

```
MINGW64:/c:/Users/lvclark/Documents/Miscanthus_genome
lvclark@CPSC-P10F82924 MINGW64 ~
$ cd Documents/Miscanthus_genome/
lvclark@CPSC-P10F82924 MINGW64 ~/Documents/Miscanthus_genome
$ ls
140513GGseqPrepareForBowtie.fasta.txt
160127PstIMspIs_sizeselect.pdf
160127SbfIIMspI_sizeselect.pdf
160127simRAD.R
160127simRAD.R~
'170801 Jose_email.pdf'
'180306 PCA from Jose'/
180405_LVC_diversity_writeup.docx
180408_LVC_diversity_writeup.docx
180409_LVC_diversity_writeup.docx
chromosome_lengths.csv
completed_Msinensis_DH1_v7.0.plasts.fa_genotypes.mergeWithGatk.vcfBIS.min5_FMT-
DP.min1000.bcf.guessClarkSacksNames.xlsx
dh1_v3.1.bt2
dh1_v3.2.bt2
dh1_v3.3.bt2
dh1_v3.4.bt2
dh1_v3.fasta
```

Introduction to bash, and using bash with R

Lecture 13 part 1, CPSC 499 Fall 2018

What is bash?

- Stands for “Bourne again shell” (open-source version of Bourne shell)
- A language for running programs command-line and interacting with your operating system
- The language used in the terminal on Linux or Mac
- On Windows, is available with Git Bash or Cygwin

Important bash commands

- `cd` – change directory
 - `..` means the parent directory
 - `~` means the home directory
 - use `/` to indicate subdirectories
- `ls` – list all files in current directory
 - `ls -lh` also shows file sizes
- `#!/bin/bash` tells computer to execute this script with bash

Viewing files in bash

- `cat` – print out whole file
- `head` and `tail` – view beginning and end of file
- `less` – scroll through file (space to go down one page, q to quit)
- `nano` – opens up a user-friendly text editor

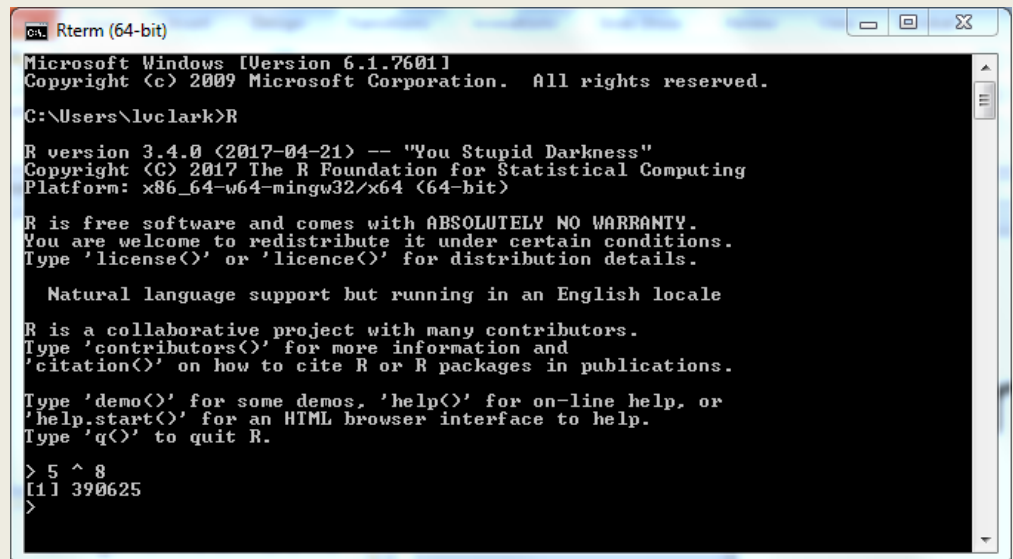
Giving arguments to bash commands

- After the command, type the arguments on same line
- Required arguments usually aren't named
 - e.g. `cat myfile.txt`
- Many arguments have one letter names
 - e.g. `head -n 20 myfile.txt`
- Some arguments have longer names
 - e.g. `head --verbose myfile.txt`

Running R from the command line

- Any operating system
- R must be in the PATH variable
- R: launches an interactive R session
- Rscript: executes a script non-interactively

q() to quit an
interactive
session



```
Rterm (64-bit)
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\lucclark>R

R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 5 ^ 8
[1] 390625
>
```

