**What to do when you get RADseq data back**

**Updated: Jan 2018 by KMD**

1) Download it from the sequencing provider as soon as you hear from them that it is done. They don’t keep files forever!

2) Store the data properly:

* Save the compressed data exactly as downloaded from the provider, to a data storage drive in the lab:
  + currently RADthistle’s /data or /data2 folders
  + The /data drive on the newer workstations
  + Please do not fill an external drive more than 80% full. (command ‘df’).
* Create the folder for the data using the following naming format:
  + Year\_Month\_Day\_Platform\_readLength
  + E.g. 2015\_10\_03\_Illumina\_100bpPE
* Immediately save the same data to our backup space on the HPC, in the same format:
  + sftp <yourNetID>@sftp.hpc.arizona.edu
  + cd /rsgrps/kdlugosch/dLab\_dataArchive
  + ‘mkdir’ your new folder and ‘put’ the \*.gz files there
  + NOTE: The HPC storage is only for archiving of compressed original data, and occasionally other small or temporary files. Do not store your split data here.
* Enter your run information on the ‘Illumina data directories’ sheet in this Drive folder.
  + This is very important for keeping track of which run is which!

3) Split the data by barcode:

* If you have not run command line scripts on our servers and workstations before, be sure to read the ‘server\_intro.pdf’ available in this Drive folder.
* Use the demultiplexing script, currently ‘Illumina\_splitting\_script\_template\_v6.pl’
  + A copy should be in this folder, see the top of the script for instructions.
  + There is also a version \*\_mod4parts.pl for cases where the download is too large to split all at once. In general, you will need RAM of at least twice the size of the uncompressed file (and uncompressed files are usually ~4x larger than compressed files).
  + If you need to split up a file, uncompress it and run the fastq-splitter.pl (in this folder) or other tool to split the R1 and R2 each into multiple files. For that script, use the setting that splits by read count, and e.g. 14,000,000 reads result in a file of about 5GB. Example command:  
    perl fastq-splitter.pl --part-size 14000000 --measure count MyRun\_R1.fastq
  + Run the demultiplexing script from your working drive on a workstation or server (not your laptop), referencing the data drive folder for the data files. The script will automatically uncompress the files as needed from the data drive onto your working drive for you, so you don’t need to make any extra copies.
* Please upload your barcode file to the folder ‘BarcodeFiles’ within this one, named in the same way as the run. This allows us all to see what was in each run and use it as needed into the future.
  + \*\*Be sure to see the splitting script for information about how your barcode file should be set up, and keep in mind that the R1 barcode will be forward, but the R2 will be reverse complement if you are using those
* Note that we have a standard naming convention for samples:
  + Species code (first two letters of genus and species)\_PopulationID-seedPacket-individual
  + E.g. CESO\_TRI-22-1

4) Clean the data

* Before using, the data need to be cleaned for low quality or adapter-contaminated sequences, and trimmed to be the same length for Stacks analysis.
* Note that trimming is something that can be optimized. While you want most of the read length, setting your trimming too long could discard many reads that became too short after low quality bases were removed from the end. Check the QC files supplied with your run to look at how quality decays over length. For example, the analysis of YST’s origins (Barker et al. 2017 Molecular Ecology) used 76bp of the 100bp reads (6 bp lost to barcode, 5 to enzyme restriction site, and 13bp to account for quality trimming). Recent runs have had little dropoff in quality at the end, so most length can be kept, but this is something that you need to look at.
* There is a pipeline script in this folder that cleans reads with SnoWhite (program available in this folder) and then trims them to length: ‘clean\_RADs\_pipe\_template\_v#.pl’
* After this, your data are in FASTA format and can be used in downstream programs such as STACKS, but the R1 and R2 reads are not necessarily in the same order anymore, as they have been processed separately and may have thrown out different reads. To order them for use of paired data, run script “sort\_R1\_R2\_v2.pl”

5) Upload the split data to NCBI

* Upload the split raw (uncleaned) data as soon as possible, and note that you do not have to make it publicly available right away.
* Do not merge any individuals that have multiple runs of data - each run should be uploaded separately.
* The NCBI upload is not strictly necessary to do immediately, but we don’t need to keep the pre-cleaned split raw data around for anything else. It takes up a lot of space, so it is ideal to upload it and then delete it, saving only the original raw, unsplit files (compressed), and your downstream files for analyses.
* Instructions are available in the file ‘Submitting to NCBI’ in this Drive folder.
* Record your sample information to ‘Dlab NCBI submission records’ in this Drive folder when you have submitted successfully.