RNA-seq and Differential Expression

Presented by: Wesley Brashear

10/27/2021

Course Outline

1. RNA-seq overview
2. Differential expression pipeline overview
3. Using Grace
4. DE pipeline on Grace

What does RNA-seq data provide?

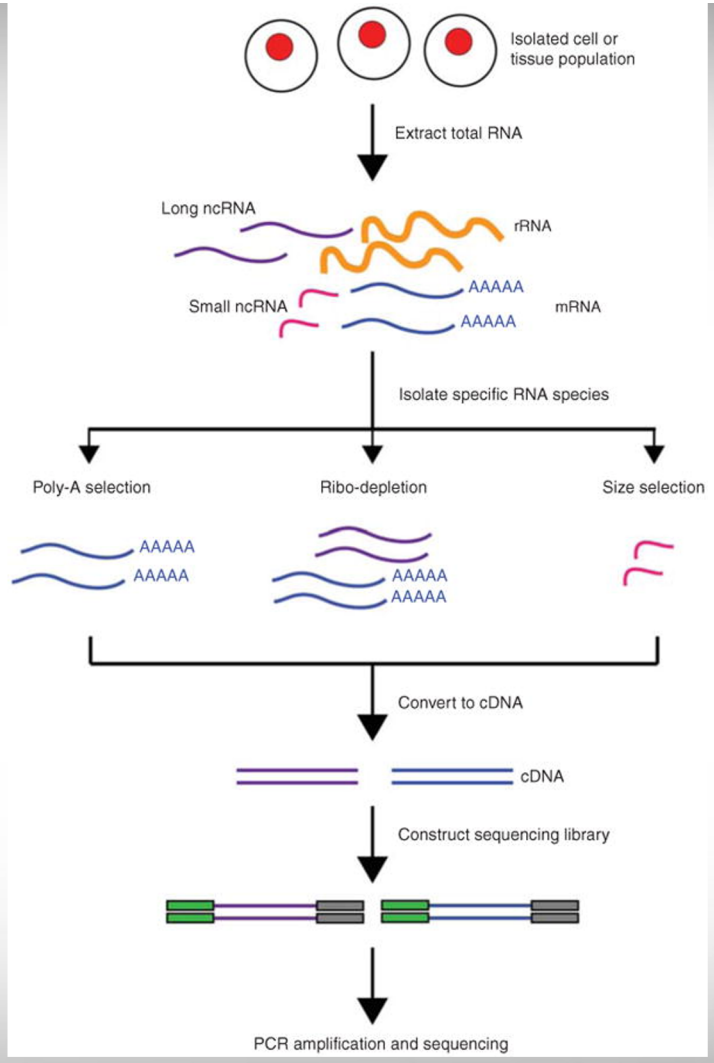
- Measure gene expression
- Detect differences in expression between groups
- Annotate genomes/transcripts
- Assemble transcriptome
- Discover nucleotide variants
- Scaffold genome assemblies

RNA-seq Applications

- Differential Expression (DE) and transcript abundance
  - HISAT2, STAR, TopHat, Cufflinks, Cuffmerge, Cuffdiff
  - DESeq and DESeq2 (R package)
  - EdgeR (R package)
- Transcriptome assembly (find isoforms and rare transcripts)
  - *de novo* (Trinity, Oases, SOAPdenovo-Trans)
  - reference based (Trinity, StringTie)
- Genome Annotation
  - Align to assembly for validation of gene models
- Variant Calling
  - STAR/Picard/GATK (Haplotype Caller (HC) in RNA-seq mode)
- *de novo* genome assembly scaffolding
  - L_RNA_scaffolder

RNA Sequencing

- Poly-A selection – enriches for mRNA
- Ribosomal depletion – removes rRNA, leaves mRNA, lncRNA, and pre-mRNA
- Size selection – used for smRNA



Kukurba and Montgomery, 2016

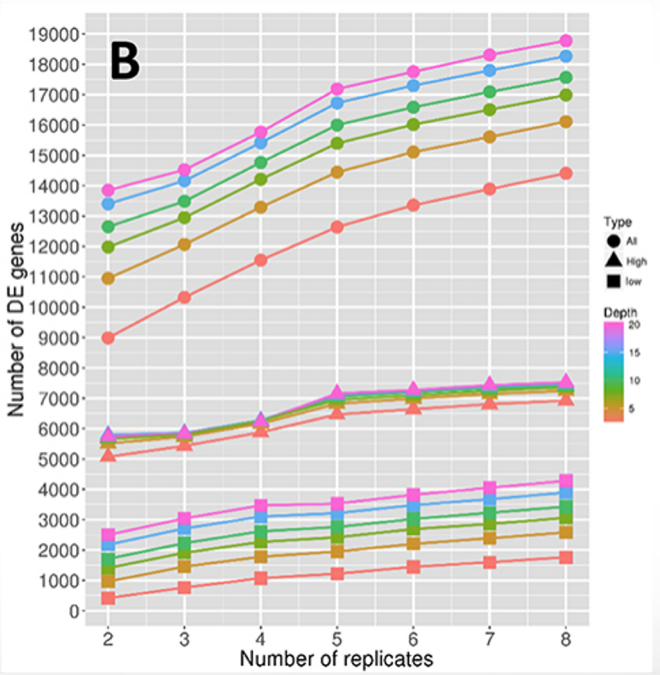
RNA-seq Experiment Design (Differential Expression)

Sequencing Depth

- Minimum 30 million aligned reads per replicate (ENCODE)
- 30-60 million reads per sample (Illumina)

Replicate Number

- 3 replicates per condition minimum (will likely recover 20-40% of DE genes)
- Schurch et al. (2016) – 6 replicates per condition minimum, recommended 12 to capture all DE genes



Schurch et al, 2016

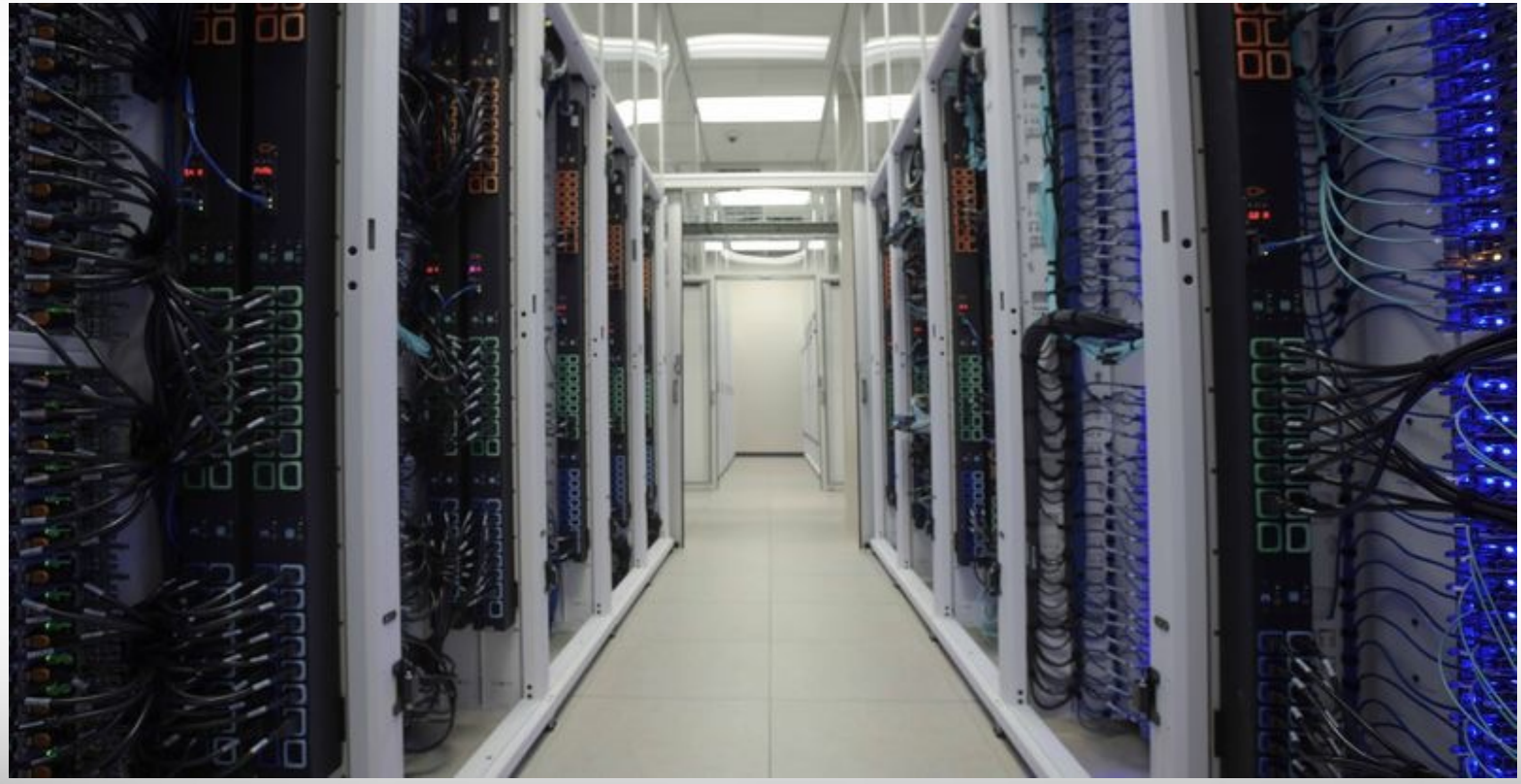
Biological vs Technical Replicates

- Biological replicates - independent samples from different populations/individuals
- Technical replicates - multiple preparations/libraries from the same individual
- Biological replicates generally increase statistical power more than technical replicates
  - Biological variability is generally greater than technical variability
  - Biological replicates contain both biological and technical variability

Differential Expression Analysis Pipeline

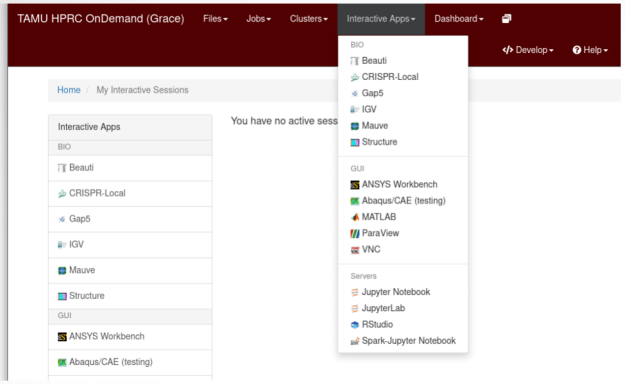
- RNA-seq library QC
- Read trimming
- Mapping reads to reference genome
- Converting and sorting alignment files
- Generating count files
- Differential expression analysis in R

Data analysis on Grace



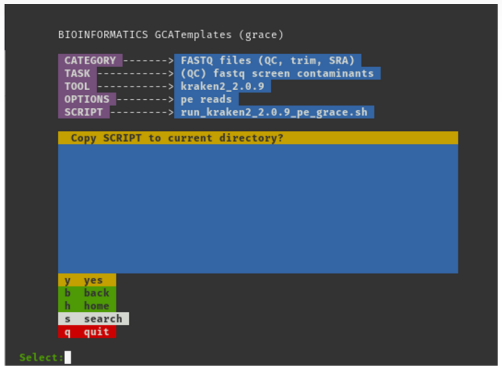
Options for Running Bioinformatics Tools

