Ecological Genomics Lab Notebook

Morgan Southgate

Overall Description of notebook

This notebook will catalog my entire work throughout the semester, including all my relevant code, results, and interpretations

Date started: January 24, 2018

Date end: (year-month-day)

Philosophy

Science should be reproducible and one of the best ways to achieve this is by logging research activities in a notebook. Because science/biology has increasingly become computational, it is easier to document computational projects in an electronic form, which can be shared online through Github.

Helpful features of the notebook

It is absolutely critical for your future self and others to follow your work.

- The notebook is set up with a series of internal links from the table of contents.
- All notebooks should have a table of contents which has the "Page", date, and title (information that allows the reader to understand your work).
- Also, one of the perks of keeping all activities in a single document is that you can **search and find elements quickly**.
- Lastly, you can share specific entries because of the three "#" automatically creates a link when the notebook renders on github.

Table of contents for 60 entries (Format is Page: Date(with year-month-day). Title)

- Page 1: 2018-01-24. Intro to Github, RMarkdown, and UNIX command-line
- Page 2: 2018-01-29. Practice logging into UNIX and basic coding.
- Page 3: 2018-01-29. Working with RNA-Seq Data
- Page 4: 2018-01-31. Continuing to work with RNA-Seq Data
- Page 5: 2018-02-05. Continuing to work with the RNA-Seq Data
- Page 6: 2018-02-07.Working with DESeq2
- Page 7: 2018-02-12.Continuing to work with DESeq2.
- Page 8: 2018-02-14. FInishing work with DESeq2.
- Page 9: 2018-02021.DESeq2 one more time!
- Page 10: 2018-02-26. Intro to Population Genomics: SNP and genotype calling
- Page 11:2018-02-28.SNP and genotype calling continued
- Page 12: 2018-03-05.Admixture and Population Structure
- Page 13: 2018-03-07.Admixture & Population Structure cont.

- Page 14: 2018-03-08. R Script for HW1 - Differential Gene Expression
- Page 15:.
- Page 16:.
- Page 17:.
- Page 18:.
- Page 19:.
- Page 20:.
- Page 21:.
- Page 22:.
- Page 23:.
- Page 24:.
- Page 25:.
- Page 26:.
- Page 27:.
- Page 28:.
- Page 29:.
- Page 30:.
- Page 31:.
- Page 32:.
- Page 33:.
- Page 34:.
- Page 35:.
- Page 36:.
- Page 37:.
- Page 38:.
- Page 39:.
- Page 40:.
- Page 41:.
- Page 42:.
- Page 43:.
- Page 44:.
- Page 45:.
- Page 46:.
- Page 47:.
- Page 48:.
- Page 49:.
- Page 50:.
- Page 51:.
- Page 52:.
- Page 53:.
- Page 54:.
- Page 55:.
- Page 56:.
- Page 57:.
- Page 58:.
- Page 59:.
- Page 60:.

Page 1: 2018-01-24. Notes on using Github, Rmarkdown, and the UNIX command-line

Today, we created our github repos for the course, and began our notebooks.

Other goals for today: * publish our notebooks * log into the UNIX server

---

Page 2: 2018-01-29. Practice logging into UNIX and basic coding.

Logged into UNIX server using Putty, accessed mydata, viewed cols_data.txt

---

Page 3: 2018-01-29. Working with RNA-Seq Data.

Goals: - Learn about the bull headed beetle (*Onthophagus taurus*) - Understand pipeline for processing RNA-Seq data - Learn how to write bash scripts and write scripts to process files in batches *.sh files* #! notation - Visualize and interpret Illumina data quality * fastq files * Phred scores

Bull headed beetle (*Onthophagus taurus*) - Native to mediterranean, deliberately introduced to Australia for pest control - Accidentally introduced to eastern US from unknown origin in 1970s - Experimental design: * Three populations reared in common garden experiment, from native range(Mediterranean), Italy (IT), western australia (WA), and North Carolina (NC) * Four developmental stages L3L (late third larval insert), PP1 (pre-pupae day 1), PD1 (pupae day 1), and AD4 (adults 4 days after ecolsion) * Both sexes (three individuals per sex) * 3 pops x 4 developmental stages x 2 sexes x 3 individuals = 72 samples * Sequenced on about 7 lanes of Illumina HiSeq 2500

The pipeline for processing transcriptomic data

1. Visualize, Clean, Visualize

- Visualize data quality using the FastQC program
- Clean raw data using the Trimmomatic program
- Visualize quality of cleaned data using FastQC

2. Download reference transcriptome assembly
3. Map (aka align) cleaned reads from each sample to the reference assembly to generate sequence alignemnt files (sam files) (Program: bwa, Input: fastq, Output: .sam)
4. Extract read count data from .sam files (the number of reads that map to each "gene")
5. Assemble a data matrix of counts for each gene for each sample
6. Analyze count data to test for differences in expression

