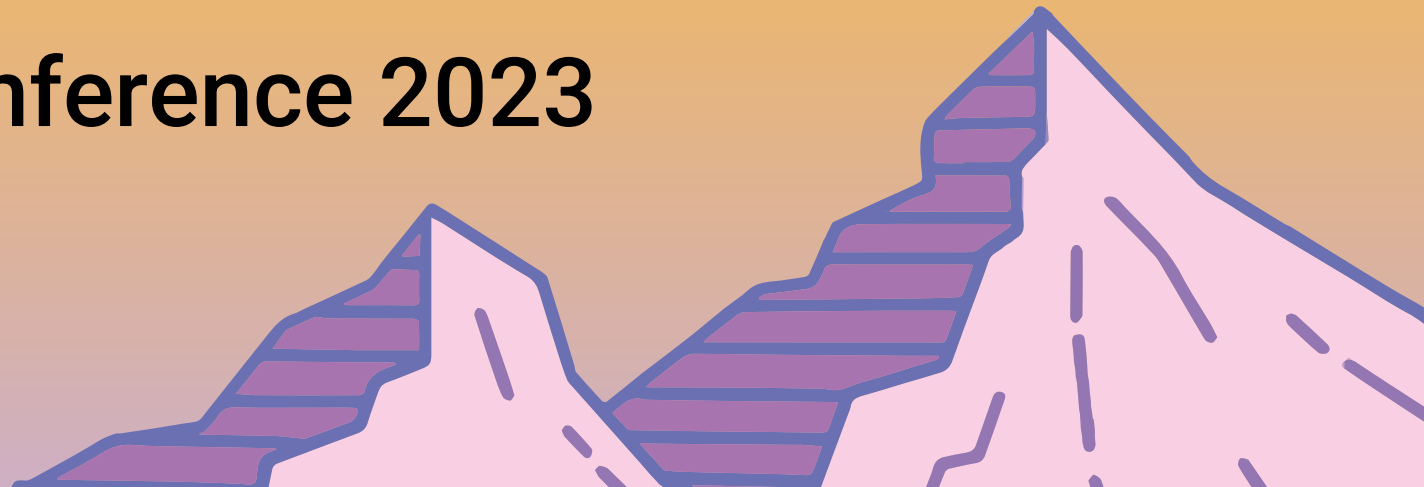
A hands-on workshop for the experimental design analysis and interpretation of RNA-seq data for interdisciplinary research

SACNAS NDiSTEM Conference 2023

Portland, OR



Seth Frietze, PhD

- Ph.D.; Harvard University
- Postdoctoral training; University of Southern California Medical School Norris Cancer Center
- Associate Professor
- Program co-leader UVM Cancer Center Cancer Cell Program

THE University of Vermont CANCER CENTER

Princess Rodriguez, PhD

- MS in Immunology
- PhD, Cell and Molecular Biology
  - Started my bioinformatic journey in 2015 as a grad student in Seth's lab.
- Assistant Professor
  - Teach intro and advanced bioinformatics at UVM

THE University of Vermont CANCER CENTER

Noelle Gillis, PhD

- BS in Microbiology & Molecular Genetics
- PhD, Cell and Molecular Biology
  - Started my bioinformatic journey in 2016 as a grad student in the co-mentored by Seth Frietze.
- Currently a postdoc at U Minnesota Masonic Cancer Center.
- Working on projects related to steroid receptor transcriptional regulation in breast and ovarian cancer models.

MASONIC CANCER CENTER

UNIVERSITY OF MINNESOTA

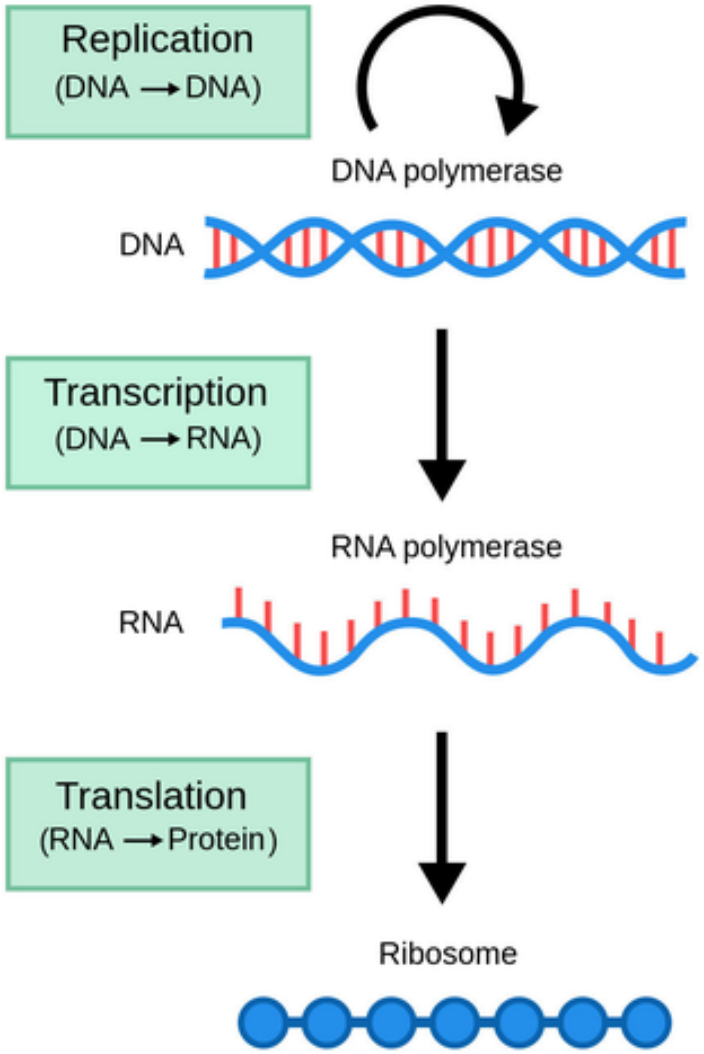
Sophie Kogut

- BS in Biological Science
- Former research tech in Seth's lab
- PhD student in the Molecular and Cellular Biology Program at the University of Washington / Fred Hutchinson Cancer Center