HPRC Portal



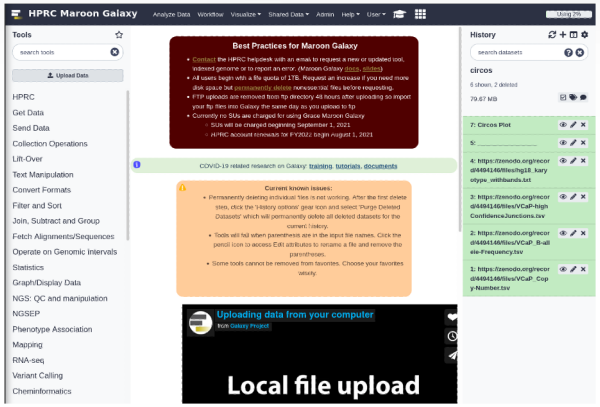
- Access is web browser based
- All HPRC software tools are available either as a GUI or via Unix command line
- Can access Unix command line
  - Grace or Terra
- Best for GUI apps
  - RStudio
  - IGV

Unix command line



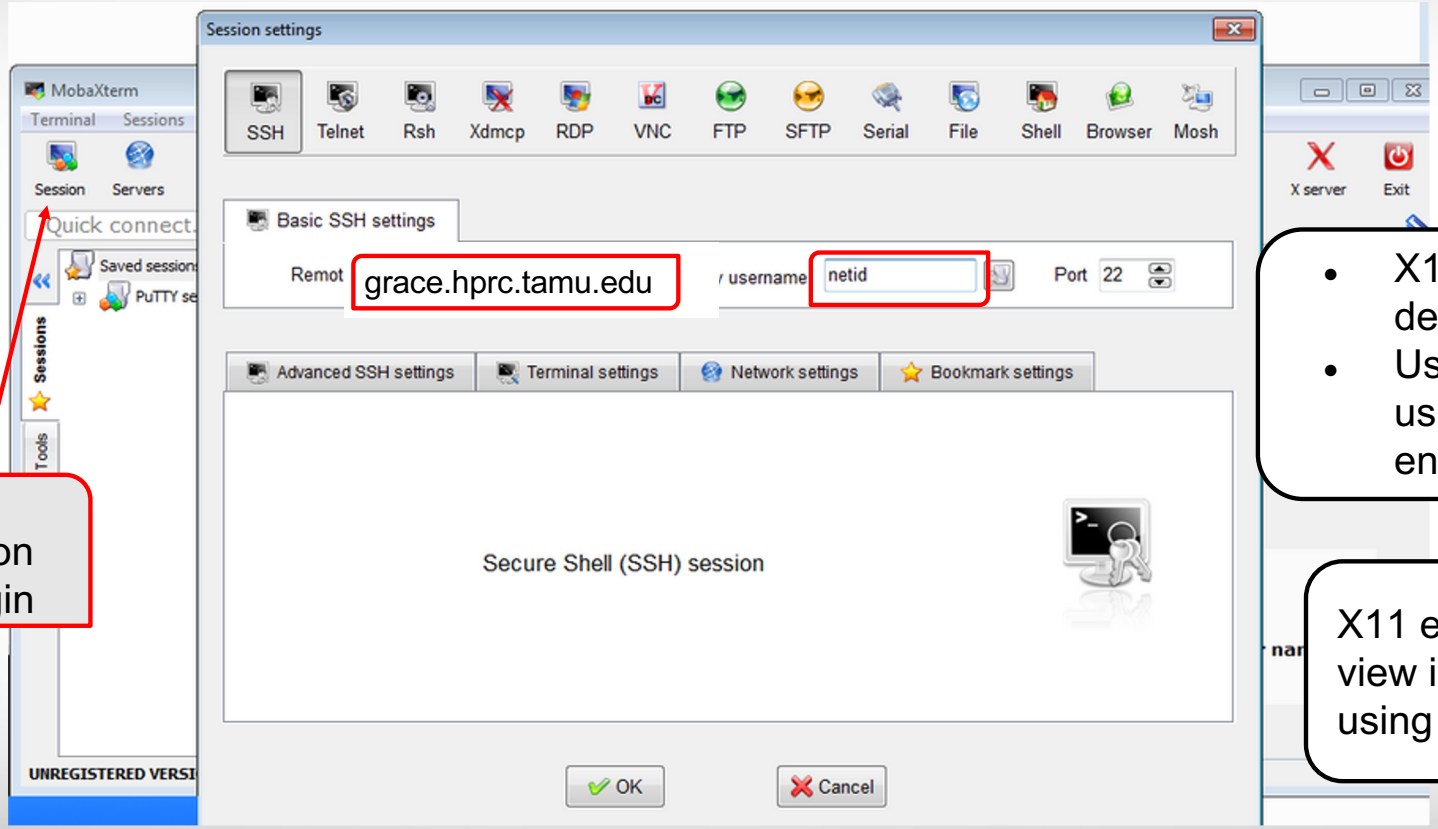
- Need to learn Unix and Slurm
- Bioinformatics template scripts available
- GUI software is not very responsive interactively
- Need SSH client on your Windows computer such as MobaXterm or use HPRC portal

HPRC Maroon Galaxy



- Access is web browser based
- First apply for an HPRC account then request a Galaxy account
- Can request HPRC to add tools from the usegalaxy.org toolshed or create a custom tool

Using SSH - MobaXterm (on Windows) to Connect to Grace



click Session to begin

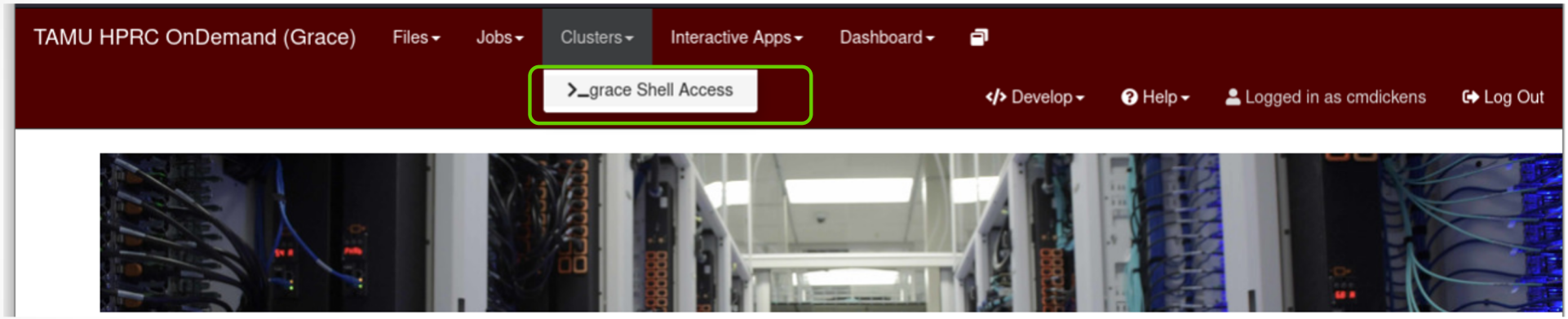
- X11 is enabled by default in MobaXterm
- Use "ssh -X" when using the terminal to enable X11

X11 enables you to view images when using the terminal

https://hprc.tamu.edu/wiki/HPRC:MobaXterm

Connect using the HPRC portal

portal-grace.hprc.tamu.edu



There are no SUs charged for using the Shell Access

Where to Find NGS Tools

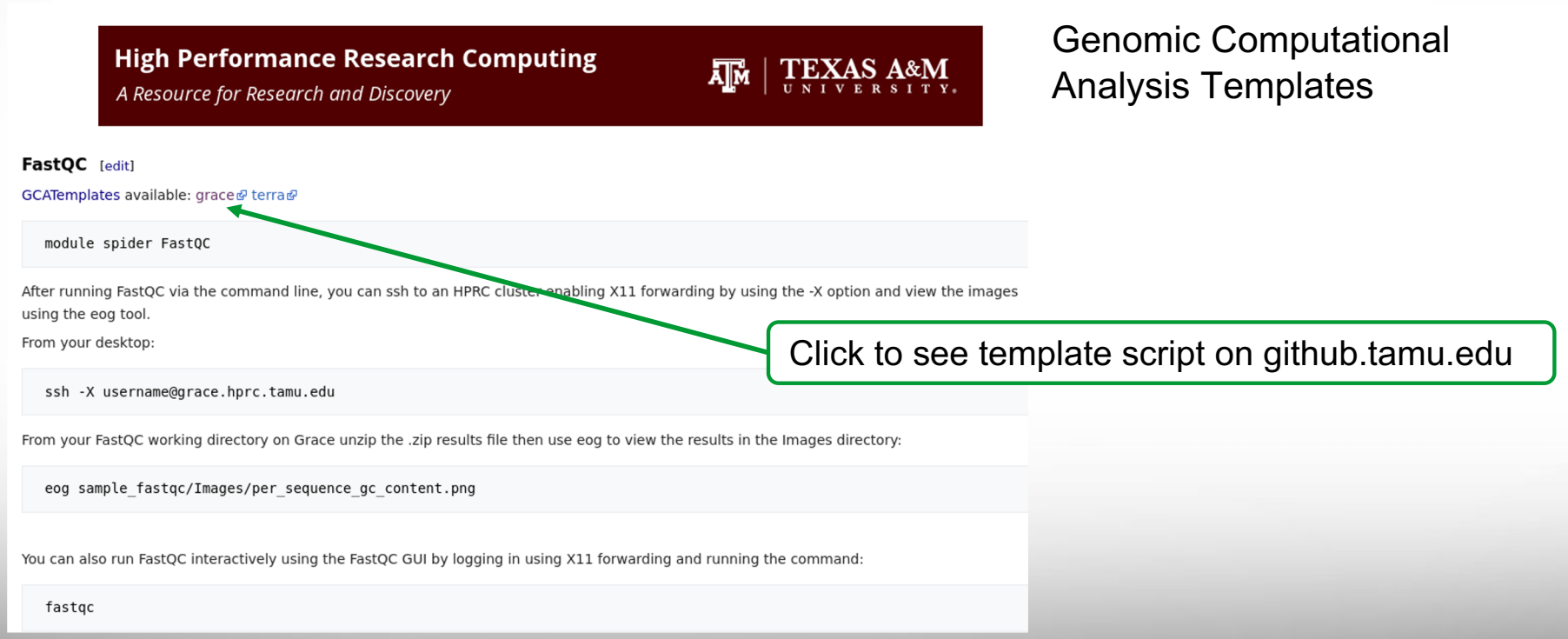
- TAMU HPRC Documentation (https://hprc.tamu.edu/wiki/Bioinformatics)
- Type the following UNIX commands to see which tools are available on Grace
  - module avail
  - module spider *toolname* (not case sensitive, but read entire output)
  - module key RNA (searches tool descriptions)
- If you would like a program installed on Grace, send an email with the URL link to help@hprc.tamu.edu