1. Visualize,Clean,Visualize

- Access shared directory through pbio381 server
- Working with WA_PP1_M1 - open file and check top four lines
- Get Quality score output
- Clean file using fastqc, save output in homedirectory/mydata

```
cd /data/project_data/beetles/rawdata/
zcat WA_PP1_M1_CCGTCC_L003_R1_001.fastq.gz | head -n4
@@@DDDDDA?DCFEGGEHHHGGEGGII+C1?EHBHGG##0:BDDFGIIDHG1=CGIGGGEHB3?D>CCAA;?:>>?BA@CB@A>ACCCBBB#+8?BBC#
WA_PP1_M1_CCGTCC_L003_R1_001.fastq.gz -o ~/mydata
```

- Move html file generated to documents/EcologicalGenomics/ using WINSCP

- Opened fastq file in html document

What is a FastQ file?

The fastq file format has 4 lines for each read:

Line	Description
1	Begins with '@' and then information about the read
2	The actual DNA sequence
3	Begins with '+' and sometimes same info as line 1
4	A string of characters representing the quality score, always with same number of characters as line 2

What is a Phred Quality Score (Q Score)?

Useful reference for Phred Q scores

P = 10^(-Q/10) Q = -10log10(P)

Phred Q Score	Prob Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

The Phred Q Score is translated to ASCII characters, meaning that a two digit number can be represented by a single character:

Quality encoding: ¡'#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI | | | | | Quality score: 0. . . . . ..10. . . . . ..20. . . . . ..30. . . . .

Page 4: Continuing to work with RNA-Seq Data

1. Visualize, Clean, and Visualize Again Continued

- Clean reads using Trimmomatic

1. Make directories “scripts” and “cleanreads”
2. Copy bash script over to ~/scripts directory

```
cp /data/scripts/trim_example.sh ~/scripts/ # copies the script to your home scripts dir
vim trim_example.sh                         # open the script with vim to edit
```

3. Open and edit the bash script using the program vim
4. Change the permissions on the script to make it executable, then run it *Trimmomatic performs cleaning steps in order presented - recommended to clip adapter early in process and clean for length at the end.

- The steps and options are from the trimmomatic website

ILLUMINACLIP:::: SLIDINGWINDOW:: windowSize: specifies the number of bases to average across requiredQuality: specifies the average quality required. LEADING: quality: Specifies the minimum quality required to keep a base. TRAILING: quality: Specifies the minimum quality required to keep a base. CROP: length: The number of bases to keep, from the start of the read. HEADCROP: length: The number of bases to remove from the start of the read. MINLEN: length: Specifies the minimum length of reads to be kept. ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read. SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold. LEADING: Cut bases off the start of a read, if below a threshold quality TRAILING: Cut bases off the end of a read, if below a threshold quality CROP: Cut the read to a specified length HEADCROP: Cut the specified number of bases from the start of the read MINLEN: Drop the read if it is below a specified length

```
#!/bin/bash
      java -classpath /data/popgen/Trimmomatic-0.33/trimmomatic-0.33.jar org.usadellab.trimmomatic.Trimm
                -threads 1 \
                -phred33 \
                 /data/project_data/beetles/rawdata/WA_PP1_M1_CCGTCC_L003_R1_00_R1.fastq.gz \
                 /data/project_data/beetles/rawdata/WA_PP1_M1_CCGTCC_L003_R1_00_R2.fastq.gz \
                 ~/cleanreads/"WA_PP1_M1_CCGTCC_L003_R1_001_clean_paired.fa" \
                 ~/cleanreads/"WA_PP1_M1_CCGTCC_L003_R1_001_clean_unpaired.fa" \
                 ~/cleanreads/"WA_PP1_M1_CCGTCC_L003_R1_002_clean_paired.fa" \
                 ~/cleanreads/"WA_PP1_M1_CCGTCC_L003_R1_002_clean_unpaired.fa" \
                 ILLUMINACLIP:/data/popgen/Trimmomatic-0.33/adapters/TruSeq3-PE.fa:2:30:10 \
                 LEADING:28 \
                TRAILING:28 \
                SLIDINGWINDOW:6:28 \
                HEADCROP:12 \
                MINLEN:35 \
```

Now run the script `chmod u+x trim_example.sh    # makes the script "executable" by the "user"`
`./trim_example.sh           # executes the script, or bash trim_example.sh` Output:

```
TrimmomaticPE: Started with arguments: -threads 1 -phred33 /data/project_data/beetles/rawdata/WA_PP1_M1_
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse onl
```

- Run FASTQC on our clean and paired reads, and store in fastqc directory
- save a copy of clean and paired reads as html file in ~EcologicalGenomics/1-31_RNASeq

```
fastqc ~/cleanreads/WA_PP1_M1_CCGTCC_L003_R1_clean-paired.fa -o ~/fastqc/
```