Fred Hutch
Cancer Center

Module 1:
Overview of RNA-seq

RNA-seq overview



Experimental design

Total RNA

Library prep

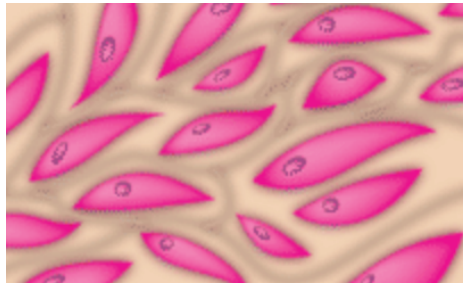
Sequence

Analyze

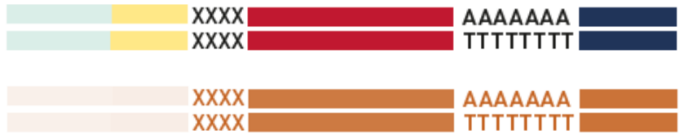
Experimental ("WET")

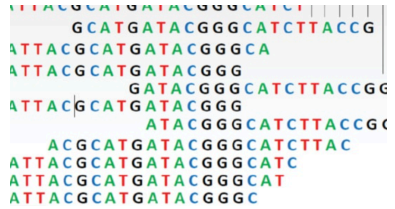
Analysis ("DRY")

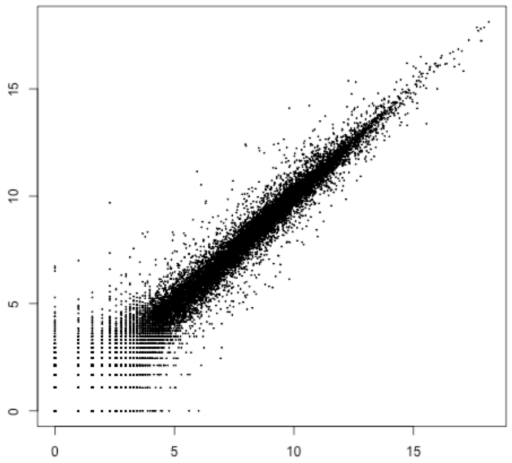
Biological material



cDNA libraries

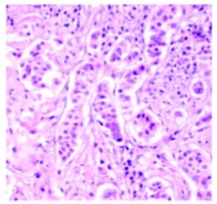


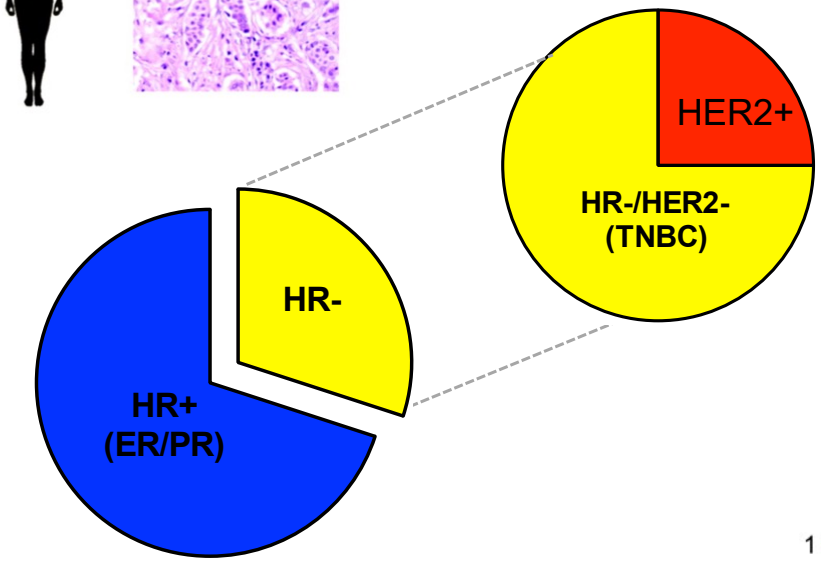




Breast cancer heterogeneity

Gene expression is diagnostic and prognostic



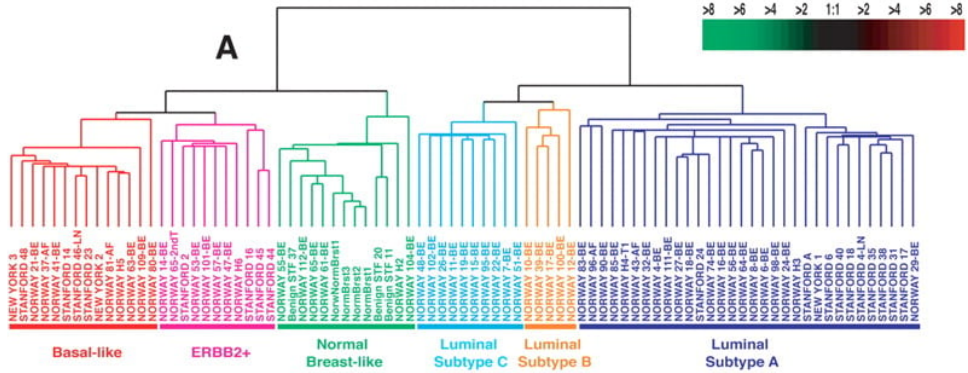


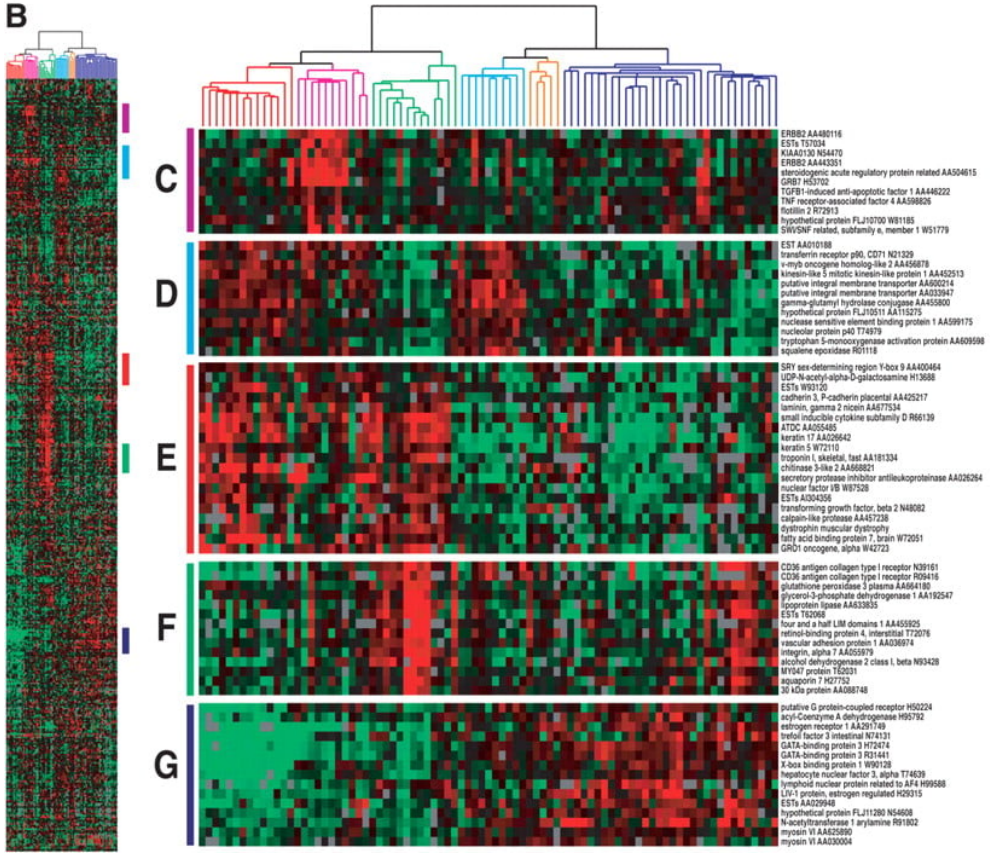
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

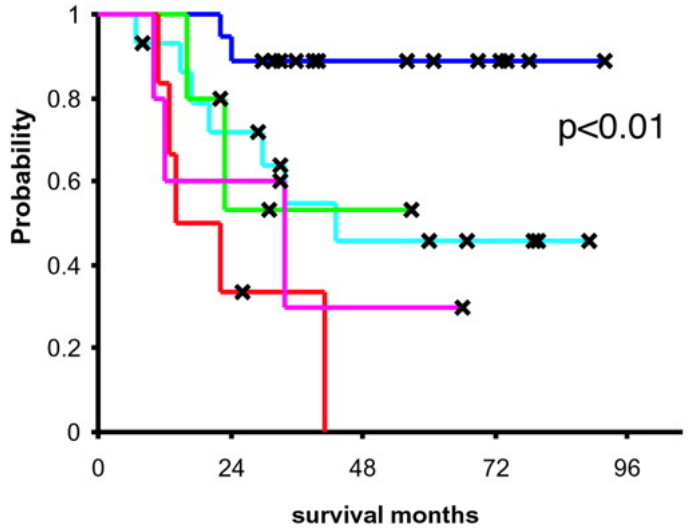
Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

Departments of [b]Genetics and [l]Surgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; [d]Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of [e]Health Research and Policy and Statistics, [c]Genetics, [f]Pathology, [i]Surgery, and [m]Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [g]Medicine (Section of Oncology), [j]Surgery, and [k]Biochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and [h]Life Sciences Division, Lawrence Orlando Berkeley National Laboratories, and Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720

