SACNAS 2023
RNA-seq Tutorial

Post-Sequencing Processing Steps

Princess Rodriguez, Noelle Gillis, Sophie Kogut, & Seth Frietze

October 27th, 2023

Steps to get you from raw RNA-seq data to DEG's and beyond!

How I learned how to do this:

- Trial and Error.
- Spending lots of time reading the program manuals.
- Googling error messages.
- Asking for help when I get stuck.

Lots of this:

But eventually this:

Steps to get you from raw RNA-seq data to DEG's and beyond!

Phase 1: Processing

Processing Step	Tools
Quality control of raw sequences	FastQC, cutadapt, trimGalore
Alignment/Mapping to the genome	STAR
Generating gene counts	HT-seq

Phase 2: Analysis

Analysis Step	Tools
Quality control of replicates	DESeq2, edgeR
Differential expression analysis	DESeq2, edgeR
Pathway analysis	GSEA, IPA, clusterProfiler

Quality Control of Raw Sequences

FASTQ File Format

Quality Control Steps

1. Generate FASTQC reports.
2. Check the quality of base-calls.
3. Check for “over-represented sequences.”
4. Trim low-quality bases.
5. Remove adapter sequences.

Example Commands:

```
fastqc 2D_WT_shGFP_E2_R50_001_S10_R1_001.fastq.gz

trim_galore --paired --Illumina 2D_WT_shGFP_E2_R50_001_S10_R2_001.fastq.gz
2D_WT_shGFP_E2_R50_001_S10_R2_001.fastq.gz
```

Alignment/Mapping to the Genome

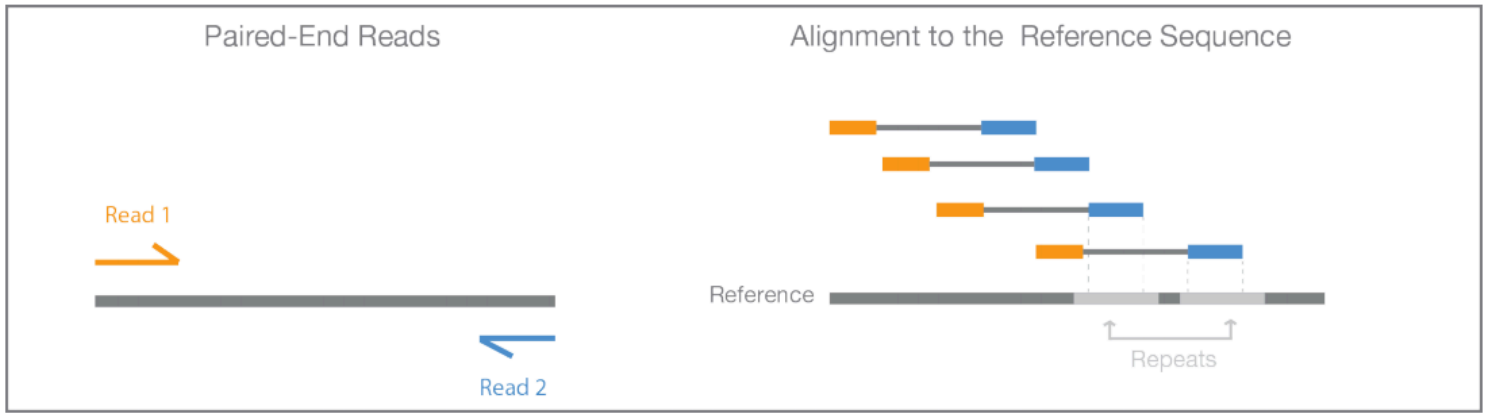


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better alignment of reads, especially across difficult-to-sequence, repetitive regions of the genome.

Alignment Steps

1. Align FASTQ files to a reference genome.
2. Check the alignment statistics.
3. Adjust alignment settings as needed.
4. Convert to BAM format.

Where do you get the reference genome files? https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

Example Command:

```
STAR --runThreadN 12 \
--genomeDir /home/langeca/gilli431/software/STAR_hg38 \
--readFilesCommand zcat \
--readFilesIn 2D_WT_shGFP_E2_R50_001_S10_R1_001.fastq.gz
2D_WT_shGFP_E2_R50_001_S10_R2_001.fastq.gz \
--outFileNamePrefix 2D_WT_shGFP_E2_R50_001
```