2: Download reference transcriptome

3: Map reads to the reference transcriptome

- First step is to index the reference transcriptome. Done already, but here's the command:

  ```
  $ bwa index /data/project_data/beetles/reference/OTAU.fna
  ```

- This generates 5 indexing files that all start with OTAU.fna in the /reference directory
- make a new directory sam in mydata
- Our first step is to map the clean reads using the bwa mem command
- output is a .sam file

  ```
  bwa mem <ref.fa> <read1.fq> <read2.fq> > <aln-pe.sam> # you fill in the inputs and outputs!
  ```

  ```
  bwa mem /data/project_data/beetles/reference/OTAU.fna ~/cleanreads/WA_PP1_M1_CCGTCC_L003_R1_clean-p
  ```

Page 5: Continuing to work with RNA-Seq Data

navigate to mydata/sam

```
tail -n 100 WA_PP1_M1_bwamem.sam > tail.sam   #take 100 last columns in file
vim tail.sam   #view tail.sam file
: set nowrap
```

A SAM file is a tab delimited text file that stores information about the alignment of reads in a FASTQ file to a reference gnome or transcriptome. For each read in a FASTQ file, there's a line in the SAM file that includes:

- the read (aka query name)
- a FLAG (number with info about mapping success and orientation and whether the read is the left or right read)
- Check FLAG scores using BROAD institue page (https://broadinstitute.github.io/picard/explain-flags.html)
- The reference sequence name to which the read mapped (if not mapped gets a *)
- The mapping quality (Phred-scaled)
- a CIGAR string that gives alignment information (how many bases Match, location of Insertion or Deletion)
- an '=', mate position, inferred insert size(columns 7,8,9)
- the query sequence and Phred-scaled quality from the FASTQ file (columsn 10 and 11)
- then lots of good information in TAGS at the end (if the read mapped) including whether it is a unique read (XT:A:U) the number of best hits (X0:i:1), and the number of suboptimal hits (X1:i:0)

The left (R1) and right (R2) reads alternate throughout the file. SAM files usually have a header section with generation information where each line starts with the '@' symbol. SAM and BAM files contain the same information, but SAM is human readable whereas BAM is in binary code, so has a smaller file size.

Here is the official SAM documentation Here is the tool for decoding the numbers in the second column of data Here is the reference for interpreting map quality scores

To get a summary of how well the reads mapped to the reference:

```
samtools flagstat *.sam
```

To see how many of the reads map uniquely: * Think about parology!

```
grep -c XT:A:U WA_PP1_M1_bwamem.sam  #use general expressions to search for specific term

samtools flagstat WA_PP1_M1_bwamem.sam

cp /data/scripts/countxpression_PE.py ~/myscripts  #move countxpression_PE.py file from shared directory

vim countxpression_PE.py  # visualize file
```

Page 6: 2018-02-07. Working with DESeq2.

Working with R on personal machine using:

-counts matrix -compiled table of uniquely mapped reads to each gene from each sample

DESeq2 - can rerun with bwa align, get python script to work with mem, or choose different way. 1) start with bwa align. Use WinSCP to transfer file into EcologicalGenomics/UNIXData/Feb7

work recorded in Feb7_DESeq2.R

Page 7: Continuing to work with DESeq2.

Page 8: Finishing work with DESeq2.

- Made a heatmap comparing expression levels
- Look at individual genes by devstage and sex using PCA

-GO enrichment using MWU test

Page 9: One more day with DESeq2!

Goals: -1) Explore how we can use DESeq2 and other packages in R to test for differences in gene expression of different models 3) To identify co-expression modules using WCGNA

- IT population didn't provide all developmental stages - so missing from some of this analysis
- made heat map of significant expression differences between WA_PP1_F and NC_PP1_F (also with corresponding male samples)
- WGNCA = weighted gene network correlation analysis, performed using package WGNCA.

_ brightest clusters and strongest correlations shown in developmental stage

Page 10: Intro to population genomics - SNP and genotype calling

Samtools: a powerful tool for manipulating sam files (and their binary equivalent, bam), and for piling up those reads across individuals to call SNPs and genotypes

Use samtools to: 1. convert from sam >> bam 2. check mapping stats 3. Fix reads that are no longer paired 4. Remove duplicates 5. Index for computational efficiency 6. Use bcftools to call SNPS and genotypes

navigate to server my data/sam open sam file

in bash: file saved as samtobam.sh in mydata/sam - make bash file executable using chmod o+x - go into screen - used to create a connection between terminal and computer, so that you can disconnect from terminal and come back later, or run a script in the background - run bash script using bash samtobam.sh - use ctrl + a + d command to detach from script

