. Inside that directory create two additional directories: `lane1` and `samples`. We will refer to the `clean` directory as the *working directory*.
   - Unarchive data set 1 (DS1):

     `/data/clean/lane1.tar`

     to the `lane1` directory.

     You can copy the file to your working directory and use `tar` to unarchive it, **or** you can change to your working directory and *untar* it without moving the file (this will save you time and will dump the unarchived files into the directory you are currently in).
4. Your decompressed files has millions of reads in it, too many for you to examine in a spreadsheet or word processor. Examine the contents of the set of files in the terminal (the `head`, `more`, and `tail` commands may be of use).

   - You should see multiple different lines with different encodings.
     - How does the FASTQ file format work?
     - How are quality scores encoded? (See the link to quality scores in Appendix.)
     - How could you tell by eye which type of encoding your data are using?
5. You probably noticed that not all of the data is high quality. In general, you will want to remove the lowest quality sequences from your data set before you proceed.

However, the stringency of the filtering will depend on the final application. In general, higher stringency is needed for *de novo* assemblies as compared to alignments to a reference genome. However, low quality data will almost always affect downstream analysis, producing false positives, such as errant SNP predictions.

6. We will use the Stacks's program `process_radtags` to clean and demultiplex our samples.

- Take advantage of the manual page for `process_radtags` on the Stacks website to find information and examples:

  http://catchenlab.life.illinois.edu/stacks/comp/process_radtags.php

- You will need to specify the set of barcodes used in the construction of the RAD library. Remember, each P1 adaptor in RAD has a particular DNA sequence (an inline barcode) that gets sequenced first, allowing data to be associated with samples such as individuals or populations.
- Enter the following barcodes into a file called `lane1_barcodes` in your working directory (make sure you enter them in the right format):
  - `AAACGG  AACGTT  AACTGA  AAGACG`
  - `AAGCTA  AATGAG  ACAAGA  ACAGCG`
  - `ACATAC  ACCATG  ACCCCC  ACTCTT`
  - `ACTGGC  AGCCAT  AGCGCA`
- Copy the remaining barcodes for this lane of samples from the file:

  `/data/clean/lane1_barcodes`

  and append them to your barcodes file in your working directory.
  - You can concatenate this file onto the end of your file using the cat command and the shell's append operator: `cat file1 >> file2`, or you can cut+paste.
  - Based on the barcode file, how many samples were multiplexed together in this RAD library? (The `wc` command can tell you this.)
- You will need to specify the restriction enzyme used to construct the library (*SbfI*), the directory of input files (the `lane1` directory), the list of barcodes, the output directory (`samples`), and specify that `process_radtags` *clean*, *discard*, and *rescue* reads.
- The `process_radtags` program will write a log file into the output directory. Examine the log and answer the following questions:
  - How many raw reads were there?
  - How many were retained?
  - Of those discarded, what were the reasons?
- What can the list of "sequences not recorded" tell you about the data analyzed and about the design of barcodes in general?

7. Rename five of the output files in the samples directory to use more meaningful names:

```
sample_AAACGG.fq    indv_01.fq
sample_AACGTT.fq    indv_02.fq
sample_AACTGA.fq    indv_03.fq
sample_AAGACG.fq    indv_04.fq
sample_AAGCTA.fq    indv_05.fq
```

Renaming files by hand is time consuming as you can tell from just working with this subset. Renaming all the files to more meaningful names can be greatly simplified by writing a shell script in an editor, such as Emacs, and then using the search/replace function. Then, you can execute the shell script to actually rename the files.

Exercise 1. Data preparation, part 2

1. We will now work with the second data set. These data contain paired-end reads that have been double-digested and dual barcoded. Each set of paired reads contains an inline barcode on the first read, and an indexed barcode on both reads. These are known as *combinatorial barcodes* as many unique combinations can be made from pairs of barcodes.

   - In `~/working/clean`, create a directory called `lane2` to contain the raw data for this exercise and create the directory `ddsamples` to contain the cleaned output.
   - Unarchive data set 2 (DS2):

     `/data/clean/lane2.tar`

     into the `lane2` directory.

2. Examine the contents of the pairs of files in the terminal again.

   - How are the FASTQ headers related between pairs of files?
   - Can you identify the indexed barcode in the FASTQ header?