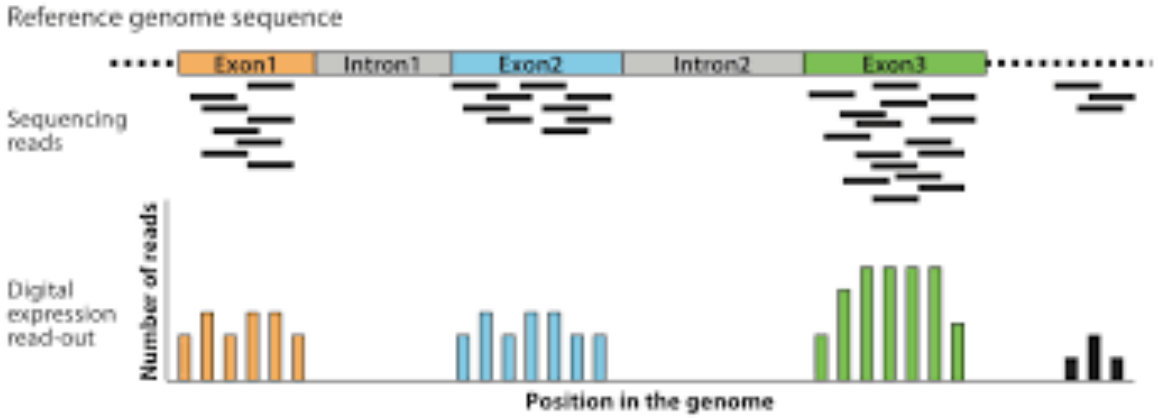
STAR Alignment Summary:

```
Number of input reads | 35202511
Average input read length | 299
                 UNIQUE READS:
   Uniquely mapped reads number | 33777023
        Uniquely mapped reads % | 95.95%
          Average mapped length | 298.68
```

Generating Gene Counts

Gene Counting Steps

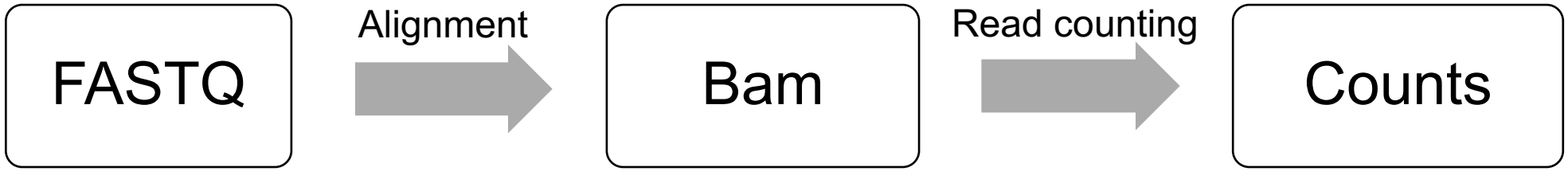
1. Check that bam files are sorted and indexed
2. Run HT-seq on bam files.
3. Save .cnts files for analysis.



Example Command:

```
htseq-count -f bam -r name -s reverse -m union -i gene_id bam_files/${base}_sorted.bam
~/software/gencode.v38.primary_assembly.annotation.gtf > count_files/${base}.counts
```

All these steps can be strung together into a “pipeline.”