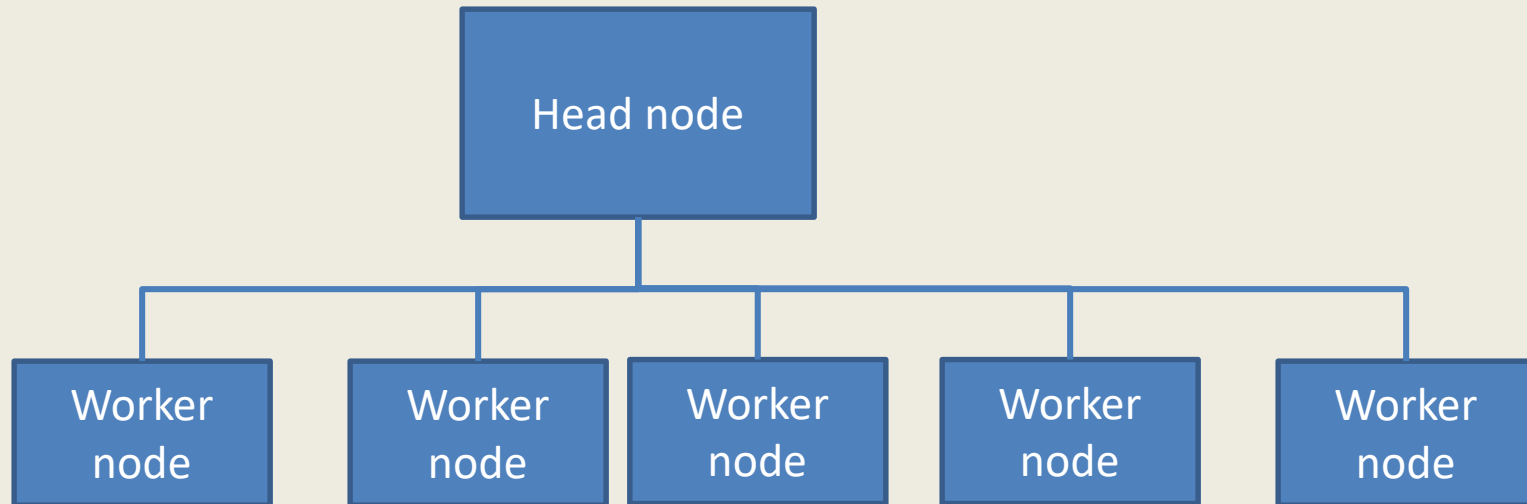
Sending arguments to your R script from bash

- E.g.
`Rscript myscript.R --be-awesome data.txt`
- Within your R script, use the function `commandArgs` to retrieve a character vector of the arguments
- `trailingOnly = TRUE` to only include the arguments you typed out

Piping

- Use `|` in bash to send output of one command to the input of the next command on the same line
 - e.g. `ls | head`
- Use `>` in bash to send output of command to a file
 - e.g. `ls > my_file_list.txt`
- Output of R script can be piped to another command (use `file = ""`)
- Output of another command can be piped to R script (use `file = file("stdin")` or `con = stdin()`)

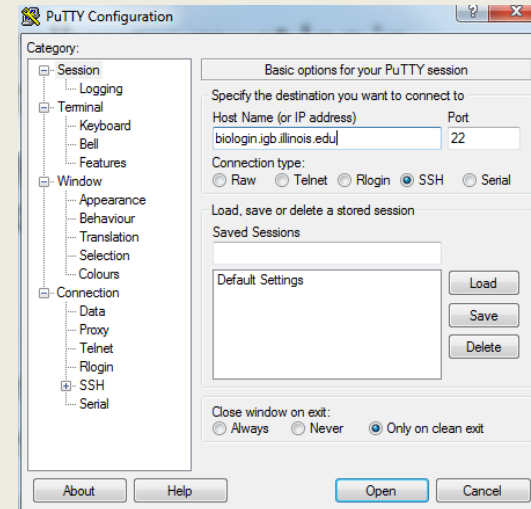
Computing clusters



- Each box represents a computer, with the same OS and software installed.
- All of them have access to the same file storage
- The head node manages the worker nodes, giving them jobs to do

Using a computer cluster

- Generally you must log in remotely
- Then you get command-line access to the head node
- Write a script for the job you want to do, submit it to the queue
- Once a worker node is available, the head node will send it your job
- Send and receive files by FTP



```
lvclark@biologin-1:~  
login as: lvclark  
lvclark@biologin.igb.illinois.edu's password:  
Last login: Mon Dec 11 13:24:36 2017 from th89-162.cropsci.uiuc.edu  
#####  
#                               #  
#      Institute for Genomic Biology      #  
#      University of Illinois Urbana-Champaign      #  
#      http://biocluster2.igb.illinois.edu      #  
#                               #  
#####  
*Please follow the guide at http://help.igb.illinois.edu/Biocluster2  
*All data on this cluster is NOT backed up. It costs $10 per terabyte  
per month  
*Please email help@igb.illinois.edu with any questions  
  
[lvclark@biologin-1 ~]$
```