Template Job Scripts

https://hprc.tamu.edu/wiki/SW:GCATemplates

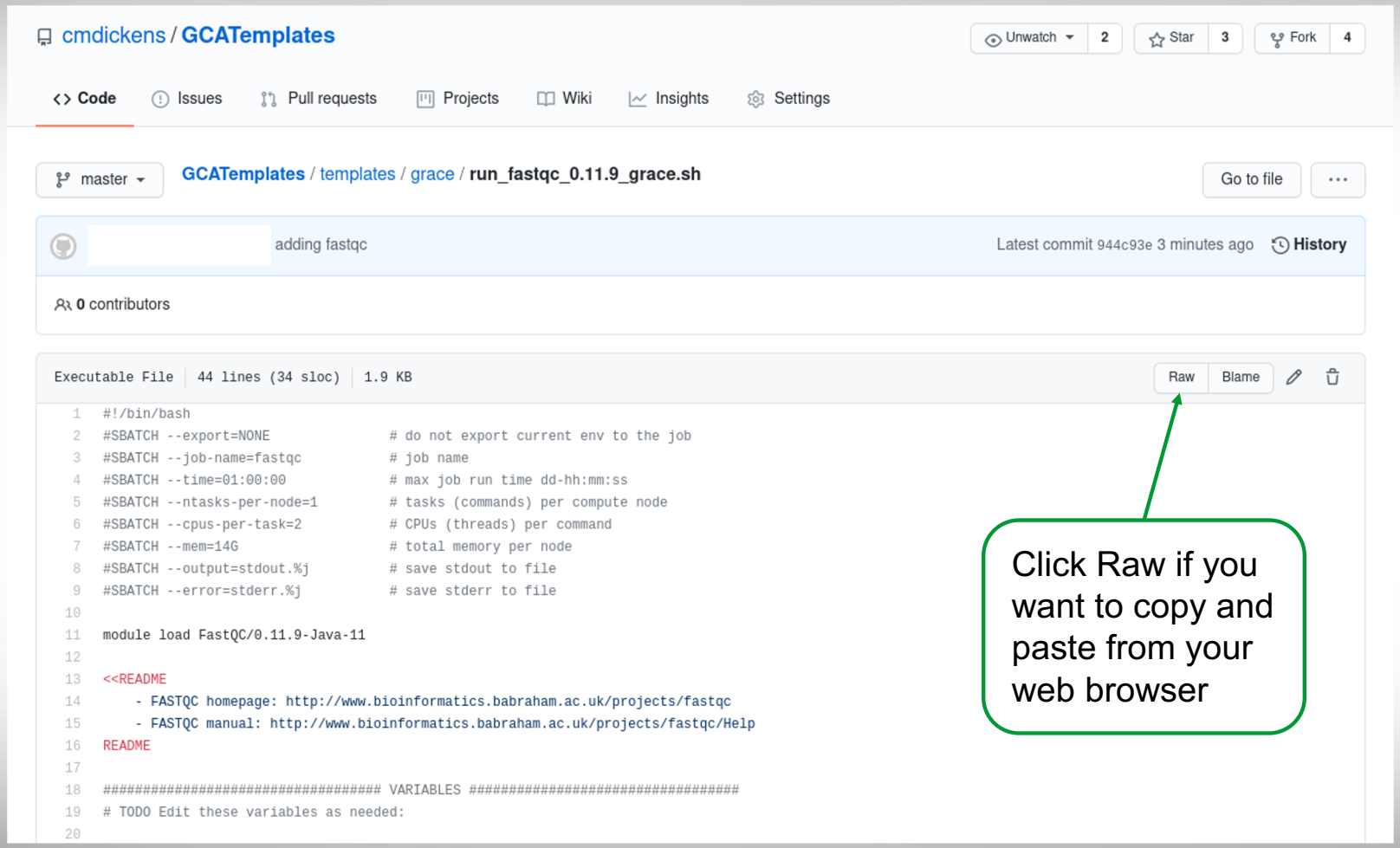
Access GCATemplates Scripts for Grace and Terra from the HPRC wiki

https://hprc.tamu.edu/wiki/Bioinformatics:Sequence_QC#FastQC



Genomic Computational Analysis Templates

Click to see template script on github.tamu.edu


cmdickens / GCATemplates
Unwatch 2
Star 3
Fork 4
Code
Issues
Pull requests
Projects
Wiki
Insights
Settings
master
GCATemplates / templates / grace / run_fastqc_0.11.9_grace.sh
Go to file
adding fastqc
Latest commit 944c93e 3 minutes ago
History
0 contributors
Executable File | 44 lines (34 sloc) | 1.9 KB
Raw
Blame
#!/bin/bash
#SBATCH --export=NONE # do not export current env to the job
#SBATCH --job-name=fastqc # job name
#SBATCH --time=01:00:00 # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1 # tasks (commands) per compute node
#SBATCH --cpus-per-task=2 # CPUs (threads) per command
#SBATCH --mem=14G # total memory per node
#SBATCH --output=stdout.%j # save stdout to file
#SBATCH --error=stderr.%j # save stderr to file
module load FastQC/0.11.9-Java-11
<<README
- FASTQC homepage: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- FASTQC manual: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help
README
################################## VARIABLES ##################################
TODO Edit these variables as needed:
Click Raw if you want to copy and paste from your web browser

Sample GCATemplate Job Script (Grace)

```
#!/bin/bash
#SBATCH --export=NONE                   # do not export current env to the job
#SBATCH --job-name=fastqc               # job name
#SBATCH --time=01:00:00                 # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1             # tasks (commands) per compute node
#SBATCH --cpus-per-task=2               # CPUs (threads) per command
#SBATCH --mem=4G                        # total memory per node
#SBATCH --output=stdout.%j              # save stdout to file (%j is jobid)
#SBATCH --error=stderr.%j               # save stderr to file (%j is jobid)

module load FastQC/0.11.9-Java-11

<<README
          - FASTQC manual: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help
README
################################### VARIABLES ###################################
# TODO Edit these variables as needed:
########## INPUTS ##########
pe1_1='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R1.fastq.gz'
pe1_2='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R2.fastq.gz'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK

########## OUTPUTS #########
output_dir='./'

################################### COMMANDS ####################################
fastqc -t $threads -o $output_dir $pe1_1 $pe1_2

<<CITATION
          - Acknowledge TAMU HPRC: https://hprc.tamu.edu/research/citations.html
          - FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
CITATION
```

Sample GCATemplate Job Script (Grace)

```
#!/bin/bash
#SBATCH --export=NONE                   # do not export current env to the job
#SBATCH --job-name=fastqc               # job name
#SBATCH --time=01:00:00                 # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1             # tasks (commands) per compute node
#SBATCH --cpus-per-task=2               # CPUs (threads) per command
#SBATCH --mem=4G                        # total memory per node
#SBATCH --output=stdout.%j              # save stdout to file (%j is jobid)
#SBATCH --error=stderr.%j               # save stderr to file (%j is jobid)

module load FastQC/0.11.9-Java-11

<<README
        - FASTQC manual: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help
README
################################### VARIABLES ###################################
# TODO Edit these variables as needed:
########## INPUTS ##########
pe1_1='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R1.fastq.gz'
pe1_2='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R2.fastq.gz'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK

########## OUTPUTS #########
output_dir='./'

################################### COMMANDS ####################################
fastqc -t $threads -o $output_dir $pe1_1 $pe1_2

<<CITATION
        - Acknowledge TAMU HPRC: https://hprc.tamu.edu/research/citations.html
        - FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
CITATION
```

These parameters are read by the job scheduler

Load the required module(s) first

This is a section of comments

This is a single line comment and not run as part of the script

This is the command to run the application

Use **$TMPDIR** whenever possible

- Use the **$TMPDIR** if the application you are running can utilize a temporary directory for writing temporary files which are deleted when the job ends
- A temp directory (**$TMPDIR**) is automatically assigned for each job which uses the disk(s) on the compute node not the **$SCRATCH** shared file system
  - Especially useful when a computational tool writes tens of thousands of temporary files which are deleted when the job is finished and are not needed for the final results
  - This is useful since files on **$TMPDIR** will not count against your file quota
  - Don't use **$TMPDIR** if your software uses temporary files for restarting where it left off if it should stop before completion
  - Will significantly speed up an mpiBLAST job

```
java -Xmx350g -jar $EBROOTPICARD/FastqToSam.jar TMP_DIR=$TMPDIR \
FASTQ=$pe1_1 FASTQ2=$pe1_2 OUTPUT=$outfile SAMPLE_NAME=$sample_name \
SORT_ORDER=$sort_order MAX_RECORDS_IN_RAM='null'
```

Example Data

```
ssh -X username@grace.hprc.tamu.edu
```

```
mkdir $SCRATCH/RNA_class
```

```
cd $SCRATCH/RNA_class
```

```
cp -r /scratch/training/bio/rna-seq/* .
```

```
ls
```

Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 355, NO. 6326 > VITAMIN B_3 MODULATES MITOCHONDRIAL VULNERABILITY AND PREVENTS GLAUCOMA IN AGED MICE

REPORT

Vitamin B_3 modulates mitochondrial vulnerability and prevents glaucoma in aged mice

PETE A. WILLIAMS, JEFFREY M. HARDER, NICOLE E. FOXWORTH, KELLY E. COCHRAN, VIVEK M. PHILIP, VITTORIO PORCIATTI, OLIVER SMITHIES, AND SIMON W. M. JOHN

Authors Info & Affiliations

SCIENCE · 17 Feb 2017 · Vol 355, Issue 6326 · pp. 756-760 · DOI: 10.1126/science.aal0092

77 198

Vitamin B_3 protects mice from glaucoma

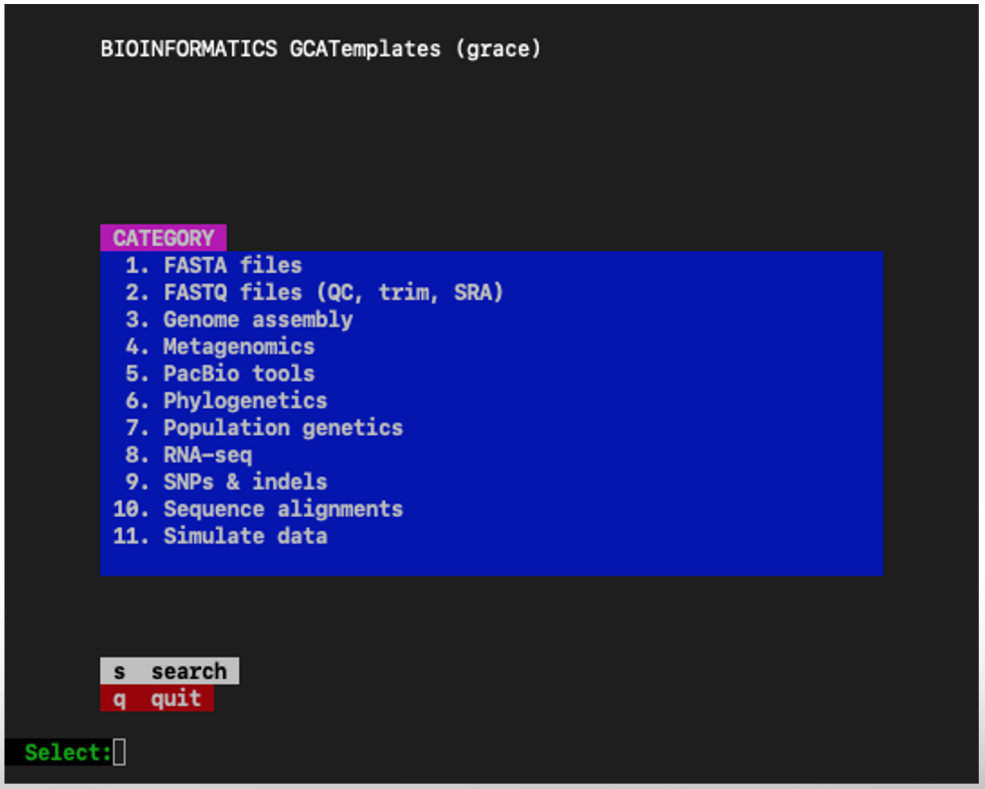
Glaucoma is the most common cause of age-related blindness in the United States. There is currently no cure, and once vision is lost, the condition is irreversible. Williams *et al.* now report that vitamin B_3 (also known as niacin) prevents eye degeneration in glaucoma-prone mice (see the Perspective by Crowston and Trounce). Supplementing the diets of young mice with vitamin B_3 averted early signs of glaucoma. Vitamin B_3 also halted further glaucoma development in aged mice that already showed signs of the disease. Thus, healthy intake of vitamin B_3 may protect eyesight.

Science, this issue p. 756; see also p. 688

Differential Expression Analysis on Grace

Quality Control

- Run FastQC to assess RNA-seq library quality
- Use GCATemplates to create job script
- Select #2 then find the template that contains fastqc or use the search to find fastqc
- Final step will save a template job script file to your current working directory



Differential Expression Analysis on Grace

Quality Control

- Edit the job script to change the path to the fastq files

```
gedit run_fastqc_0.11.9_grace.sh
```

```
################################### VARIABLES ###################################
# TODO Edit these variables as needed:

########## INPUTS ##########
pe1_1='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R1.fastq.gz'
pe1_2='/scratch/data/bio/GCATemplates/data/miseq/c_dubliniensis/DR34_R2.fastq.gz'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK

########## OUTPUTS #########
output_dir='./'
```

Differential Expression Analysis on Grace

Quality Control

- Edit the job script to change the path to the fastq files

```
################################### VARIABLES ###################################
# TODO Edit these variables as needed:

########## INPUTS ##########
pe1_1='/scratch/user/username/RNA_class/Control1_R1.fastq.gz'
pe1_2='/scratch/user/username/RNA_class/Control1_R2.fastq.gz'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK

########## OUTPUTS #########
output_dir='./'
```

```
sbatch run_fastqc_0.11.9_grace.sh
```

Differential Expression Analysis on Grace

Quality Control

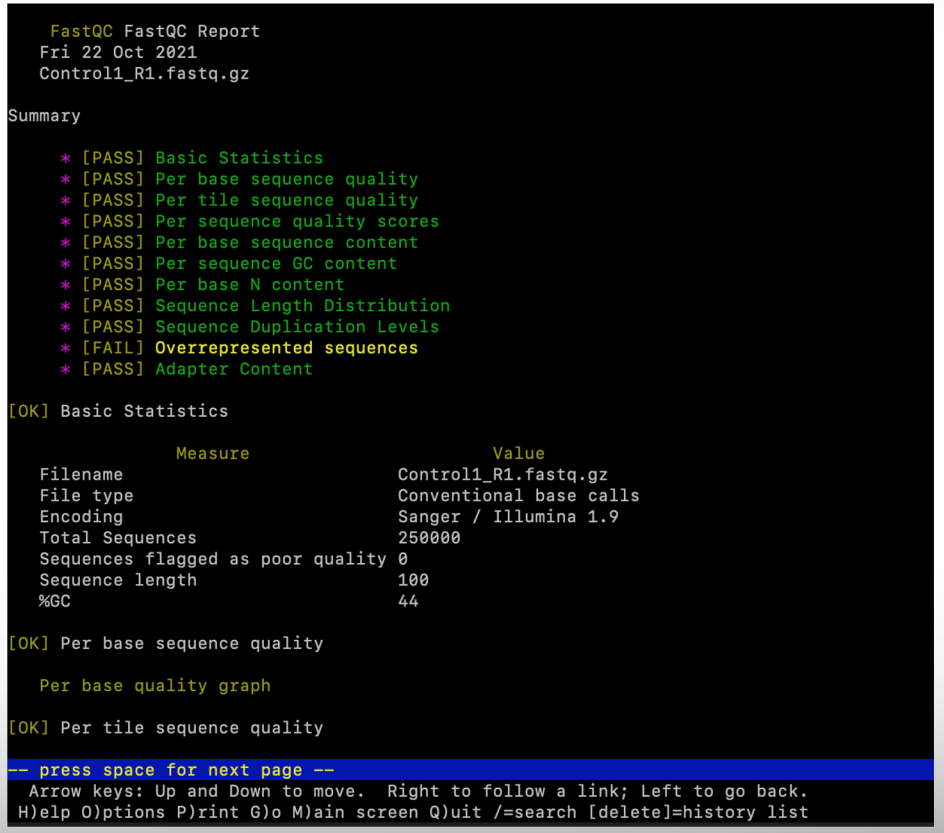
- Unzip the results file and you can view the results with lynx and eog (eog requires X11 login; if using the portal, use the Files app to view the images)

```
unzip Control1_R1_fastqc.zip
```

Differential Expression Analysis on Grace

Quality Control

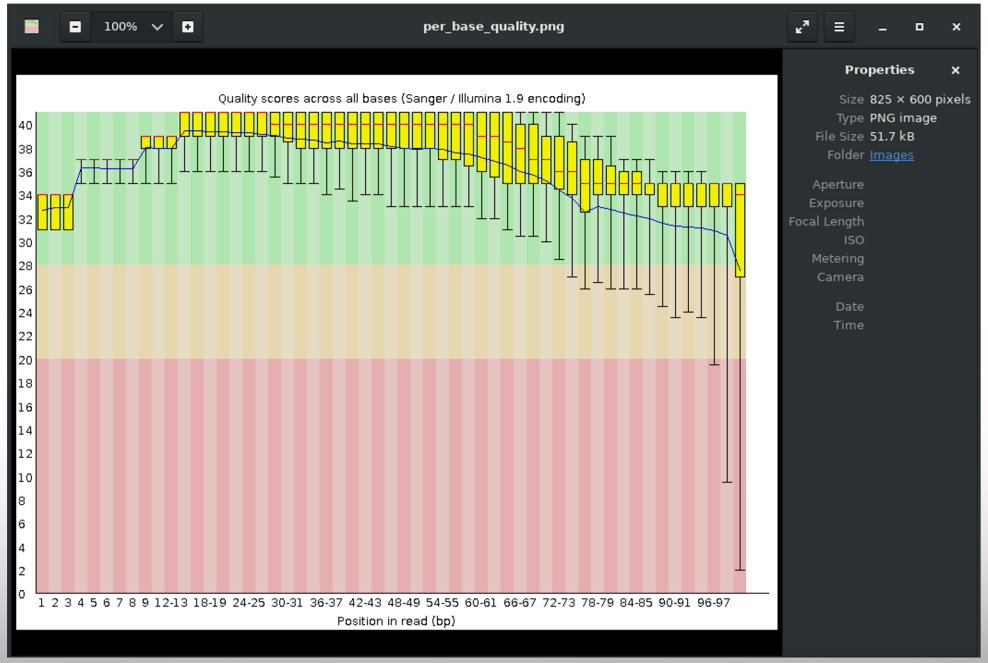
```
lynx Control1_R1_fastqc.html
```



Differential Expression Analysis on Grace

Quality Control

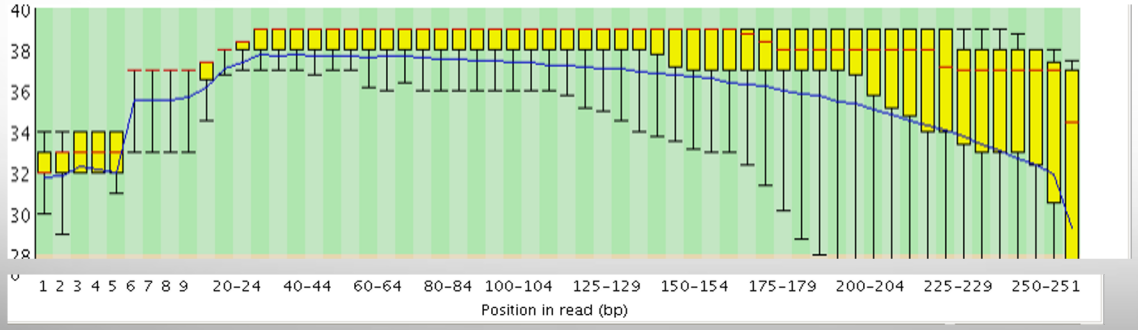
```
eog Control1_R1_fastqc/Images/per_base_quality.png
```



FastQC Output Image Quality Distribution

FASTQ format

```
@ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAACAAAAAACATAATGCCGTAAA
+ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
=:=4AD=B8A:+<A::1<:AE<C3*?F<B???<?:8:6?B*9BD;/638.=-'-.@7=).=A:6?DDDCBBBB9555&&)+((+2&&+(((((((&&&+
@ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
GATGTTTTGTTACTGATTGGAACCATGATTGGTGCTTTACTTGGTTTCTTCCTATTTAACCACAAGCCTGCCAAAGTATTTATGGGAGATGTAGGTAGTT
+ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
BCCFDEFFHHHHHJJJJIJJJJJJJJJIJJJFHIJJJIJJJJJIJJJJJJJIJJJIJIIJJJJJFHJJJJIJJHH=CHHFFFFFFEDDEDEEEDCCDCE
@ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
TATTTTAAGTGACCAAGGAATGACTCCCCAATCATGGCTGTATCAACTCCAAAATTTTCTGCAACAGTCGCTGAAATATCTGCAAAATGCCCTTGTGGAA
+ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJHIJHHGFFFFFFFEEEEEEDDDDDDDDDCDDDD
```

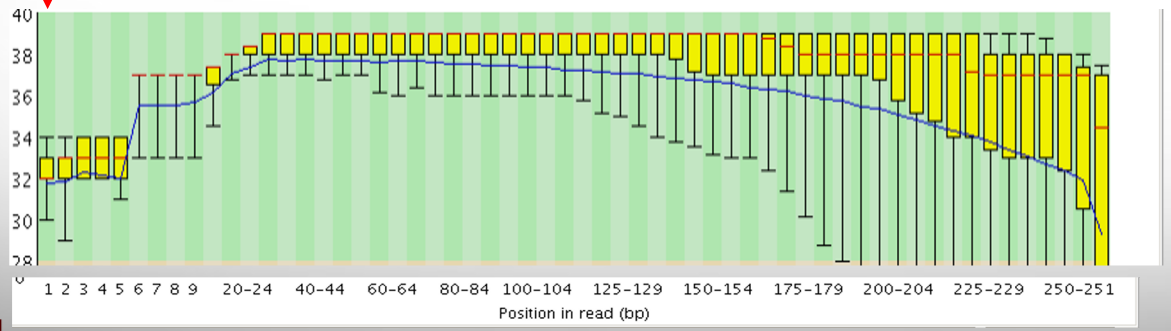


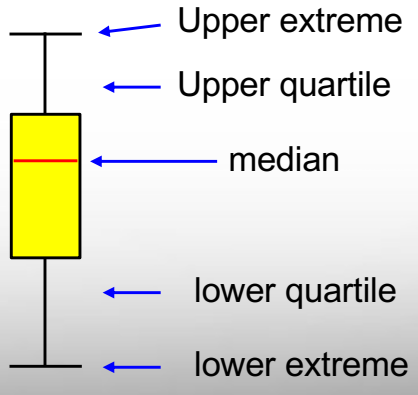
FastQC Output Image Quality Distribution

FASTQ format

```
@ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAACAAAAAACATAATGCCGTAAA
+ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
=:=4AD=B8A:+<A::1<:AE<C3*?F<B???<?:8:6?B*9BD;/638.=-'-.@7=).=A:6?DDDCBBBB9555&&)+((+2&&+(((((((&&&+
@ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
GATGTTTTGTTACTGATTGGAACCATGATTGGTGCTTTACTTGGTTTCTTCCTATTTAACCACAAGCCTGCCAAAGTATTTATGGGAGATGTAGGTAGTT
+ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
BCCFDEFFHHHHHJJJJIJJJJJJJJJIJJJFHIJJJIJJJJJIJJJJJJJIJJJIJIIJJJJJFHJJJJIJJHH=CHHFFFFFFEDDEDEEEDCCDCE
@ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
TATTTTAAGTGACCAAGGAATGACTCCCCAATCATGGCTGTATCAACTCCAAAATTTTCTGCAACAGTCGCTGAAATATCTGCAAAATGCCCTTGTGGAA
+ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJHIJHHGFFFFFFFEEEEEEDDDDDDDDDCDDDD
```