---

Page 11: SNP and genotype calling continued

- open samtobam.sh
- comment out all lines run in previous session
- error: WA_PP1_M1 accidentally renamed to WA_PP1_F1 partway thru script in previous session - fixed here by making bcftools output (when calling SNPs) with correct name
- | pipeline command in bash that puts output from previous command as input into next
- / allows splitting of command between multiple lines
- VCF = variant calling format
- columns associated with the number of samples in the data
- rows associated with SNPs present
- analyzing .vcf file
- head to check top
- wc to check length
- use tail to look at opposite end of file
- then open .vcf file using vim filename
- jump to a given line using :number
- and set no wrap using : set nowrap

PART 2: - navigate to /data/project_data/beetles/snps - these files contain data for all individuals - minDP (minimum depth to call SNPs) = 5 - minGP (minimum genotype probability) = 0.9

- make new folder called myresults in home directory, to receive output of analysis from shared directory
-

vcftools: –command format

Page 12: Admixture & Population Structure

In server: access data/project_data/beetles/snps Use vcftools to analyze vcf files produced by reads2snps

```
vcftools --vcf OTAU_2018_reads2snps_DP10GP95.vcf --min-alleles 2 --max-alleles 2 --maf 0.01 --max-missi
```

Use R in server to read in data, and calculate summary and df of F - F represents the inbreeding coefficent (heterozygosity in individual relative to subpopulation). Low inbreeding coefficient = excess of heterozygotes, high inbreeding coefficent = excess of homozygotes - Actual value depends on biology and life histroy strategies of organisms

```
>R
> df <- read.table("reads2snpsmiss0.8.het", header=T)
> head(df)
        INDV O.HOM.  E.HOM. N_SITES        F
1 IT_AD4_F1_  98985 98336.4  127352  0.02235
2 IT_AD4_F2_  88580 90908.7  117665 -0.08704
3 IT_AD4_F3_  99355 99213.6  128482  0.00483
4 IT_AD4_M1_  92147 90344.0  116911  0.06787
5 IT_AD4_M2_  93096 97362.4  125674 -0.15070
6 IT_AD4_M3_ 100851 99243.7  128559  0.05483

> summary(df$F)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.15070  0.02253  0.06285  0.05485  0.09509  0.16704

> sd(df$F)
[1] 0.06320409
```

Now run analysis using chosen model and parameters (reads2snps, max missing = 0.8)

```
vcftools --vcf OTAU_2018_reads2snps_DP10GP95.vcf --min-alleles 2 --max-alleles 2 --maf 0.01 --max-missi
```

Now pull out from vcf file individual populations

```
grep "IT" /data/project_data/beetles/metadata/cols_data.txt | cut -f 1 > IT.inds
```

And now

```
vcftools --vcf OTAU_2018_reads2snps_DP10GP95_biallelic_MAF01_Miss0.8.vcf.recode.vcf --keep IT.inds --fr

Parameters as interpreted:
        --vcf OTAU_2018_reads2snps_DP10GP95_biallelic_MAF01_Miss0.8.vcf.recode.vcf
        --keep IT.inds
        --freq2
        --out IT

Keeping individuals in 'keep' list
After filtering, kept 24 out of 72 Individuals
Outputting Frequency Statistics...
After filtering, kept 8730 out of a possible 8730 Sites
Run Time = 0.00 seconds
```

Page 13: Admixture & Population Structure Continued

- First, repeat steps from session 12 for NC & WA populations on server
- Next, load data into new R script (March7_PopGen.R) and calculate SFS for each popualtio

next: - thin vcf file, convert vcf to .geno, write bash script and run admixture

```
vcftools --vcf OTAU_2018_reads2snps_DP10GP95_biallelic_MAF01_Miss0.8.vcf.recode.vcf --thin 1000 --recod
```

- file is different than what other people have, so go back to last session and overwrite .vcf.recode.vcf file
- then move three PGD spider files (beetle.pop, beetle.spid, and bash script) to ~/myscripts
- edit beetle.pop and beetle.spid using vim script name
- try to run script in bash, not working, so:

- access copy of .geno file in /data/project_data/beetles/snps, move to /myresults
- make new bash script (ADMIX.sh) to run for loop with different values of K for admixture analysis

---

Page 14: HW-1 R Script

2 RScripts used: One to analyze differential gene expression between populations as a whole (HW1_wholePop), and one to analyze differential expression between each gender of each population separately (HW1_sexPop).

Page 15:

Page 16:

Page 17:

Page 18:

Page 19:

Page 20:

Page 21:

Page 22:

Page 23:

Page 24:

Page 25:

Page 26:

Page 27:

Page 28:

Page 29:

Page 30:

Page 31:

Page 32:

Page 33:

Page 34:

Page 35:

Page 36:

Page 37:

Page 38:

Page 39:

Page 40:

Page 41:

Page 42: