Interested in RADseq in the McDaniel Lab?

EXECUTIVE SUMMARY – HAVE ANSWERS TO THESE QUESTIONS BEFORE COMING BY:

What is your question? What statistical approach will you use to answer that question? Do you have the computer skills necessary to implement that approach, given the oodles of data that you will generate?

What is your organism’s genome size? What genomic resources are available for your organism? Is your organism a recent polyploid? (hopefully no!).

Evaluate your DNA extraction method before you extract all of your samples. We need to see a gel (seriously!) – review the images of good and bad extractions at the end of this doc.

Familiarize yourself with the Illumina and RADseq literature (suggestions below).

Here is a collection of things to consider as well as pieces of information you will need to have prior to starting and/or planning a RADseq project.

1) Data analysis and computer skills
   a. All Illumina based sequencing projects generate an immense amount of data. Raw data files ranging in size from 20-50 gigabytes or larger are normal. If you don't have a plan to process and analyze the data you will be instantly overwhelmed.
   b. Realistically evaluate your computer skills. If you are unfamiliar with Unix, command line programs, Perl or Python, cluster based computing, compiling programs, and Excel it is 100-percent-absolutely-critical you develop these skill prior to beginning a RADseq project.
   c. We often tell people that with 2 weeks of lab work you will generate 6-12 months of computer work for yourself. The data analysis portion is consistently what people struggle with most and often underestimate.

2) What exactly are the questions you are interested in trying to answer with a RADseq data set and what statistics are you planning to generate to answer those questions.

3) How may samples can I run?
   a. Like so many things in science the answer to this question is “well it depends”. Specifically it depends on these factors: The size of your organism’s genome, the questions you are trying to answer, the read depth needed to answer those questions, and the amount of sequencing you can afford.
   b. For example: you want to investigate hybridization and you have enough funding for library prep and a mid throughput NextSeq run (which will generate ~130 million reads). Since you are interested in hybrids detecting heterozygous loci is of paramount importance so you decide you need a read depth of at least 15 reads per loci for you to feel comfortable analyzing your data with your chosen method. If your organism’s genome
is small, say 500mb, and you estimate that EcoRI will cut on average 50,000 times per genome this will yield ~50,000 loci per individual. If you want 15x coverage per locus you will need 750,000 reads worth of sequence per individual to achieve this (50,000 x 15). If your sequencing run can generate 130M reads this means you could multiplex up to 173 samples (15M / 750,000). This of course is just an estimate to get you in the ball park, as with all Illumina sequencing there can be a large amount of stochastic variation in both number of reads per individual as well as in read depth per locus. Were this a real scenario we might suggest multiplexing 125 individuals. Additional if the genome size is larger or the frequency of EcoRI cut sites is higher, or you desire greater read depth, or you are trying to detect signatures of rare events, the number of individuals you can include changes.

4) What is the size of your organisms genome?

5) What genomic resources, if any, exist for your organism?

6) Is your organism a recent polyploid? (Hopefully your answer is no)

7) How familiar are you with RADseq literature?
   a. We use a modification of the ddRADseq protocol by Peterson et al 2012. Take some time to thoroughly read the paper as well as look at their supplementary information.
   b. Molecular Ecology consistently publishes high quality research using RADseq as well as discussion of new methods and limitations of RADseq data. If you are looking for good papers, start looking in this journal.

8) What the McDaniel lab can and can not help with:
   a. Can help with:
      i. Building the Illumina libraries. We can barcode up to 400 separate samples for multiplexing in a single Illumina run.
      ii. Equipment, but not direct assistance, for large scale DNA extractions/DNA evaluation:
         1. 96 sample tissue grinder
         2. Refrigerated dual plate centrifuge
         3. 184 well slab gel electrophoresis
      iii. Answering quick or specific questions about RADseq computer programs, data analysis, data processing.
   b. Can not help with:
      i. Data processing or analysis, it simply is too time consuming.
      ii. Teaching basic molecular lab techniques, you need to come with a solid well practiced background in basic molecular biology lab skills (making and running agarose gels, PCR, enzyme based reactions, proper pipette skills for handling very small volumes,
9) DNA quality and quantity.
   a. While this seems pretty straight forward it is often something people have a difficult time with. In general you want to have the highest quality DNA possible (ie. intact high molecular weight DNA).
   b. Run an agarose gel of your DNA extractions to visualize the quality of the DNA. There is not a more efficient way to assess quality that through a gel. I can not stress the importance of this enough, **run gels of the DNA as you are doing the DNA extractions** that way you can address problems early on, not after all the extractions are complete. Quantification is no substitute for a gel as gels assess quality (Fig. 1). We use 0.7-1% agarose run at 120 volts for 25 min. Make sure you take a good, high quality photo of your gel (one you would be proud to send to your Mom or Advisor, ie. not overexposed, streaky, crooked, broken, insufficiently separated, ect…)
   c. The quantity is important but a large quantity of DNA is not essential, our protocol requires a minimum of ~100ng of DNA in a volume of 6ul.
   d. Assess your DNA quantity using a fluorometric method, such as PicoGreen or Qubit. Do Not use Nanodrop. Standardization of DNA quantities will be needed prior to library preparation, so be prepared to adjust sample concentrations as needed.

10) Here are a collection of useful programs (by no means exhaustive):
   a. STACKS (become intimately familiar with this program if you plan to use it for analysis, read the manual in your spare time, put it under your pillow at night, join the online forum and see what people are having a hard time with, ect..) *http://creskolab.uoregon.edu/stacks/*
   b. FASTX-Toolkit (for processing fastq data)  
   
   http://hannonlab.cshl.edu/fastx_toolkit/
   c. FastqUtils (an extension of NGS utilities also for processing data)  
   
   http://ngsutils.org/modules/fastqutils/
   d. Python script for insilico genome digestion  
   
   https://github.com/BU-RAD-seq/Digital_RADs

11) How familiar are you with Illumina sequencing technology?
   a. This is really cool stuff, search YouTube for videos of how the sequencing actually works. Compare that to the diagram in Peterson et al 2012 protocol to see the connection between what you make in the lab and how it is used by the sequencing instrument.
   b. Investigate the different platforms Illumina offers and what the differences, advantages, and drawbacks are of each, the three most common are MiSeq, HiSeq, and NextSeq.
Fig 1. Examples of DNA extractions of varying quality. In each lane 2μl of DNA was loaded in a 0.7% agarose gel.

A) high quality DNA with little to no fragmentation, this is the best case scenario for DNA extractions.
B) High quality DNA with some fragmentations, as seen in the smear below the bright band at the top, these samples would be good for library prep.
C) DNA of similar quality to B however these samples contain large quantities of RNA contamination as seen by the bright smears at the bottom. These samples would be good for library prep once the RNA is digested.
D) Mixture of DNA quality, lane 3 contains highly fragmented DNA and can not be used for library prep, lane 5 contains no high molecular weight DNA but does contain some moderate sized DNA, this sample may work for library prep, lane 7 is similar to lane 5 but the fragments are slightly larger.
E) Mixture of DNA qualities, lane 4-8 are very low quality DNA and can not be used for library prep, lanes 1, 2, and 3, contain a mixture of large and moderate sized fragmented DNA these three may work for library prep however they are low concentration as seen by the faintness of the smears.