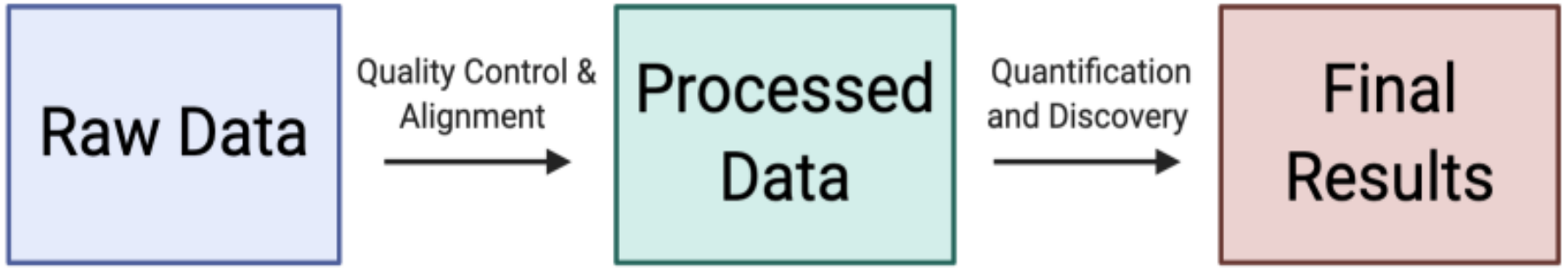


RNA-seq_pipeline_slurm.sh | functional_RNA-seq_pipeline_slurm.sh

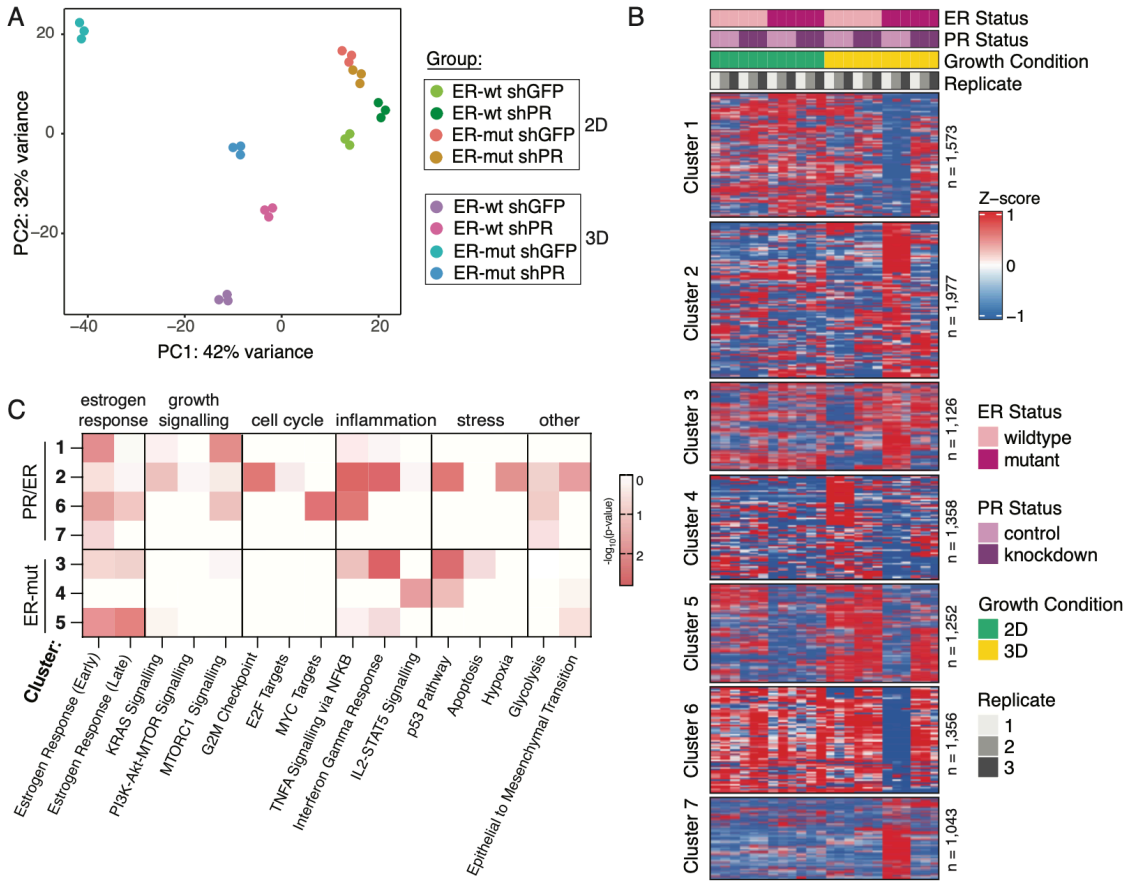
```
# Load necessary modules - check what you have installed on your server vs. what is in your bash
# module load fastqc
# module load cutadapt
# module load star
# module load samtools
# module load htseq
# module load multiqc

# Create directory for FastQC reports
mkdir fastqc_reports

# Create directory for trimmed fastq files
mkdir trimmed_fastq

# Run FastQC on all input FASTQ files and perform adapter trimming with TrimGalore
for f in *_R1_001.fastq.gz; do  #check the base name of your files and modify as needed
base=$(basename ${f} _R1_001.fastq.gz)
fastqc -o fastqc_reports ${base}_R1_001.fastq.gz ${base}_R2_001.fastq.gz; # Generate FastQC report
trim_galore --paired --illumina --fastqc --output_dir trimmed_fastq ${base}_R1_001.fastq.gz ${base}_R2_001.fastq.gz; # Trim
done

# Make a directory for the bam files
mkdir bam_files

# Run STAR on the trimmed FASTQ files
for f in trimmed_fastq/*_R1_001_val_1.fq.gz; do
  base=$(basename $f _R1_001_val_1.fq.gz); # Extract basename of input file
  r2="$base"_R2_001_val_2.fq.gz"
STAR --runThreadN 12 \
     --genomeDir /home/langeca/gilli431/software/STAR_hg38 \
     --readFilesCommand zcat \
     --readFilesIn $f trimmed_fastq/$r2 \
     --outFileNamePrefix bam_files/${base}_ \
     --outSAMtype BAM \

# Sort and index the BAM files
samtools sort -o bam_files/${base}_sorted.bam bam_files/${base}_Aligned.sortedByCoord.out.bam
samtools index bam_files/${base}_sorted.bam

# Make a directory for the count files
mkdir count_files
```