Biocluster2

- See instructions at <https://help.igb.illinois.edu/Biocluster2>
- Maintained at the Carl Woese Institute for Genomic Biology
- Fee-for-use
- Has a lot of bioinformatics software installed
- You can request software to be installed

Example Slurm script to run an R script

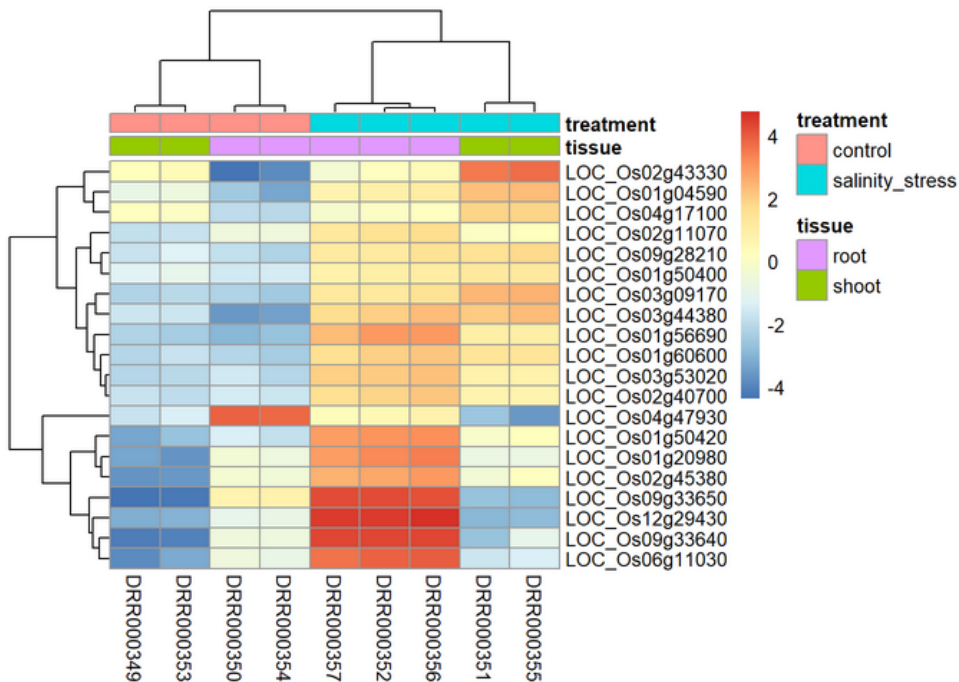
```
#!/bin/bash
#SBATCH -p normal
#SBATCH -mem=1g
#SBATCH -N 1
#SBATCH -n 1

module load R

Rscript testscript.R
```

Differential gene expression analysis with RNA-seq

Lecture 13, part 2

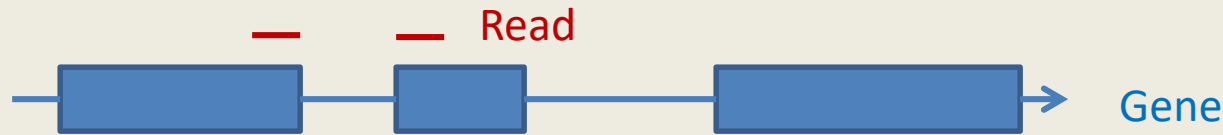


Analysis overview

- Input data: FASTQ files from sequencing all mRNA from several samples
- Goal: what genes are differentially regulated across:
 - Tissues
 - Treatments
 - Environments
 - Genotypes
 - Etc.

