Announcements

- 12 November - New Notes on clustering in Week 9-10
- 11 November - My entire annotated DESeq2 procedure is in the DESeq2/EdgeR how-to.
- 4 November - Erythrobasidium data is available on globus. SRR4061760*.fastq.gz
  Use this data for a de novo assembly using Trinity, SOAPdenovo-Trans, Trans-ABySS, Oases, IDBA-Trans, or MIRA (or another program of your choice).
- 2 October - Check SRA for more data!
- 23 September - cleaned data in /depot/mgribsko/genomics/erythrobasidium/cleaned_data
  - 8044.7.89795.TTAGGC_r1.paired - paired end read 1
  - 8044.7.89795.TTAGGC_r2.paired - paired end read 2
  - 8044.7.89795.TTAGGC_r1.unpaired - SE (unpaired mate-pair)
  - 8044.7.89795.TTAGGC_r2.unpaired - SE (unpaired mate-pair)
  - 8255.2.95835.GTCCGC_r1.paired - paired end read 1
  - 8255.2.95835.GTCCGC_r2.paired - paired end read 2
  - 8255.2.95835.GTCCGC_r1.unpaired - SE (unpaired paired-end)
  - 8255.2.95835.GTCCGC_r2.unpaired - SE (unpaired paired-end)
- 21 September - see kmer plot and estimated fragment size on Assemblers and Group page
- 31 August - real data for the class is in /scratch/scholar/mgribsko/class_data/erythrobasidium/dna_raw. This is actual data from Erythrobasidium hasegawanum ATCC 9536.
  In case you missed it in class, the two files are different runs, the data in each file is interleaved.
- 31 August - use the previously posted data, now located in class_data_for_cleaning, for the week 3 tasks

Leaderboard

Insert your results here, order by best assembly first. These statistics come from Quast. use the paired end data for reads, e.g., 8255.2.95835.GTCCGC_r?.paired.fastq
Some of the Quast output are repetitive, just copy the ones below.

Keep at least one entry for each assembler, otherwise, if you have multiple assemblies, keep the ones that are "best" in at least on column.
Keep complete results on your group page under "Assemblers and Groups"

Use this color for gapfilled/scaffolded assemblies bests
1. reapr, gapfiller, ssGap, gapfiller of mer8 assembly

Questions

I think the slack page works better for questions

Participants

- Michael Gribskov, mgribsko
- Lizz Allmon, allmon
- Megan Fenton, fentom
- Sandra Gomez Guiterrez, sgomezgu
- Khaill Jaheed, kjahed
- Pedro Pablo Parra, pparrag
- Homa Salami, salami
- Claire Schraidt, cschraid
- Morgan Sparks, sparks35
- Omar Zayed, ozayed
- Tahira Fatima, fatima
- Yu Li, li2932
- Ramirez Camelo Luis

Week 2 - Introduction to Unix

Learn to run manipulate data and run programs on the Purdue RCAC clusters

- Read the Introduction to Unix, Manual, 1807unix_introduction.18.3.pdf
- Before 30 August - Complete the "unix Activities", UnixActivities.pdf
  - there was a typo in the activities, in #1, to see my bash profile type less ~mgribsko/.bash_profile

- "there was a typo in the activities, in #1, to see my bash profile type less ~mgribsko/.bash_profile"
Intro to Unix tutorial

Megan told me about this course, https://www.datacamp.com/courses/introduction-to-shell-for-data-science, which gives an introduction to unix and shell scripting. It can give you some practice in using some basic Unix commands, and an intro to shell scripting. I did the course in about 2 hrs (they say 4 hours), but I already know the material.

Week 3 - Data Cleaning, WGS

Complete cleaning your project data and be ready to choose a genome assembly program

I moved the tasks and reading to a subpage.