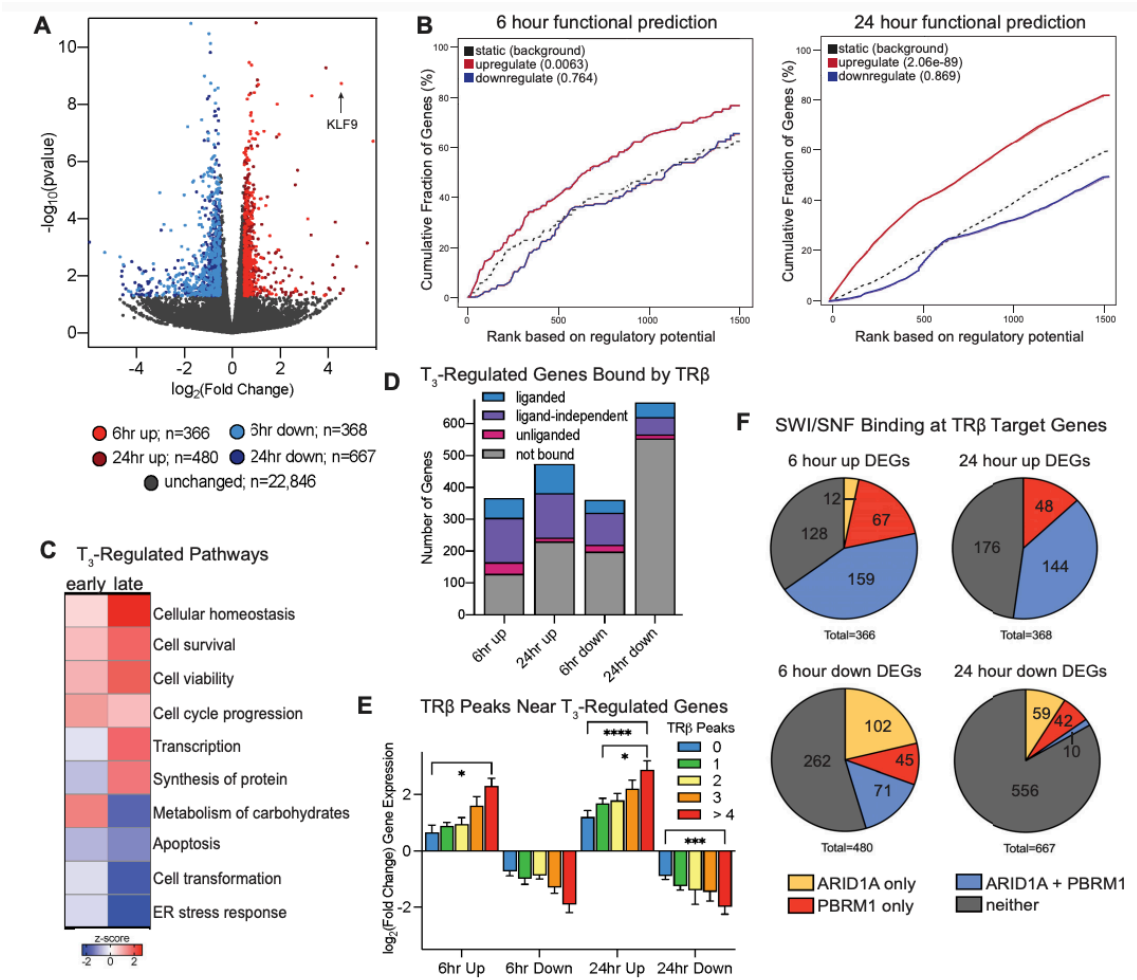
1. Put the raw data for the experiment you are analyzing into a single folder (on MSI).
2. Modify the SLURM header in the pipeline script
3. Send the job out for analysis.
4. Wait 12-24 hours.
5. Check your QC metrics and retrieve the count files.

What can be done with the processed data?



What can be done with the processed data?





Resources, References, and Manuals

Tool:	Reference:	Manual:
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/
STAR	Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21.	https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
HT-Seq	Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics, 31, 166–169	https://htseq.readthedocs.io/en/master/overview.html
DESeq2	Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550	http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

R for Data Science Ebook: https://r4ds.had.co.nz/
Unix Command Line Basics: https://www.unixtutorial.org/basic-unix-commands
Bioconductor: http://bioconductor.org/
Guide to Using R Markdown Files: https://bookdown.org/yihui/rmarkdown/
Guide to Making and Using Shiny Apps: https://shiny.rstudio.com/
ChatGPT: https://chat.openai.com/chat