Average quality score distribution at position one



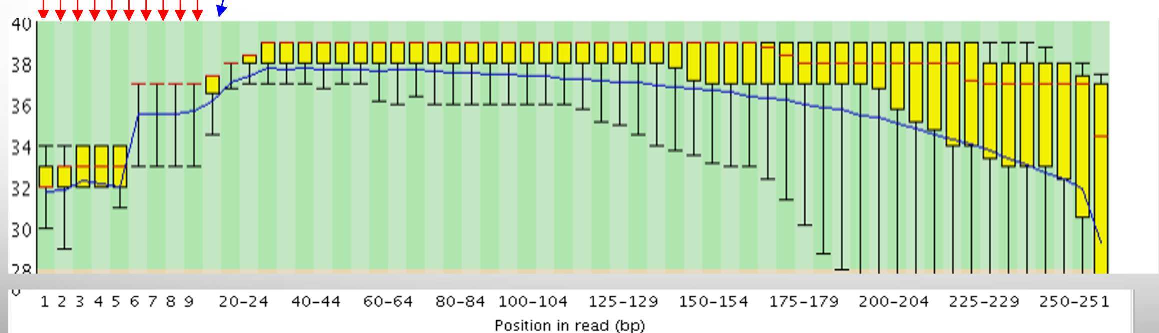


FastQC Output Image Quality Distribution

FASTQ format

```
@ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAACAAAAAACATAATGCCGTAAA
+ERR504787.2.1 M00368:15:000000000-A0HKH:1:5:21261:10968-1 length=100
=:=4AD=B8A:+<A::1<:AE<C3*?F<B???<?:8:6?B*9BD;/638.=-'-.@7=).=A:6?DDDCBBBB9555&&)+((+2&&+((((((()&&&+
@ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
GATGTTTTGTTACTGATTGGAACCATGATTGGTGCTTTACTTGGTTTCTTCCTATTTAACCACAAGCCTGCCAAAGTATTTATGGGAGATGTAGGTAGTT
+ERR504787.3.1 M00368:15:000000000-A0HKH:1:3:12724:25677-1 length=100
BCCFDEFFHHHHHJJJJIJJJJJJJJJJIJJJFHIJJJIJJJJJIJJJJJJJIJJJIJIIJJJJJFHJJJJIJJHH=CHHFFFFFFFEDDEDEEEDCCDCE
@ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
TATTTTAAGTGACCAAGGAATGACTCCCCAATCATGGCTGTATCAACTCCAAAATTTTCTGCAACAGTCGCTGAAATATCTGCAAAATGCCCTTGTGGAA
+ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJHIJHHGFFFFFFFEEEEEEDDDDDDDDDCDDDD
```

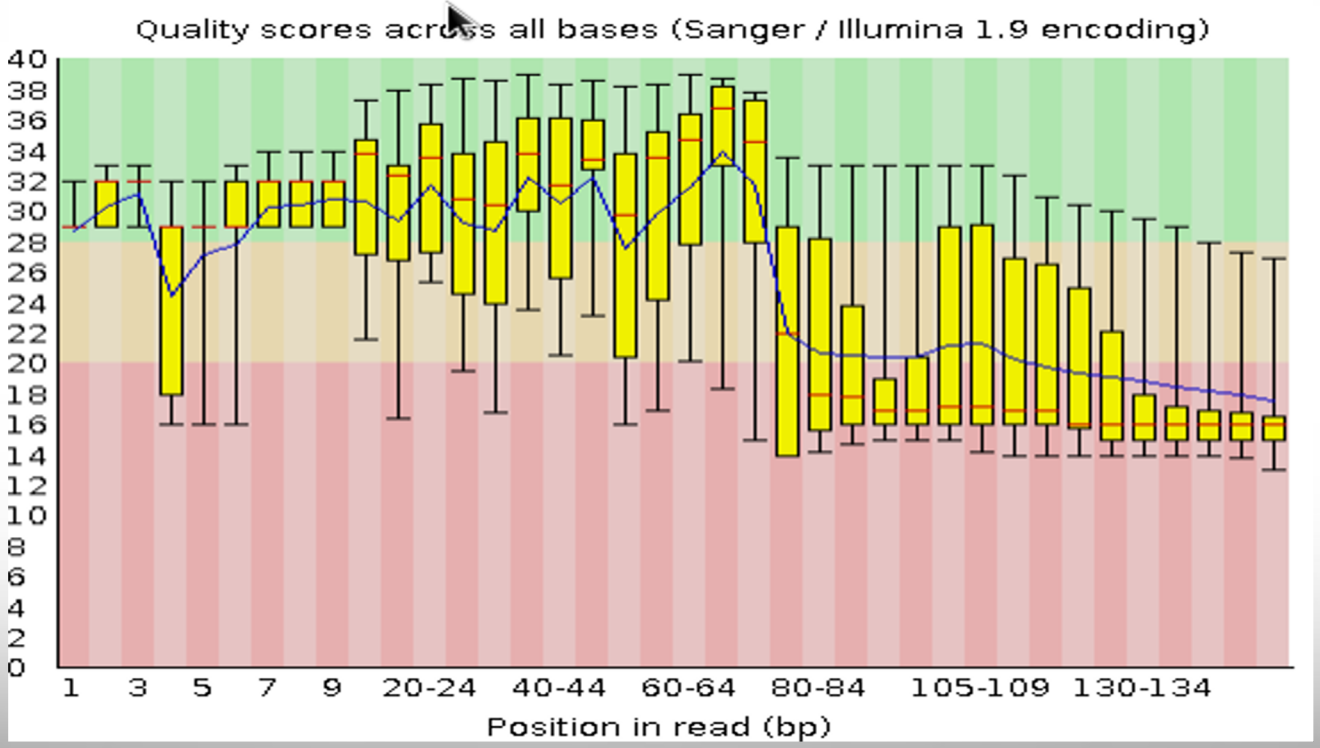
Positions are ‘binned’ after the first few positions



Failed QC Examples

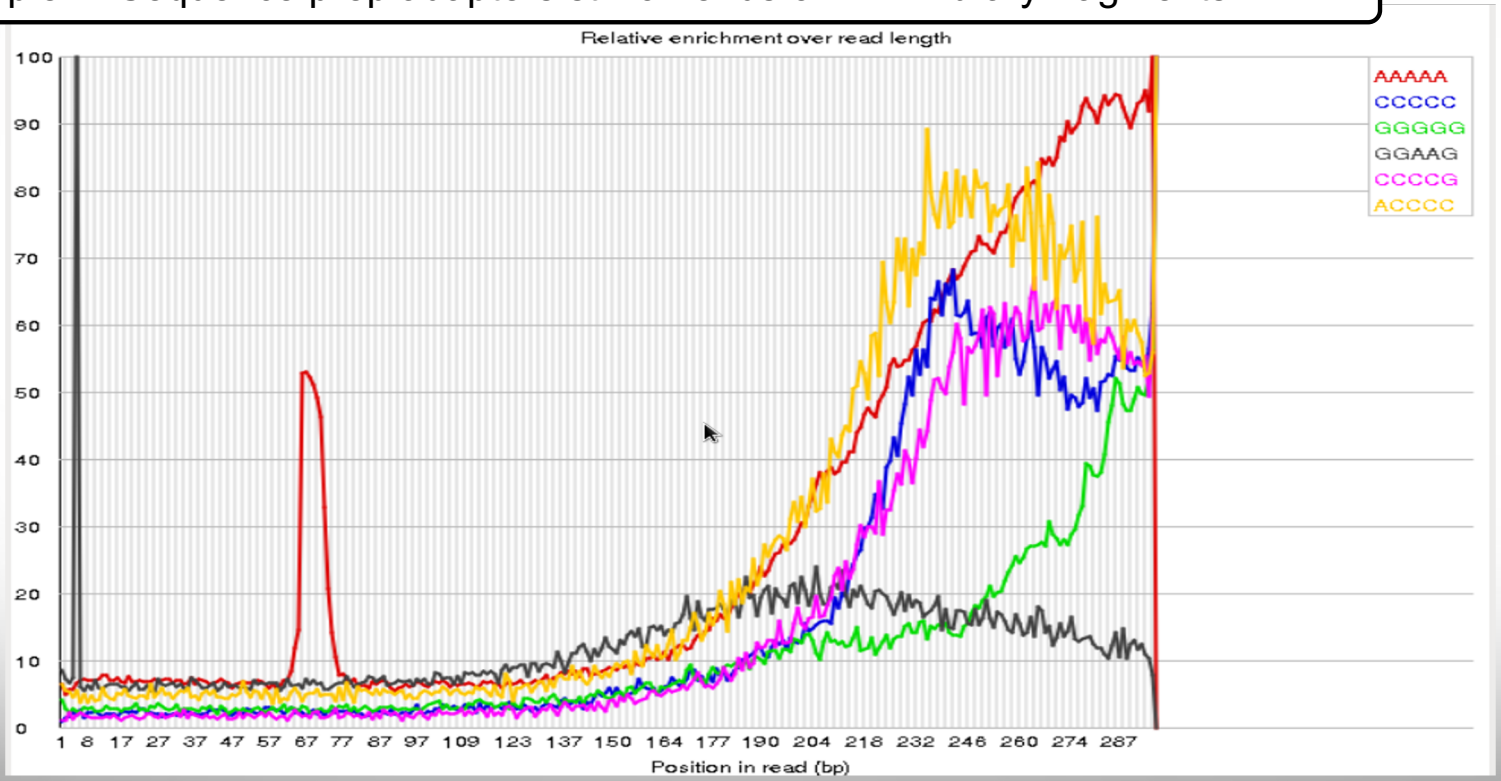
FastQC Output Image
Failed Per base sequence quality

Example 1. Expired MiSeq mate-pair kit (9 months expired)



FastQC Output Image
Failed Kmer Content

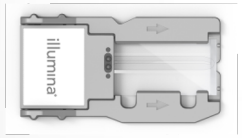
Example 2. Sequence prep adapters still on ends of DNA library fragments

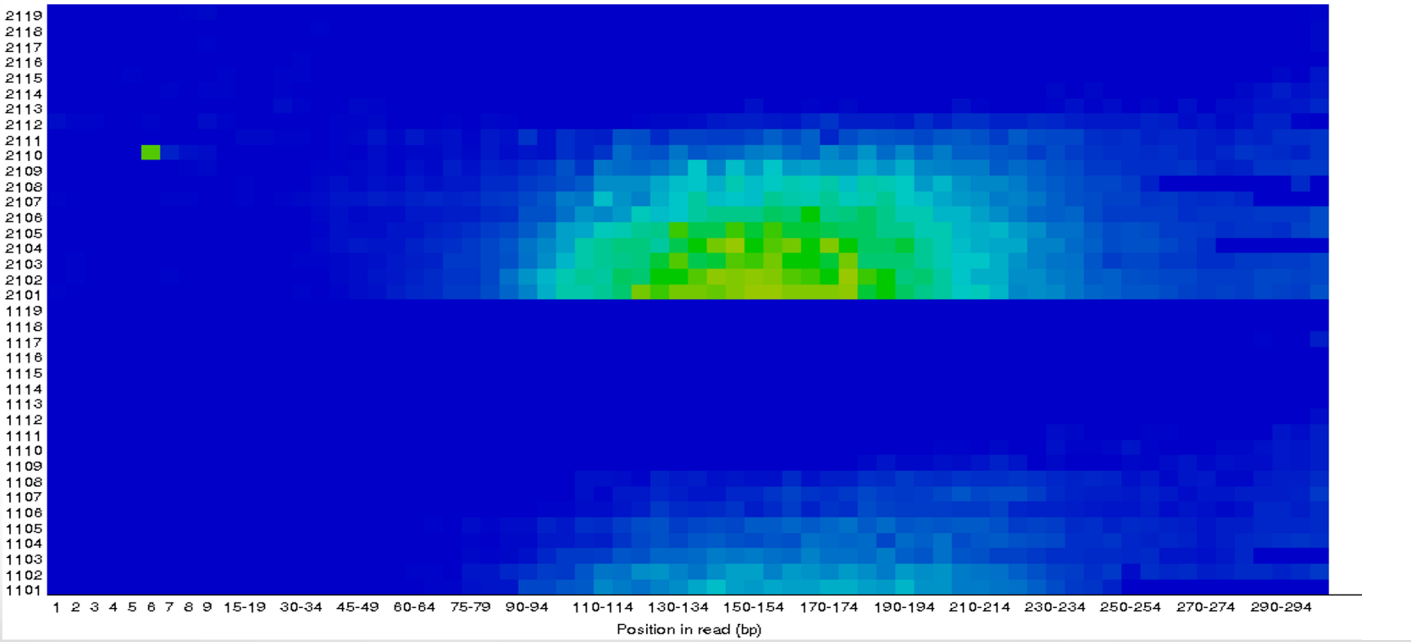


FastQC Output Image
Flowcell: not good per_tile quality

Example 3. Faulty flowcell

MiSeq flowcell

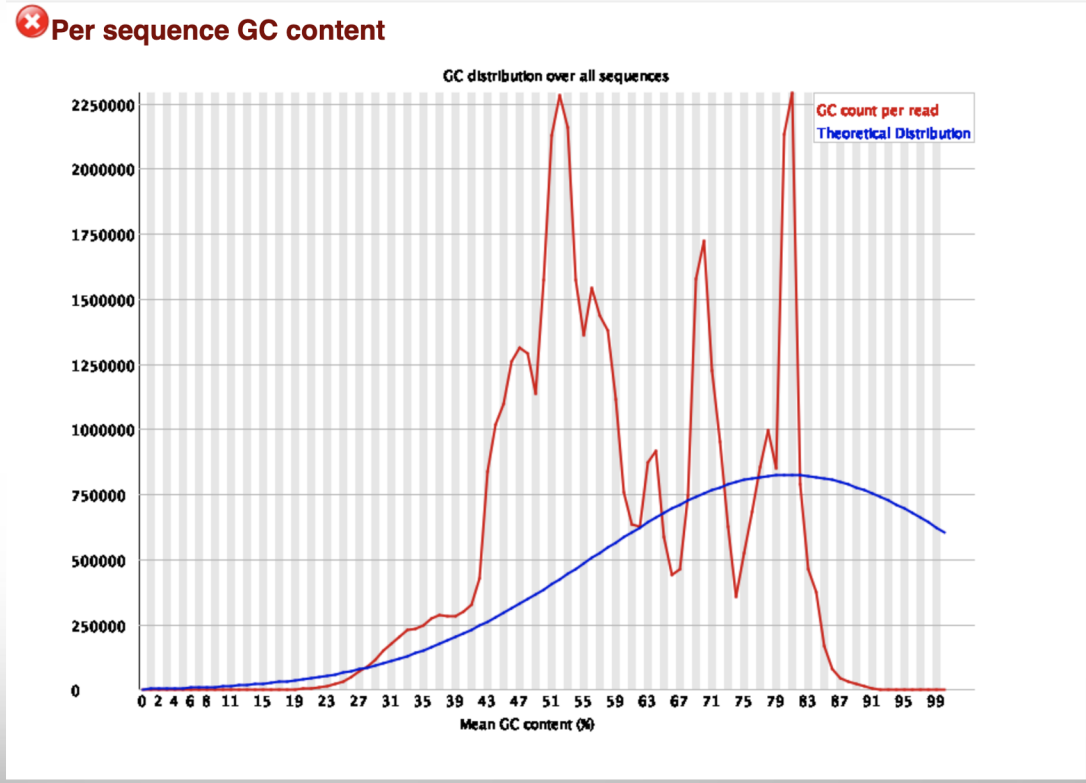




good quality poor quality

FastQC Output Image
Failed GC content

Example 4. Contamination



Differential Expression Analysis on Grace

Read Trimming

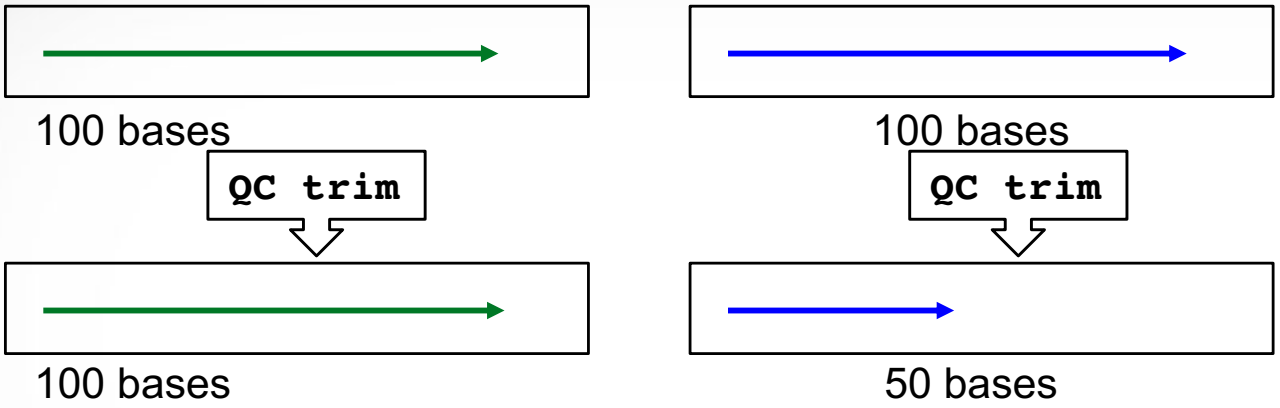
- Find and load the required modules

```
module purge
module spider trim_galore
module spider Trim_Galore/0.6.6-Python-3.8.2
module load GCCcore/9.3.0
module load Trim_Galore/0.6.6-Python-3.8.2
trim_galore --paired Control1_R1.fastq.gz Control1_R2.fastq.gz
```

Trimming PE Short Sequence Reads

File 1 from sequencer

File 2 from sequencer



minimum read length = 40

Resulting FASTQ Files with trimmed reads

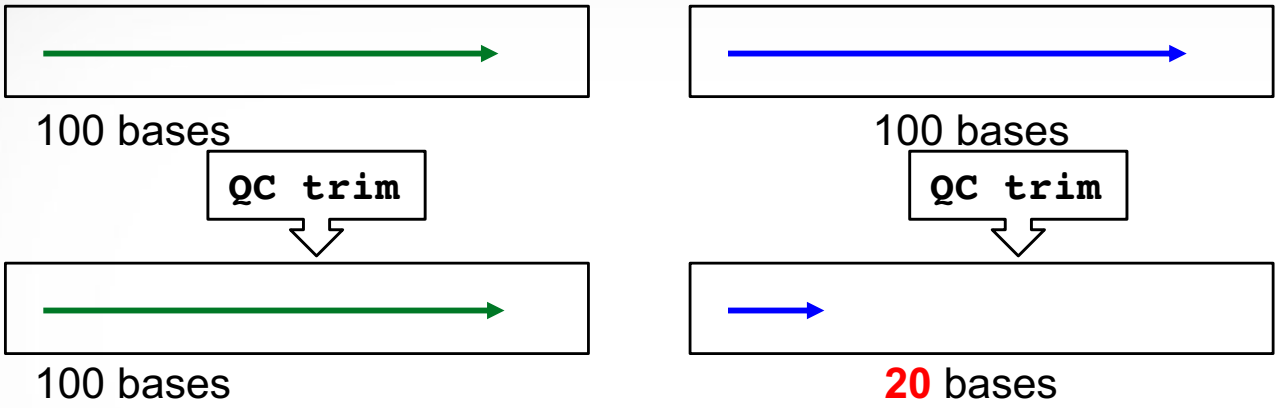
Paired end 1 trimmed file

Paired end 2 trimmed file

Trimming PE Short Sequence Reads

File 1 from sequencer

File 2 from sequencer



minimun read length = 40

Resulting FASTQ Files with trimmed reads

Paired end 1 trimmed file reads

Paired end 2 trimmed file

Single end

Differential Expression Analysis on Grace

Read Trimming

- Check the results with FastQC

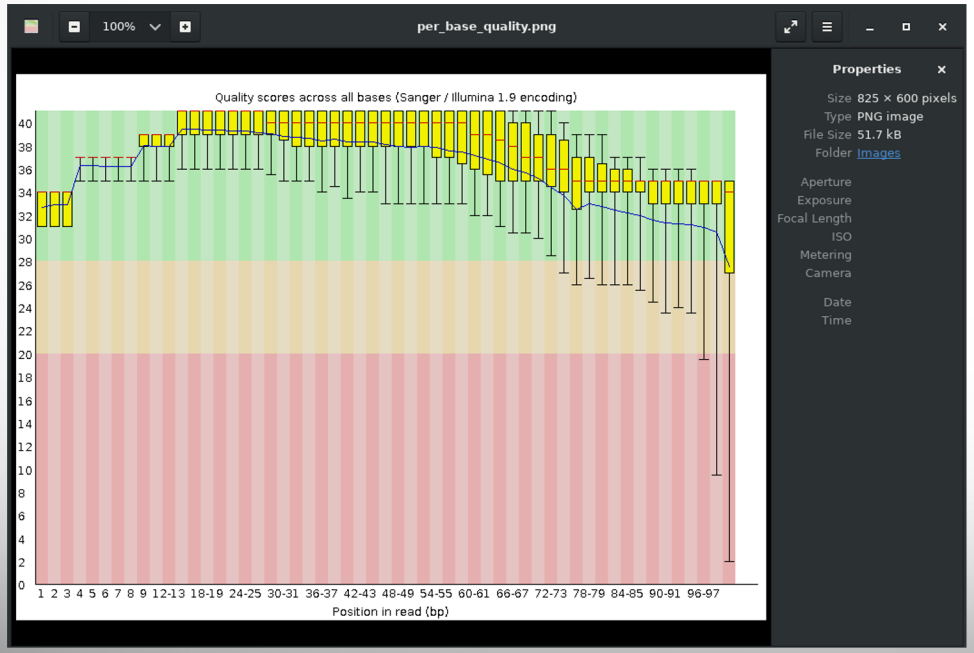
```
module load FastQC/0.11.9-Java-11
fastqc Control1_R1_val_1.fq.gz
unzip Control1_R1_val_1_fastqc.zip
eog Control1_R1_val_1_fastqc/Images/per_base_quality.png
```

Differential Expression Analysis on Grace

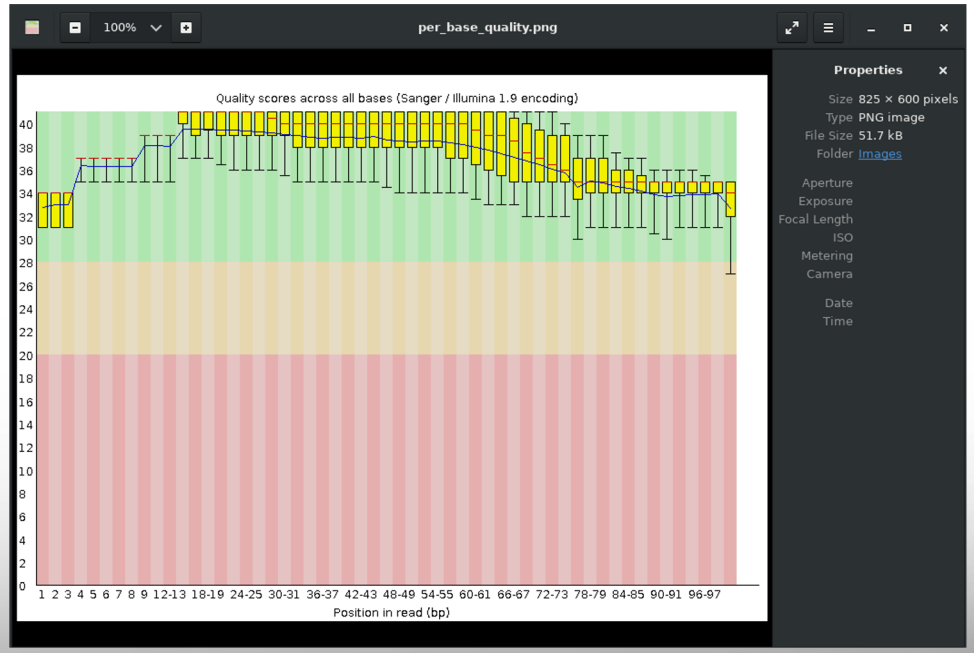
Read Trimming

- Check the results with FastQC

Before



After



Differential Expression Analysis on Grace

Read Mapping

- Popular splice-aware aligners
  - STAR
  - HISAT2
- Both programs need to index genome before aligning reads
  - Only needs to be done once
  - HISAT2 faster and more memory efficient
  - Some genomes already indexed on Grace

```
/scratch/data/bio/genome_indexes/
```