For tuesday:

- run fastqc on the practice data
- try to make a command file for trimmomatic

For thursday:

- complete quality and adapter trimming
- complete materials and methods statement
- be prepared to discuss materials and methods and your data cleaning results

Week 4 - Genome Assembly

tuesday

- complete cleaning of erythrobasidium data
- compare cleaned data and choose one to continue
- be prepared to discuss your trimming command and why you chose the option you chose

thursday

- discuss what assembly programs you want to use
- form groups

Week 5 - Genome Assembly

Week 6 - Genome Assembly

Tuesday - finish kmer analysis

Thursday - debug assemblies

Week 7 - Assembly quality

Tuesday - discuss assembly quality, compare assemblies

Make sure that you give the mate-pair library in the correct orientation with reapr. Here is a comparison of the results for my mer5 assembly (not the best assembly)
Week 9 - 10 - Differential expression

R script

```r
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
library(DESeq2)

# check where you are, find your data files
getwd()
dir()
dir("029251.salmon")

# try reading one file to see if the format is ok
t1<-read.delim("029251.salmon/quant.sf", sep="\t", row.names=1)
dim(t1)
head(t1)

# if it looks good, cut and paste to read all 9 files
t1.1<-read.delim("029251.salmon/quant.sf", sep="\t", row.names=1)
t1.2<-read.delim("029252.salmon/quant.sf", sep="\t", row.names=1)
t1.3<-read.delim("029253.salmon/quant.sf", sep="\t", row.names=1)
t2.1<-read.delim("029254.salmon/quant.sf", sep="\t", row.names=1)
t2.2<-read.delim("029255.salmon/quant.sf", sep="\t", row.names=1)
t2.3<-read.delim("029256.salmon/quant.sf", sep="\t", row.names=1)
t3.1<-read.delim("029257.salmon/quant.sf", sep="\t", row.names=1)
t3.2<-read.delim("029258.salmon/quant.sf", sep="\t", row.names=1)
t3.3<-read.delim("029259.salmon/quant.sf", sep="\t", row.names=1)

# copy the count files into a single matrix called raw
# the number of rows is the number of genes in the quant.sf files
# the number of columns is the number of columns
raw<-matrix(nrow=112229,ncol=9)
dim(raw)

raw[,1]<-t1.1[,4]
raw[,2]<-t1.2[,4]
raw[,3]<-t1.3[,4]
raw[,4]<-t2.1[,4]
raw[,5]<-t2.2[,4]
raw[,6]<-t2.3[,4]
raw[,7]<-t3.1[,4]
raw[,8]<-t3.2[,4]
raw[,9]<-t3.3[,4]

# give the columns names
colnames(raw)<-c("1.1","1.2","1.3","2.1","2.2","2.3","3.1","3.2","3.3")

# check to see it looks ok, and calculate some simple stats
head(raw)
sum(raw)                    # total read counts
colSums(raw)                # counts / sample

# a boxplot is a good way to see if some samples are very different in size
```
# but
# due to the wide range of scale, log(counts) is usually better
# can't take logs of zeroes so i make a new copy with 0 -> NA
rawna<-raw
rawna[rawna==0]<-NA
boxplot(log10(rawna),pch=20, cex=0.2)
boxplot(rawna, log="y",pch=20, cex=0.2)       # how are these two different?

# count histograms are useful for deciding where to prefilter
hist(log10(raw[,1]),breaks=50)                # histogram counts of sample
hist(log10(raw),breaks=50)                    # histogram counts of all
sample
hist(log10(rowSums(raw)),breaks=50)           # histogram counts of each
gene

# scatterplots are good for seeing non-linear effects, and to ge a general
# idea of noise level
pairs(log10(raw[,1:9]),pch=20,cex=0.1)

# now we need to put the data into the form DESeq2 uses, and define the
# experimental design. Note that the I round the raw counts to convert to
integer
metadata<-data.frame(sampleNames=colnames(raw),stage=c(rep("preclim",3),rep(  
"climacteric",3),rep("postclim",3)))
metadata
count<-DESeqDataSetFromMatrix(round(raw),colData=metadata,design=~stage)

# to prefilter, we need to first scale the samples to each other. DESeq2
does this
# to see the scaled counts, include normalized=T
pre<-DESeq(count)
head(counts(pre))
head(counts(pre,normalized=T))

prena<-counts(pre,normalized=T)
prena[prena==0]<-NA
boxplot(prena,log="y")
pairs(log10(counts(pre[,1:9],normalized=T)),pch=20,cex=0.1)