3. We will again use the Stacks' program `process_radtags` to clean and demultiplex our samples.

   - You will need to specify the set of barcode pairs used in the construction of the RAD library.
   - Enter the following barcodes into a file called `lane2_barcodes` in your working directory (make sure you enter them in the right format):

     - `AACCA<tab>ATCACG` `CATAT<tab>ATCACG` `GAGAT<tab>ATCACG`
     - `TACCG<tab>ATCACG` `AAGGA<tab>CGATGT` `CAACC<tab>CGATGT`
     - `GACAC<tab>CGATGT` `TACGT<tab>CGATGT`

   - Copy the remaining barcodes for this lane of samples from the file:

     `/data/clean/lane2_barcodes`

and append them to your barcodes file in your working directory.

- You can concatenate this file onto the end of your file using the cat command and the shell's append operator: `cat file1 >> file2`, or you can cut+paste.
- How many samples were multiplexed together in this RAD library? (The `wc` command can tell you this.)

- You will need to specify the two restriction enzymes used to construct the library (*NlaIII and MluCI*), the directory of input files (the `lane2` directory), the list of barcodes, the output directory (`ddsamples`) and specify that `process_radtags` *clean*, *discard*, and *rescue* reads.
- The `process_radtags` program will write a log file into the output directory. Examine the log and answer the following questions:
  - What is the purpose of the four different output files for each set of barcodes?
  - How many raw reads were there?
  - How many were retained?

4. For one of the sets of output files in the `ddsamples` directory rename the *.1.fq file to use a more meaningful name. Then, concatenate the other output files onto the end, so that for the set of files:

```
sample_AACCA-ATCACG.1.fq
sample_AACCA-ATCACG.2.fq
sample_AACCA-ATCACG.rem.1.fq
sample_AACCA-ATCACG.rem.2.fq
```

We are left with a single output file:

```
indv_01.fq
```

- Why are we able to concatenate all the data together from both the single and paired-end reads?
- How will Stacks interpret the single-end reads versus the paired-end reads?

Exercise 1. Data preparation, part 3 [Optional]

1. We will now work with a third data set. These data contain paired-end reads that have been double-digested. Each set of paired reads contains an index barcode on the first read and an index barcode on the second read, however, the second index barcode is actually a randomly generated oligo sequence. This oligo can be used to identify sequences generated by PCR amplification (or any type of amplification process) and reduce them to a single copy.

- In `~/working`, create a directory called `clone`, and within `clone` create the directory `raw` to contain the unprocessed data for this exercise and create the directory `decloned` to contain the cleaned output.

- Copy the two files that make up data set 3 (DS3):

  `/data/clean/Undetermined_S0_R1_001.fastq.gz`

  and

  `/data/clean/Undetermined_S0_R2_001.fastq.gz`

  into the `raw` directory.

2. Examine the contents of the pairs of files in the terminal again.
   - Where is the index barcode and where is the random oligo sequence?
3. Run the `clone_filter` program from Stacks to identify and remove the PCR duplicates. It will be helpful to consult the manual page for `clone_filter` on the Stacks website:

   http://catchenlab.life.illnois.edu/stacks/comp/clone_filter.php
4. Examine the output of clone_filter.
   - What is the distribution of PCR duplicates and how many total duplicates were found?

On your own outside of class

1. Remind yourself of the use of shell tools and regular expressions in Unix:
   - Identify all of the barcodes in one of the sequencing files.
   - Count the number of occurrences of each barcode.

     (`cut, grep, sort, uniq`, and the pipe "|" are the commands you need)
   - Using a single shell command, extract out the restriction site from the single end reads of DS2 and print the distribution of restriction sites. Repeat this for the paired end reads. If you are experiencing lots of ambiguous RAD sites when running `process_radtags`, this procedure can help you diagnose if your restriction enzyme site is intact.
   - Determine the distribution of read lengths in your data set (you'll need to use `awk` in addition to `grep`). This is useful if you have variable length data, say from a MiSeq or IonTorrent machines and need to choose where to trim the data.
2. Write a shell script in Emacs to rename all the files in one execution.
   - Use a shell loop to do the hard work.
3. Try using the `process_radtags` program with a range of parameters.
   - Specify a different (incorrect) restriction enzyme.
     - How many reads were retained this time?
     - Of those discarded, what were the reasons?
   - Vary the sliding window score threshold, and the size of the window.
     - How do these changes to the parameter affect the number or retained reads?

Citations and Readings

Core readings in preparation for the lecture and workshop

Amores, A., et al. 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing. **Genetics** 188:799-808.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. **Nature Reviews Genetics** 1–12.

Arnold, B. et al. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. **Molecular Ecology** 22: 3179–3190.

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

Catchen, J. et al. 2013a. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. **Molecular Ecology**, 22:2864–2883.

Catchen, J. et al. 2013b. Stacks: an analysis tool set for population genomics. **Molecular Ecology** 22:3124–3140.

Davey, J. W., et al. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature Reviews Genetics** 12:499-510.

Davey, J. W. et al. Special features of RAD Sequencing data: implications for genotyping. **Molecular Ecology** 22: 3151–3164.

Etter, P. D., et al. 2011. SNP Discovery and Genotyping for Evolutionary Genetics using RAD sequencing. *in* Molecular Methods in Evolutionary Genetics, Rockman, M., and Orgonogozo, V., eds.

Ekblom, R., and J. Galindo. 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity** 107:1-15.

Hohenlohe, P. A. et al. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. **PLoS Genetics** 6: 1-23.

Lescak, E. A., Bassham, S. L., Catchen, J., gelmond, O., Sherbick, M. L., Hippel, von, F. A., & Cresko, W. A. 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. **Proceedings of the National Academy of Sciences** 112(52), E7204–12.

Population genomics background, concepts and statistical considerations

Cariou, M. et al. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. **Ecol Evol** 3, 846–852.

Gompert, Z., and C. A. Buerkle. 2011a. A hierarchical Bayesian model for next-generation population genomics. **Genetics** 187:903-917.

Hohenlohe, P. A., et al. 2010. Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. **International Journal of Plant Sciences** 171:1059-1071.

Luikart, G.,et al. 2003. The power and promise of population genomics: from genotyping to genome typing. **Nature Reviews Genetics** 4:981-994.

Lynch, M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. **Genetics** 182:295-301.

Nielsen, R., et al. 2005. Genomic scans for selective sweeps using SNP data. **Genome Research** 15:1566-1575.

Nielsen, R., et al. 2011. Genotype and SNP calling from next-generation sequencing data. **Nature Reviews Genetics** 12:443-451.

Rubin, B. E. R. et al. 2012. Inferring Phylogenies from RAD Sequence Data. **PLoS ONE** 7, e33394–e33394.

Stapley, J., et al. 2010. Adaptation genomics: the next generation. **Trends in Ecology and Evolution** 25:705-712.

Empirical studies using RRL and RAD sequencing

Altshuler, D., et al. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. **Nature** 407:513-516.

Baxter, S. W., et al. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. **PLoS ONE** 6:e19315.

Chutimanitsakun, Y., et al. 2011. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. **BMC Genomics** 12: 1-13.

Emerson, K. J., et al. 2010. Resolving postglacial phylogeography using high-throughput sequencing. **Proceedings of the National Academy of Sciences** 107:16196-16200.

Gore, M. A., et al. 2009. A first-generation haplotype map of maize. **Science** 326:1115-1117.

Richards, P. M. et al. 2013. RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. **Molecular Ecology** 22, 3077–3089.

RAD-seq genotyping methodology

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

Etter, P. D., et al. 2011. Local De Novo Assembly of RAD Paired-End Contigs Using Short Sequencing Reads. **PLoS ONE** 6:e18561

Hohenlohe, P. A., et al. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. **Molecular Ecology Resources** 11 Suppl 1:117-122.

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

Willing, E. M., et al. 2011. Paired-end RAD-seq for de novo assembly and marker design without available reference. **Bioinformatics** 27:2187-2193.

Related reduced representation library (RRL) methodologies

Andolfatto, P., et al. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. **Genome Research** 21:610-617.

Elshire, R. J., et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. **PLoS ONE** 6:e19379.

Peterson, B. K. et al. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. **PLoS ONE** 7, e37135.

Rigola, D., et al. 2009. High-Throughput Detection of Induced Mutations and Natural Variation Using KeyPoint™ Technology. **PLoS ONE** 4:e4761.

van Orsouw, N. J., et al. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. **PLoS ONE** 2:e1172.

van Tassell, C. P., et al. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. **Nature Methods** 5:247-252

Wang, S. et al. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. **Nature Methods** 9, 808–810.