Send an email to help@hprc.tamu.edu if you need a genome indexed that is not found in the genome_indexes directory

Differential Expression Analysis on Grace

Read Mapping

- We'll use HISAT2 to align our library to the mouse reference genome

```
gcatemplates
```

  - Type 10 to select "Sequence alignments"
  - Type 2 to select "align mRNA reads to a reference"
  - Type 1 to select "hisat2_2.2.1"
  - Type 1 to select "pe library"
  - Type y to copy the SCRIPT to your current directory

```
gedit run_hisat2_2.2.1_pe_grace.sh
```

Differential Expression Analysis on Grace

Read Mapping

- Modify the script

Change path to our trimmed reads

```
######### SYNOPSIS #########
# This template script aligns paired end reads and sorts the output into a bam file

################################### VARIABLES ###################################
# TODO Edit these variables as needed:

########## INPUTS ##########
pe_1='/scratch/data/bio/GCATemplates/miseq/a_fumigatus/DRR022927_1.fastq.gz'
pe_2='/scratch/data/bio/GCATemplates/miseq/a_fumigatus/DRR022927_2.fastq.gz'

# you can use an already prefixed genome found at: /scratch/data/bio/genome_indexes/
genome_index_prefix='/scratch/data/bio/genome_indexes/gmod_genomes/Aspergillus_fumigatus_Af293/hisat2/
A_fumigatus_Af293'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK
# read group information
id='af_amp'
library='sra'
platform='ILLUMINA'
sample='DRR022927'

########## OUTPUTS #########
output_bam="${sample}_pe_aln.bam"
```

Differential Expression Analysis on Grace

Read Mapping

- Modify the script

```
######### SYNOPSIS #########
# This template script aligns paired end reads and sorts the output into a bam file

################################### VARIABLES ###################################
# TODO Edit these variables as needed:

########## INPUTS ##########
pe_1='/scratch/user/username/RNA_class/Control1_R1_val_1.fq.gz'
pe_2='/scratch/user/username/RNA_class/Control1_R2_val_2.fq.gz'

# you can use an already prefixed genome found at: /scratch/data/bio/genome_indexes/
genome_index_prefix='/scratch/data/bio/genome_indexes/gmod_genomes/Aspergillus_fumigatus_Af293/hisat2/
A_fumigatus_Af293'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK
# read group information
id='af_amp'
library='sra'
platform='ILLUMINA'
sample='DRR022927'

########## OUTPUTS #########
output_bam="${sample}_pe_aln.bam"
```

Change path to the indexed reference sequence

/scratch/data/bio/genome_indexes/ncbi/mm39/hisat2/GCF_000001635.27_GRCm39_genomic

Differential Expression Analysis on Grace

Read Mapping

- Modify the script

```
######### SYNOPSIS #########
# This template script aligns paired end reads and sorts the output into a bam file

################################### VARIABLES ###################################
# TODO Edit these variables as needed:

########## INPUTS ##########
pe_1='/scratch/user/username/RNA_class/Control1_R1_val_1.fq.gz'
pe_2='/scratch/user/username/RNA_class/Control1_R2_val_2.fq.gz'

# you can use an already prefixed genome found at: /scratch/data/bio/genome_indexes/
genome_index_prefix='/scratch/data/bio/genome_indexes/ncbi/mm39/hisat2/GCF_000001635.27_GRCm39_genomic'

######## PARAMETERS ########
threads=$SLURM_CPUS_PER_TASK
# read group information
id='af_amp'
library='sra'
platform='ILLUMINA'
sample='DRR022927'

########## OUTPUTS #########
output_bam="${sample}_pe_aln.bam"
```

Change the read group information

- ID = ‘SRR5061328’
- library = ‘sra’
- platform = ‘ILLUMINA’
- sample = ‘Control1’

Differential Expression Analysis on Grace

Read Mapping

- Now submit the job and examine the output

```
sbatch run_hisat2_2.2.1_pe_grace.sh
```

```
more stderr.jobid
```

```
236499 reads; of these:
  236499 (100.00%) were paired; of these:
    32979 (13.94%) aligned concordantly 0 times
    194913 (82.42%) aligned concordantly exactly 1 time
    8607 (3.64%) aligned concordantly >1 times
    ----
    32979 pairs aligned concordantly 0 times; of these:
      3583 (10.86%) aligned discordantly 1 time
    ----
    29396 pairs aligned 0 times concordantly or discordantly; of these:
      58792 mates make up the pairs; of these:
        33529 (57.03%) aligned 0 times
        22657 (38.54%) aligned exactly 1 time
        2606 (4.43%) aligned >1 times
92.91% overall alignment rate
[bam_sort_core] merging from 0 files and 48 in-memory blocks...
```

Differential Expression Analysis on Grace

Generating Count Files

- Several options are available
  - Salmon
  - Sailfish
  - RSEM
  - htseq-count
  - summarizeOverlaps (from R package GenomicsAlignments)

```
module load GCC/10.2.0 OpenMPI/4.0.5 HTSeq/0.11.3 SAMtools/1.11
```

```
samtools index Control1_pe_aln.bam
```

```
htseq-count -f bam -r pos -i gene Control1_pe_aln.bam GCF_000001635.27_GRCm39_genomic.gff > Control1_counts.txt
```

Differential Expression Analysis on Grace

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

05/19/2021

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, and mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.32.0

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#htseq-count-input

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Open RStudio through the Grace portal or on your laptop
- Set your working directory

```
setwd("/scratch/user/username/RNA_class/counts")
```

- You might need to install some packages

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("EnhancedVolcano")
```

https://bioconductor.org/packages/release/bioc/html/EnhancedVolcano.html

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Load the necessary libraries

```
library("DESeq2")
library("ggplot2")
library("EnhancedVolcano")
library("pheatmap")
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Load the count and sample information

```
sampleTable <- read.csv("sampleTable.csv", header = TRUE)
sampleTable <- as.data.frame(sampleTable)
sampleTable$condition <- factor(sampleTable$condition)
sampleTable
```

	sampleName	fileName	condition
1	Control1_counts.txt	Control1_counts.txt	Control
2	Control2_counts.txt	Control2_counts.txt	Control
3	Control3_counts.txt	Control3_counts.txt	Control
4	Control4_counts.txt	Control4_counts.txt	Control
5	Control5_counts.txt	Control5_counts.txt	Control
6	NAD1_counts.txt	NAD1_counts.txt	NAD_supplement
7	NAD2_counts.txt	NAD2_counts.txt	NAD_supplement
8	NAD3_counts.txt	NAD3_counts.txt	NAD_supplement
9	NAD4_counts.txt	NAD4_counts.txt	NAD_supplement
10	NAD5_counts.txt	NAD5_counts.txt	NAD_supplement

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Build the DESeqDataSet

```
dds <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,
                                  directory = ".",
                                  design= ~ condition)
dds
```

```
> dds
class: DESeqDataSet
dim: 46316 10
metadata(1): version
assays(1): counts
rownames(46316): 0610005C13Rik 0610006L08Rik ... n-TYgta9 n-Tcgca44
rowData names(0):
colnames(10): Control1_counts.txt Control2_counts.txt ... NAD4_counts.txt NAD5_counts.txt
colData names(1): condition
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Filter out genes with less than 10 total reads

```
keep <- rowSums(counts(dds)) >= 10
dds <- dds[keep,]
```

- Run the differential expression analysis

```
dds <- DESeq(dds)
res <- results(dds)
res
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526   1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365 -0.46059738889209 0.272726760296267    -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153   1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                  baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                 <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526   1.0536757176372  0.803957398183298    0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365 -0.46059738889209 0.272726760296267    -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153   1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

mean of normalized counts for all samples

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526  1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365  -0.46059738889209 0.272726760296267    -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153  1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

log2 fold change: NAD supplement vs Control

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526  1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365  -0.46059738889209 0.272726760296267   -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153  1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

standard error: NAD supplement vs Control

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat             pvalue              padj
                    <numeric>          <numeric>         <numeric>           <numeric>          <numeric>         <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526   1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708               NA
0610009B22Rik  229.572854136365 -0.46059738889209 0.272726760296267   -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153   1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

Wald statistic: NAD supplement vs Control

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526   1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365 -0.46059738889209 0.272726760296267    -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158   -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153   1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

Wald test p value: NAD supplement vs Control

Differential Expression Analysis on Grace

Differential Expression using DESeq2

```
> res
log2 fold change (MLE): condition NAD supplement vs Control
Wald test p-value: condition NAD supplement vs Control
DataFrame with 46316 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
0610005C13Rik  5.99463012842517  0.847110388480526   1.0536757176372  0.803957398183298   0.4214215791485 0.626767086797856
0610006L08Rik 0.595406936513421  -1.33338402962542  2.80181545117752 -0.475900020133387  0.634145607845708                NA
0610009B22Rik  229.572854136365 -0.46059738889209 0.272726760296267   -1.688860265827 0.0912462113650002 0.227131423458176
0610009E02Rik  52.7148015454124  -1.18516447577791 0.483501805720158  -2.45121003015211 0.0142376849790884 0.058533268492583
0610009L18Rik  5.27096640148362  0.500548878654153   1.0060551707554  0.497536211933899  0.618810973397835 0.779869206330055
```

BH corrected p values

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- How many genes are differentially expressed at a significant level

```
sum(res$padj < 0.05, na.rm = TRUE)
```

```
> sum(res$padj < 0.05, na.rm = TRUE)
[1] 5374
```

- Collect all DE genes and write the results to file

```
sigGenes <- res[ which(res$padj < 0.05), ]
sigGenes
write.csv(sigGenes, "Differentially_Expressed.csv", row.names = TRUE)
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- PCA plot
  - Log transform the results and calculate variance

```
rv <- rowVars(assay(logTran))
select <- order(rv, decreasing = TRUE)[seq_len(min(100, length(rv)))]
```

  - Run the PCA and look at the results

```
PCA <- prcomp(t(assay(logTran)[select, ]), scale = F)
summary(PCA)
```

```
> summary(PCA)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5    PC6     PC7     PC8     PC9     PC10
Standard deviation     13.3632 2.44136 1.93651 1.50803 1.45379 1.2831 1.10025 0.51589 0.38353 3.16e-15
Proportion of Variance  0.9114 0.03042 0.01914 0.01161 0.01079 0.0084 0.00618 0.00136 0.00075 0.00e+00
Cumulative Proportion   0.9114 0.94178 0.96092 0.97252 0.98331 0.9917 0.99789 0.99925 1.00000 1.00e+00
```

https://www.huber.embl.de/users/klaus/Teaching/DESeq2Predoc2014.html

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- PCA plot
  - Set up everything for ggplot

```
percentVar <- round(100*PCA$sdev^2/sum(PCA$sdev^2),1)
ggPCA_out <- as.data.frame(PCA$x)
ggPCA_out <- cbind(ggPCA_out, sampleTable)
head(ggPCA_out)
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