Short read-alignment for RNA-seq

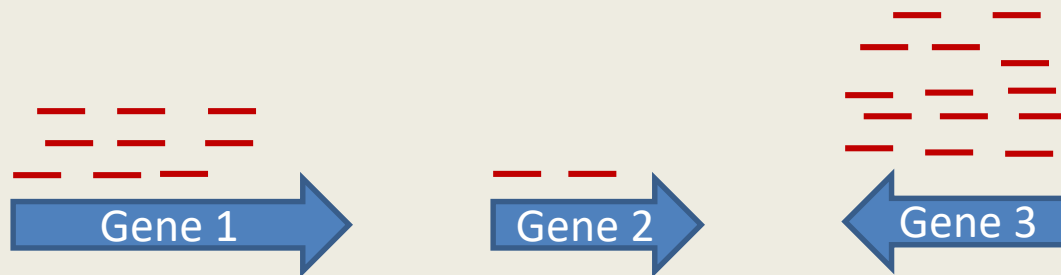
- A sequencing read might span two exons



- Use TopHat (which uses Bowtie2) to align reads to genome across junctions
- Other splice-aware mappers like SOAP-splice, gsnap, STAR
- Can give them genome annotation, or let them find splice sites de-novo

Transcript counting

- Next step after alignment: how many reads per gene?
- Cufflinks (part of Bowtie-TopHat suite)
- Multiple options within Bioconductor (we'll use GenomicAlignments)



Comparison of aligners for RNA-seq

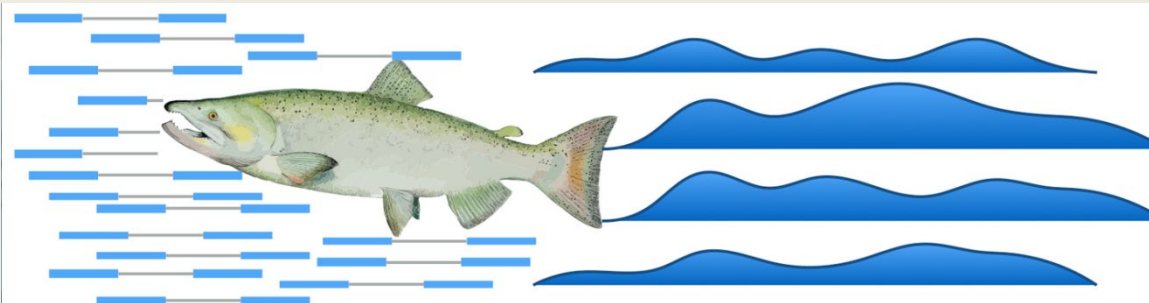
- TopHat was the standard for a long time but is falling out of use (slow)
- STAR is fast if you have a lot of RAM and hard drive space
- GSNAP can be used from within R (on Linux only)

Alternatives to traditional alignment

- Salmon, Sailfish, and kallisto
- Count up k-mers unique to transcripts, or do pseudo-alignment
- Much faster
- Match read to transcript but don't return an exact alignment
- Return matrix of counts, gene x sample
- Quickly becoming more popular than old methods

Salmon

- Index transcriptome rather than genome; doesn't find new transcripts or splice sites
- Does “quasi-mapping” to transcriptome
- Corrects for bias from GC-content and gene length
- If a read could belong to multiple transcripts, gets probability of it belonging to each one



Normalization of read counts

- Some seq. libraries will have higher depth overall than others, so counts need to be adjusted for that
- Among genes, expression varies on a log scale, so transformation is necessary for visualization

Small sample size

- RNA-seq is expensive, so often there are only a few biological replicates
- How can you test for significant difference between two means when $N < 10$?
- Software like DESeq and edgeR use data across all genes to estimate variance for each individual gene

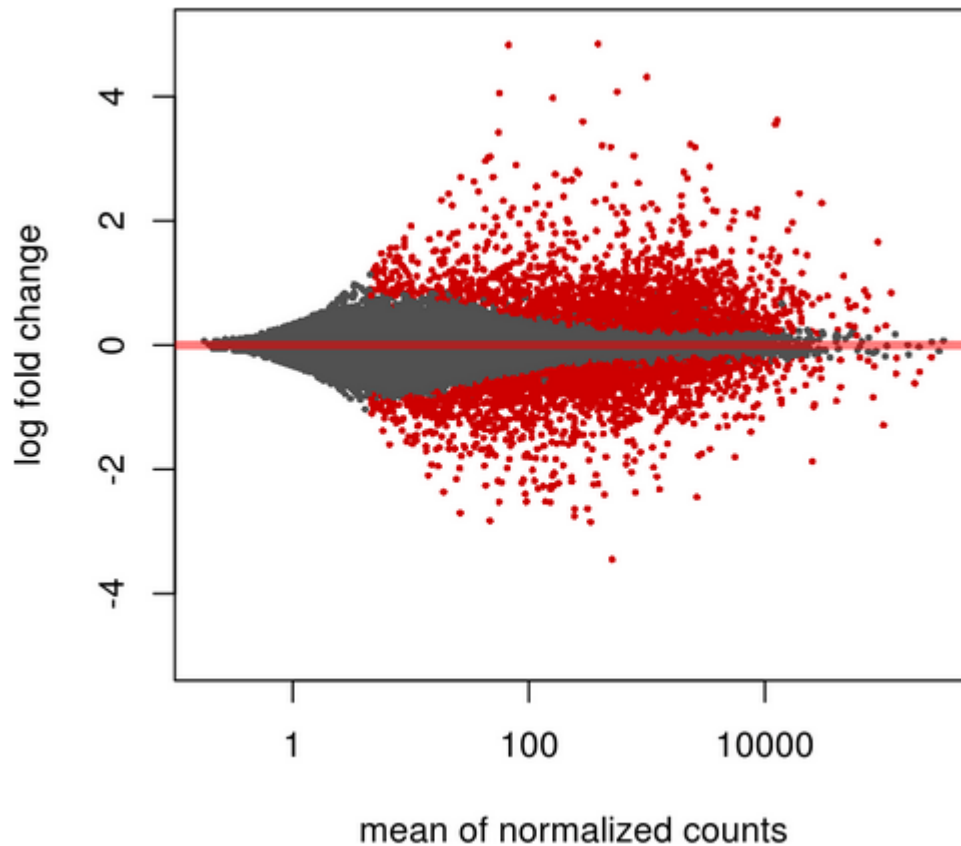
	Treatment 1			Treatment 2		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	205	150	172	220	231	243