Contributed by David Botstein, July 17, 2001

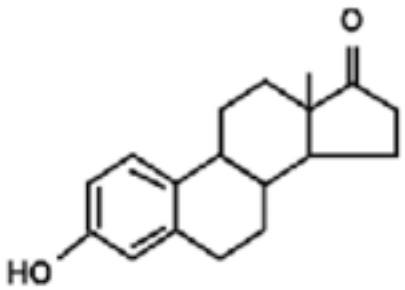




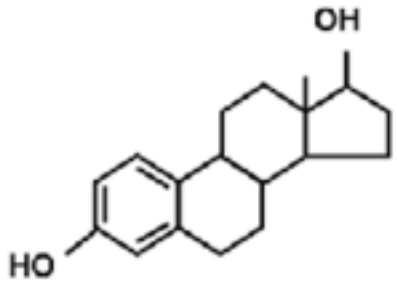


Endocrine therapies are based on estrogen ligands

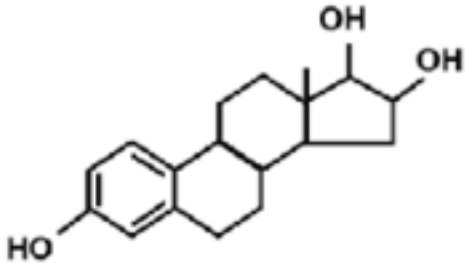
Estrogens



Estrone

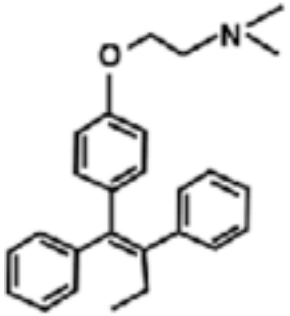


17β-Estradiol

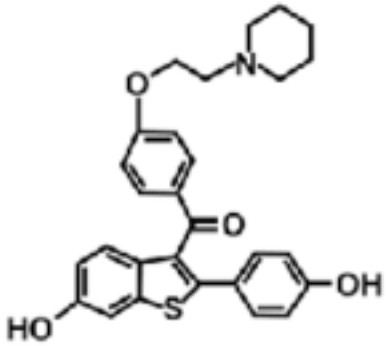


Estriol

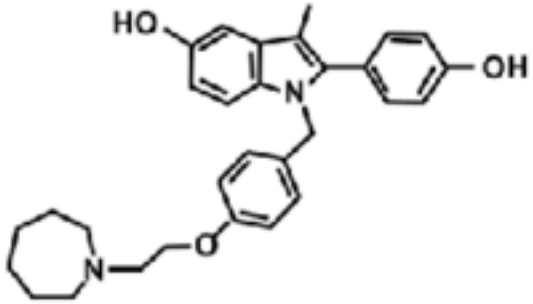
SERMs



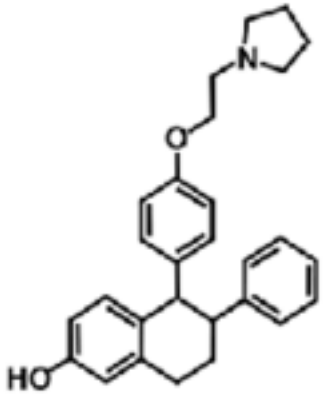
Tamoxifen



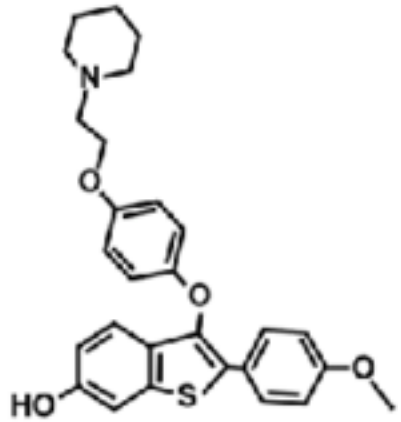
Raloxifene



Bazedoxifene

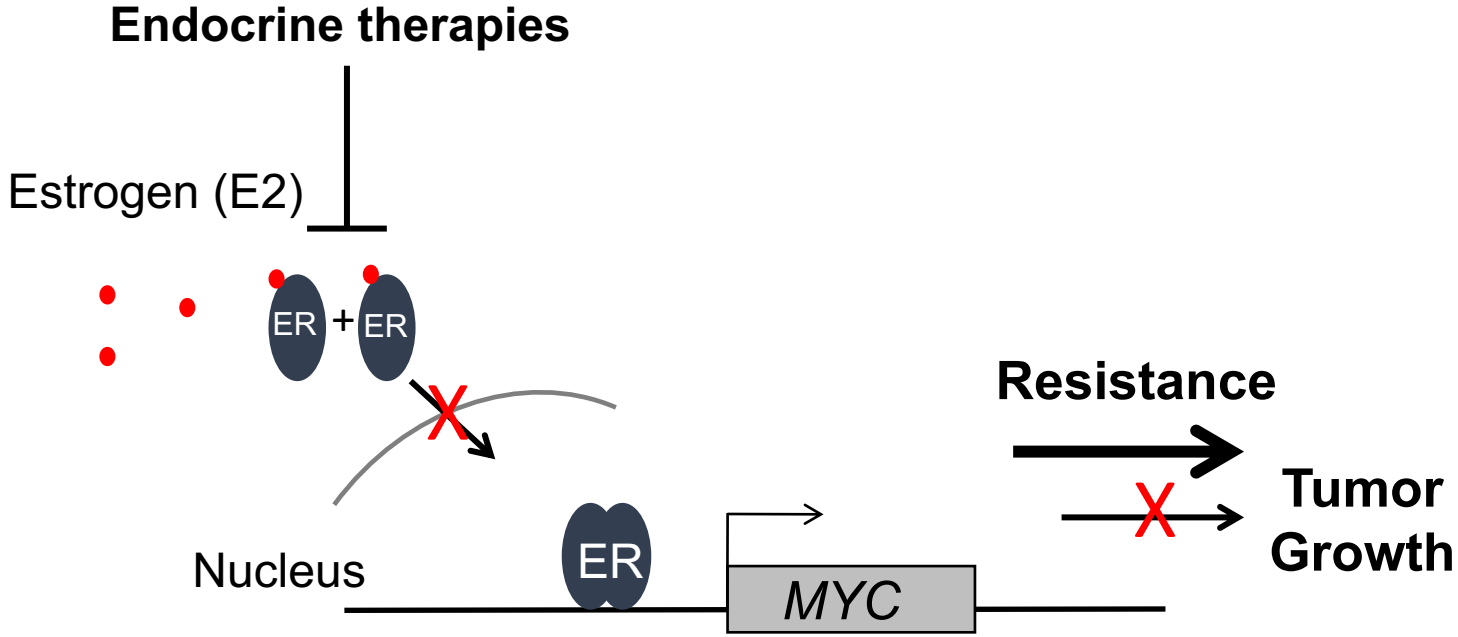


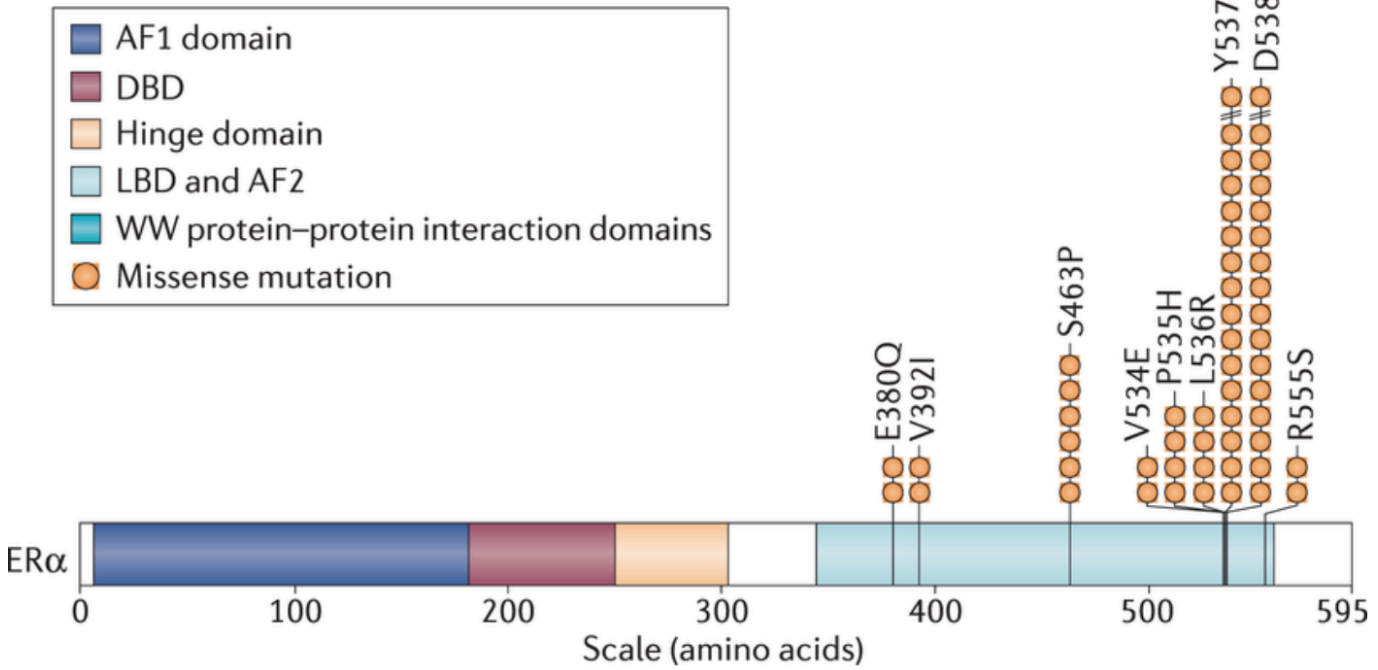
Lasofoxifene



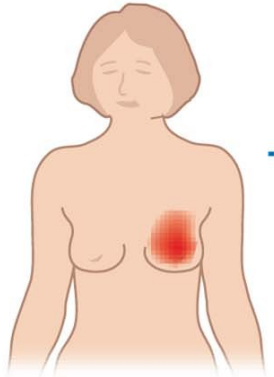
Arzoxifene

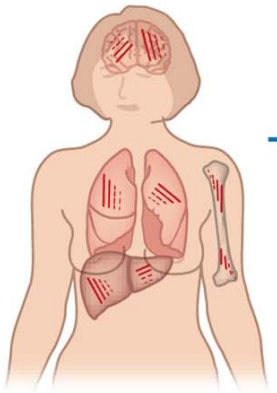
ESR1 mutations are common in endocrine therapy resistance (ETR)





Cell models to study ETR: CRISPR engineering *ESR1* Y573S





Primary untreated ER+ breast cancer

Rare *ESR1* mutation

Tamoxifen

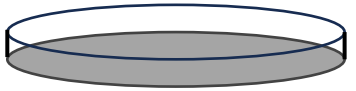
Aromatase inhibitor

Treatment-resistant metastatic ER+ breast cancer

Rare *ESR1* mutation

ESR1 mutations (~20–50%)

MCF-7 (ER+)

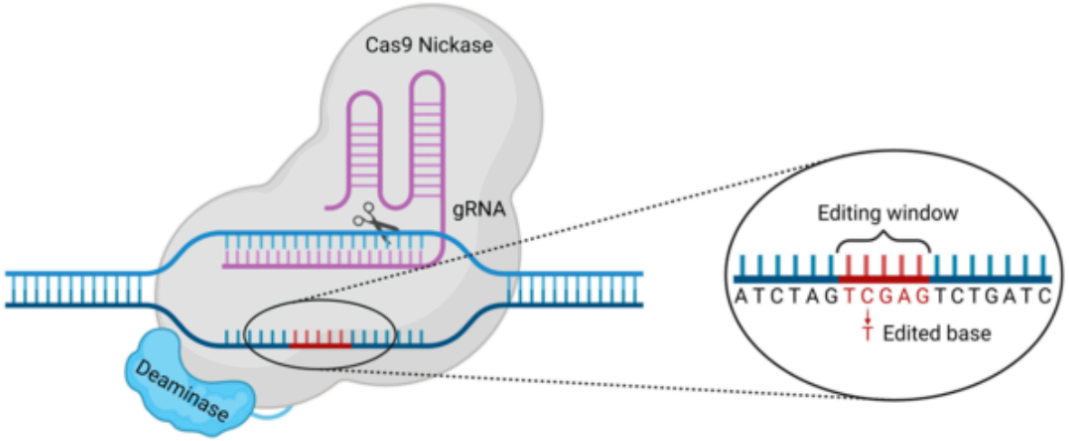


parental

CRISPR base-edit

MCF-7 (ER+)

Y537S



CRISPR-base editing method

1. Find *ESR1* sequence on Genome browser
2. Protein sequence on Uniprot
3. Import sequence into benchling, design gRNAs
4. Generate vectors, deliver and select stable cell lines
5. Characterize (RNA-seq)

genotype

Parental (wild-type)

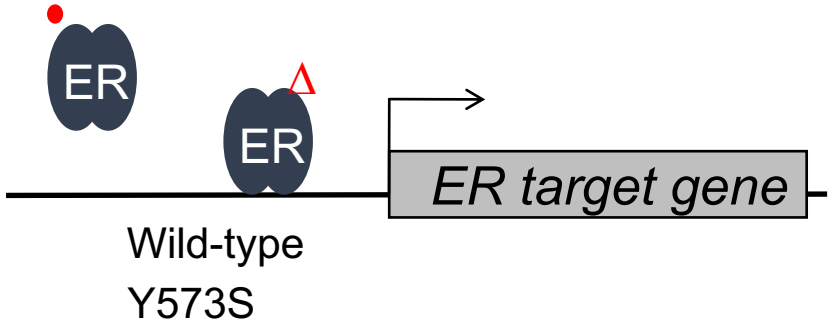
Y573S

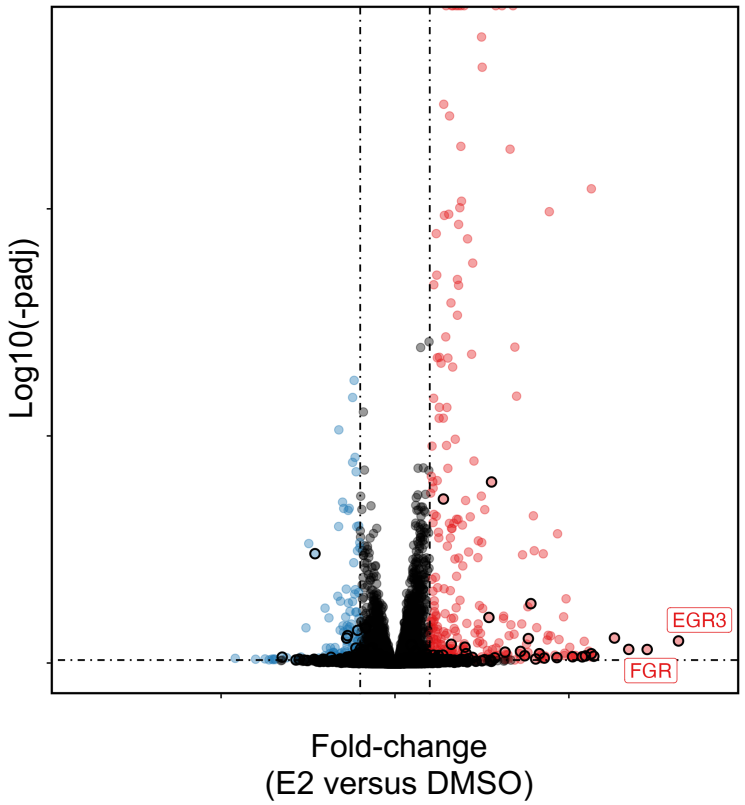
treatment

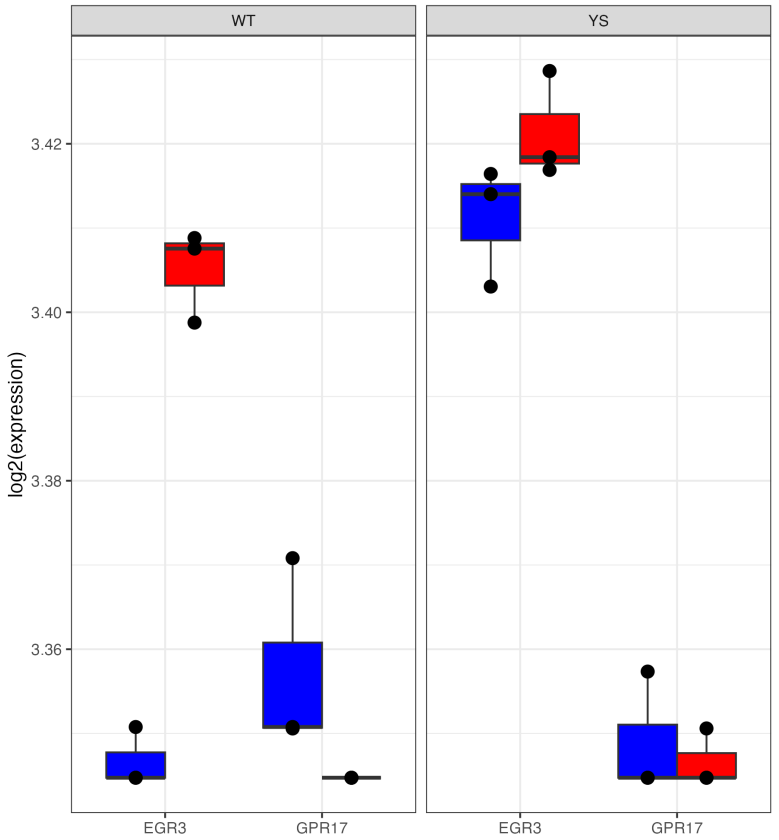
DMSO/E2

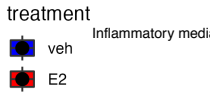
6hr

RNA-seq analysis









Module 2:
Post-Sequencing Processing Steps

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

<u>How I learned how to do this:</u>

- 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, Salmon

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

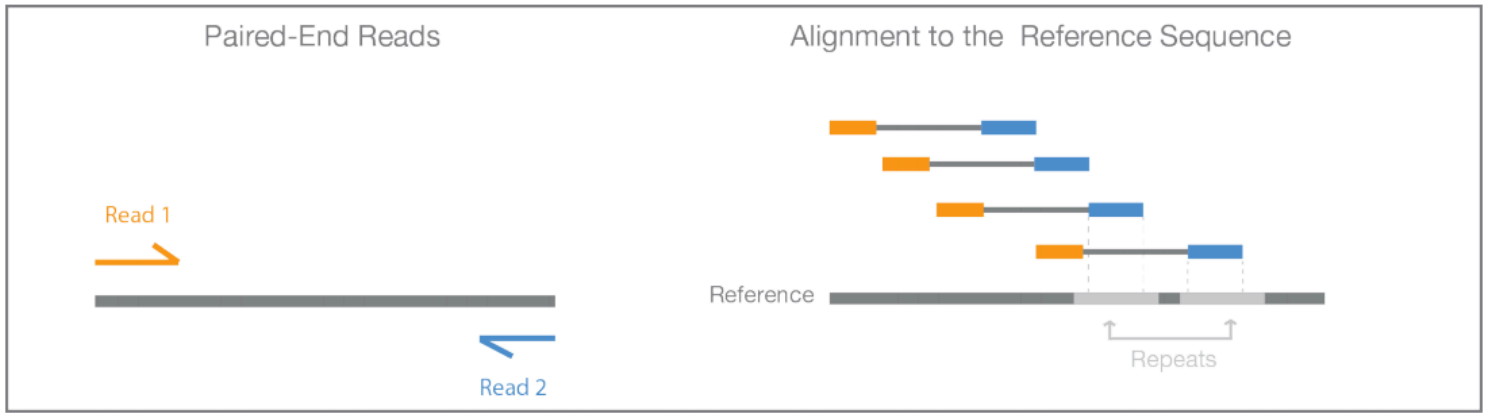


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better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