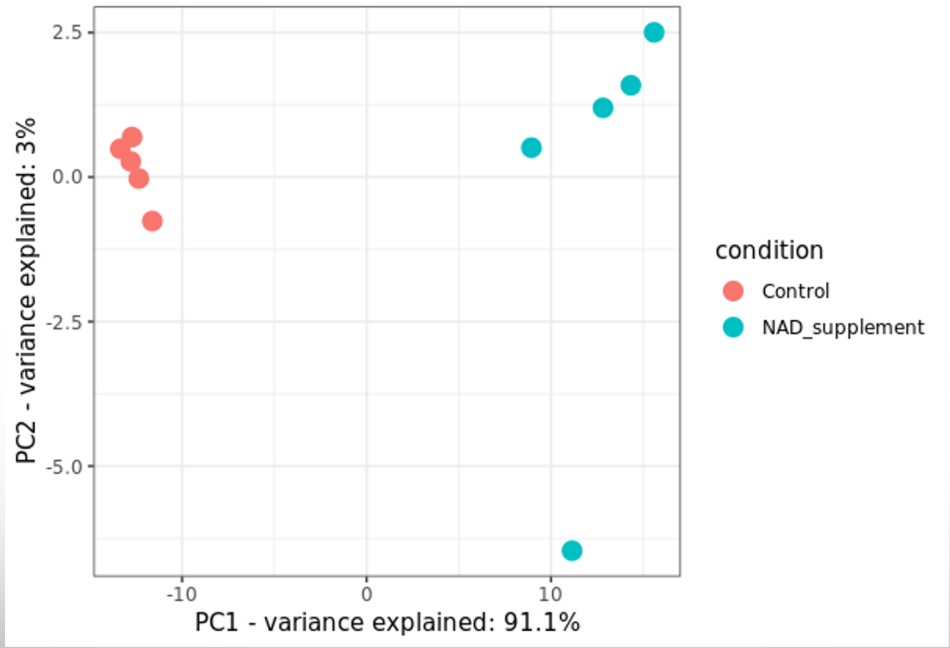
- PCA plot

```
ggplot(ggPCA_out, aes(x=PC1,y=PC2,color=condition)) +
  geom_point(size=4) +
  labs(x = paste0("PC1 - variance explained: ", round(percentVar[1],4), "%"),
       y = paste0("PC2 - variance explained: ", round(percentVar[2],4), "%")) +
  theme_bw()
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- PCA plot

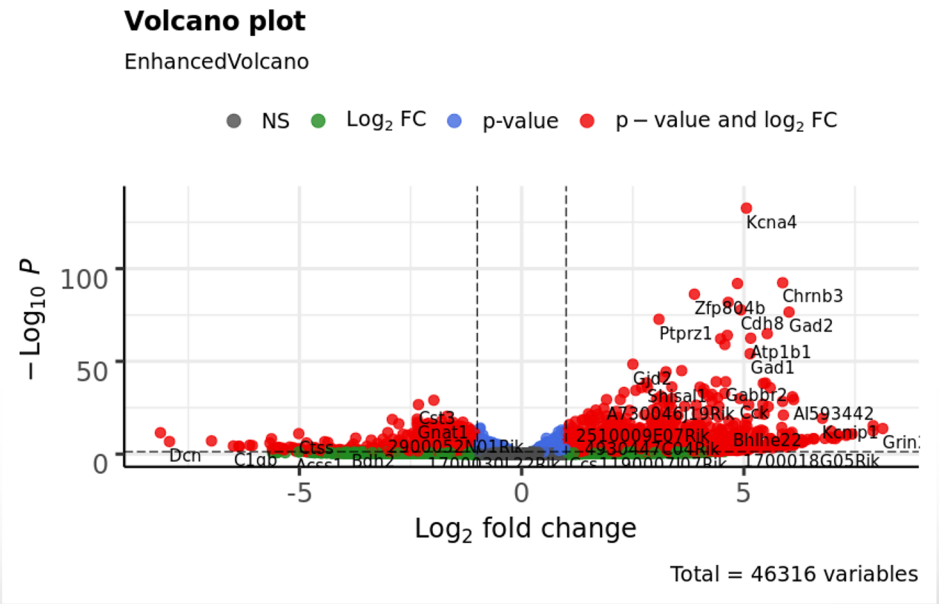


Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Volcano plot

```
EnhancedVolcano(res,
                lab = rownames(res),
                x = 'log2FoldChange',
                y = 'padj',
                pCutoff = 0.05,
                FCcutoff = 1.0,
                pointSize = 3.0,
                labSize = 4.0,
                colAlpha = 4/5,
                drawConnectors = FALSE)
```



Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Volcano plot
  - Change categories and plot again

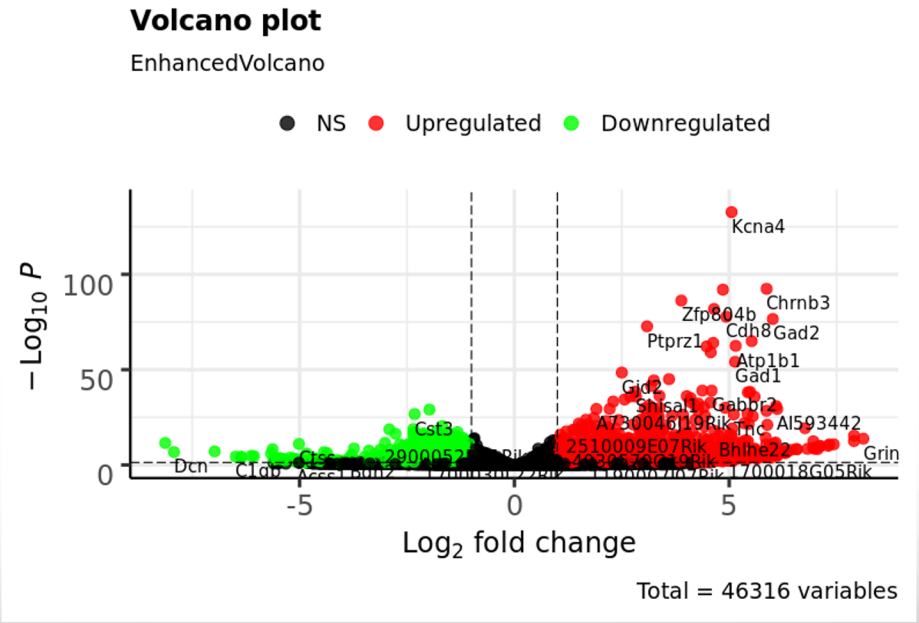
```
keyvals <- ifelse(
  res$log2FoldChange >= 1.0 & res$padj <= 0.05, 'red',
  ifelse(res$log2FoldChange <= -1.0 & res$padj <= 0.05, 'green', 'black'))

keyvals[is.na(keyvals)] <- 'black'
names(keyvals)[keyvals == 'red'] <- 'Upregulated'
names(keyvals)[keyvals == 'green'] <- 'Downregulated'
names(keyvals)[keyvals == 'black'] <- 'NS'
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Volcano plot
  - Change categories and plot again

```
EnhancedVolcano(res,
                lab = rownames(res),
                x = 'log2FoldChange',
                y = 'padj',
                pCutoff = 0.05,
                FCcutoff = 1.0,
                pointSize = 3.0,
                labSize = 4.0,
                colAlpha = 4/5,
                colCustom = keyvals,
                drawConnectors = FALSE)
```

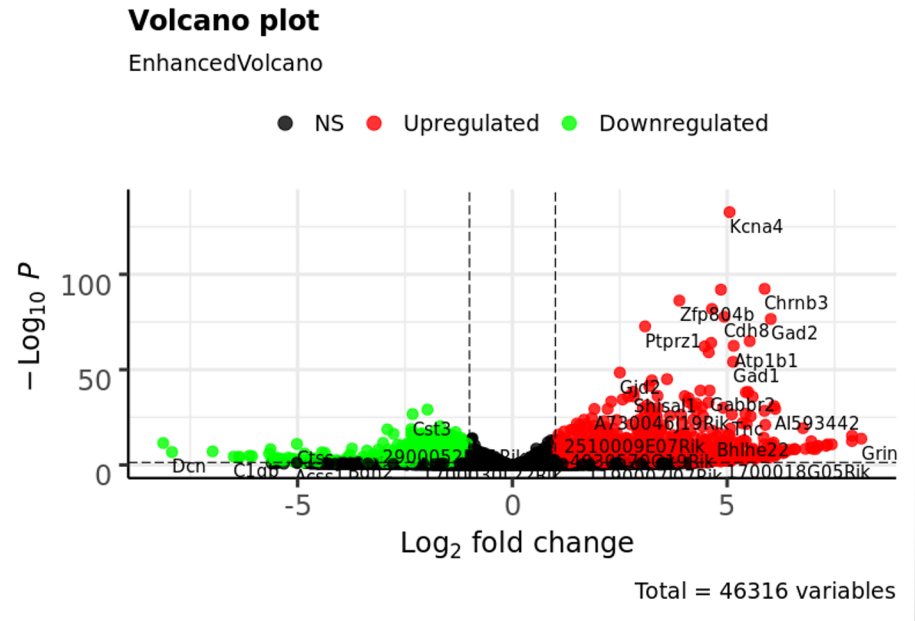


Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Volcano plot
  - Change categories and plot again

```
EnhancedVolcano(res,
                lab = rownames(res),
                x = 'log2FoldChange',
                y = 'padj',
                pCutoff = 0.05,
                FCcutoff = 1.0,
                pointSize = 3.0,
                labSize = 4.0,
                colAlpha = 4/5,
                colCustom = keyvals,
                drawConnectors = FALSE)
```



Differential Expression Analysis on Grace

Differential Expression using DESeq2

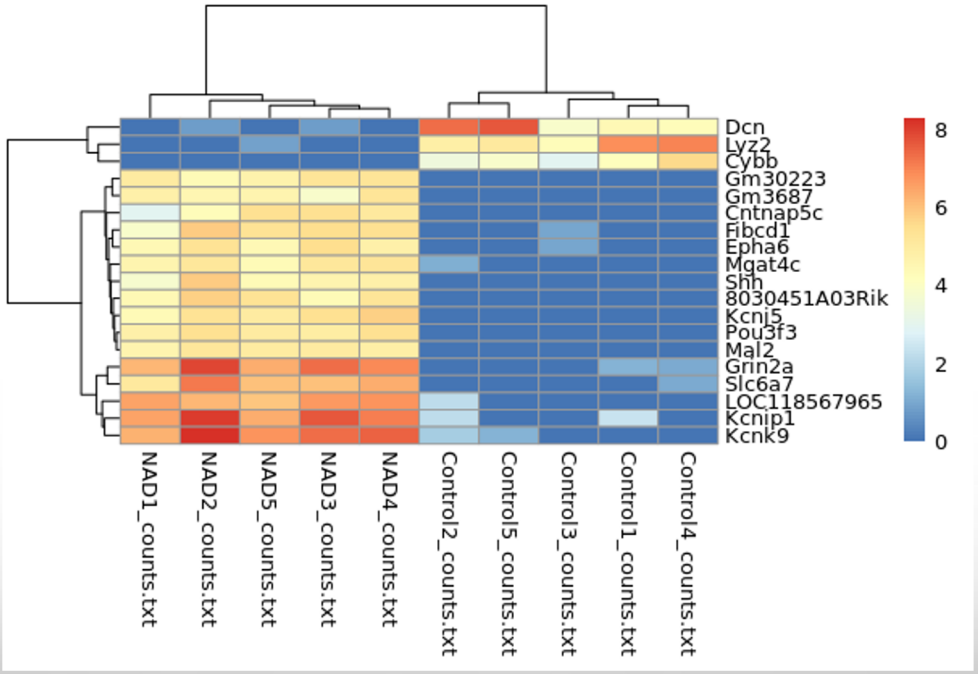
- Heatmap

```
resorted_deresults <- res[order(res$padj),]
sig <- resorted_deresults[!is.na(resorted_deresults$padj) &
                          resorted_deresults$padj < 0.05 &
                          abs(resorted_deresults$log2FoldChange) >= 6.5,]

selected <- rownames(sig);selected
ntd <- normTransform(dds)

pheatmap(assay(ntd)[selected,], cluster_rows = TRUE, show_rownames = TRUE,
cluster_cols = TRUE, labels_col = colData(dds)$sampleName)
```

Differential Expression Analysis on Grace

Differential Expression using DESeq2

- Heatmap



Differential Expression Analysis on Grace

More downstream analyses in R...

- GO enrichment
  - gprofiler2
- Pathway analysis
  - gage
  - pathview

QUESTIONS?