$$P = ??$$

Output of differential expression analysis

- LFC = log fold change = effect size:
 - $\log\left(\frac{readcounts_{treatment}}{readcounts_{control}}\right)$
- *P*-value (significance; probability of data under null hypothesis that LFC = 0)
- Adjusted *P*-value: multiple testing correction reduces significance.

LFC and significance versus depth



Red = significant hits
Grey = not significant

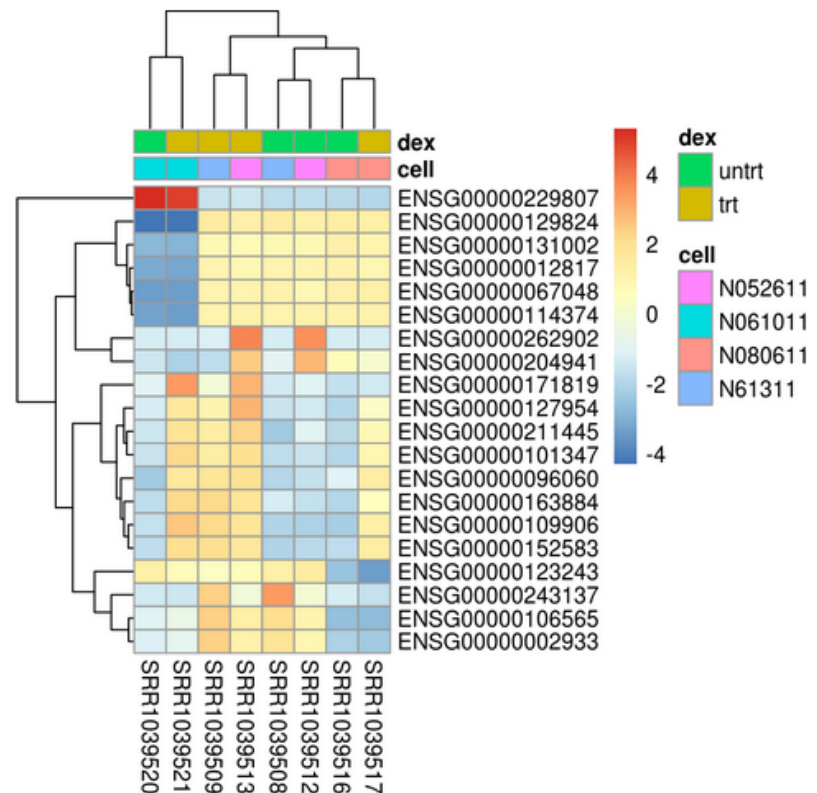
Detection of differential expression is more sensitive for highly expressed genes.

What do we do with significant genes?

- Gene ontology (GO) terms:
 - Numbers that correspond to gene functions
 - <http://www.geneontology.org/>
 - **Are particular terms enriched in the differentially expressed genes?**
- GO term categories
 - Biological processes
 - Cellular components
 - Molecular functions

What do we do with significant genes?

- Are there groups of genes upregulated or downregulated together?
 - Related functions
 - Part of a signaling pathway



Tuesday's lab

- We'll run through differential expression analysis with the same dataset, but using Salmon instead of TopHat