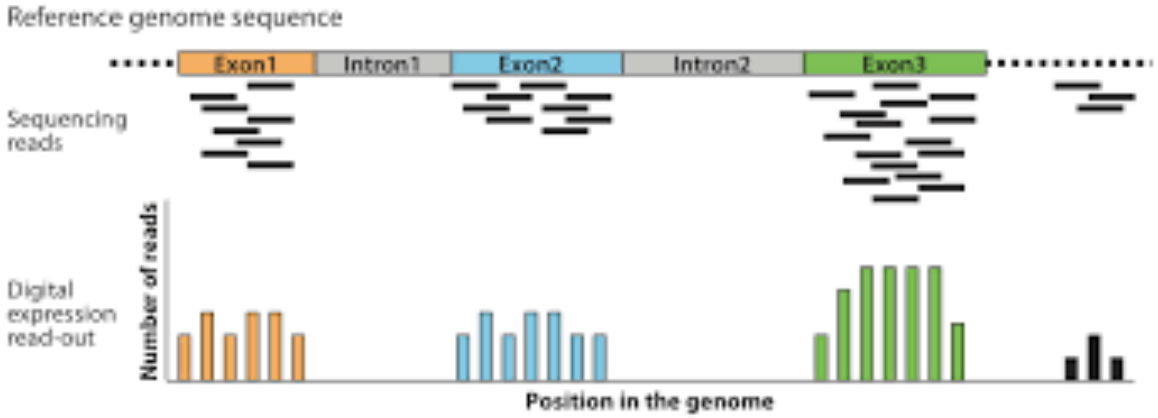
STAR Alignment Summary:

```
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

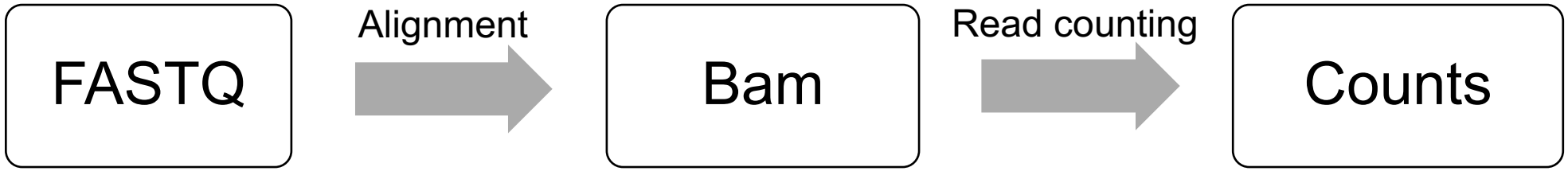
1. Check that bam files are sorted and indexed
2. Run HT-seq on bam files.
3. Save .cnts files for analysis.



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.”

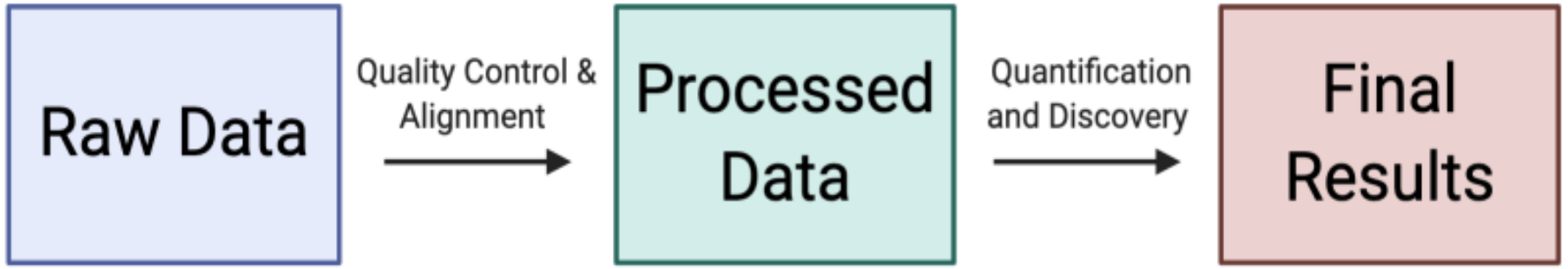


RNA-seq_pipeline_slurm.sh | functional_RNA-seq_pipeline_slurm.sh

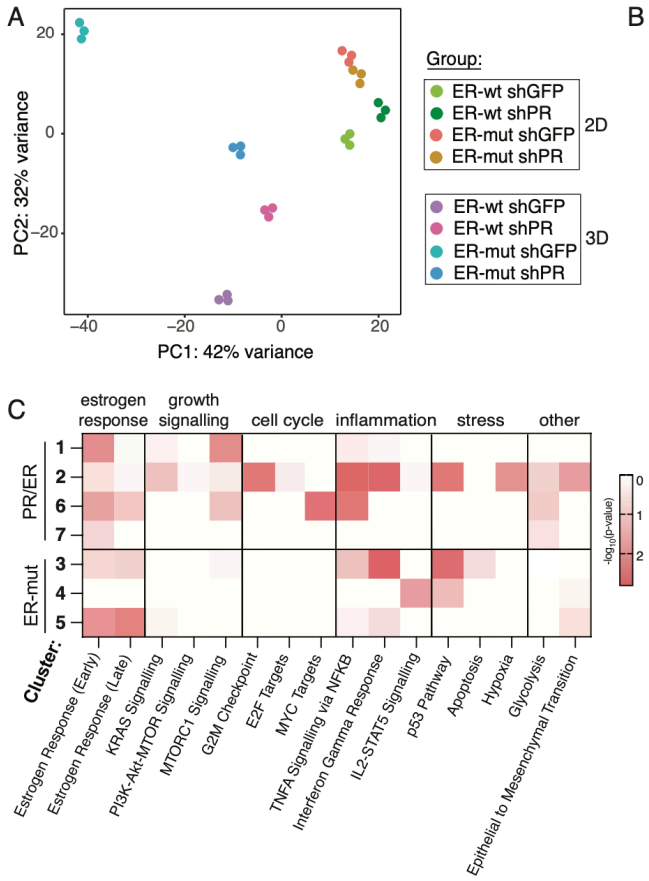
```
# Load necessary modules - check what you have installed on your server vs. what is in your bash
# module load fastqc
# module load cutadapt
# module load star
# module load samtools
# module load htseq
# module load multiqc

# 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
```

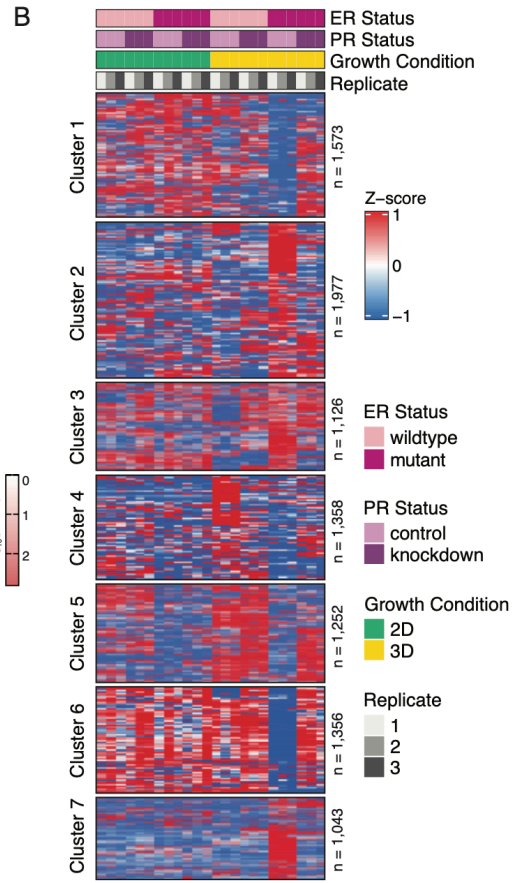
1. Put the raw data for the experiment you are analyzing into a single folder (on MSI).
2. Modify the SLURM header in the pipeline script
3. Send the job out for analysis.
4. Wait 12-24 hours.
5. Check your QC metrics and retrieve the count files.

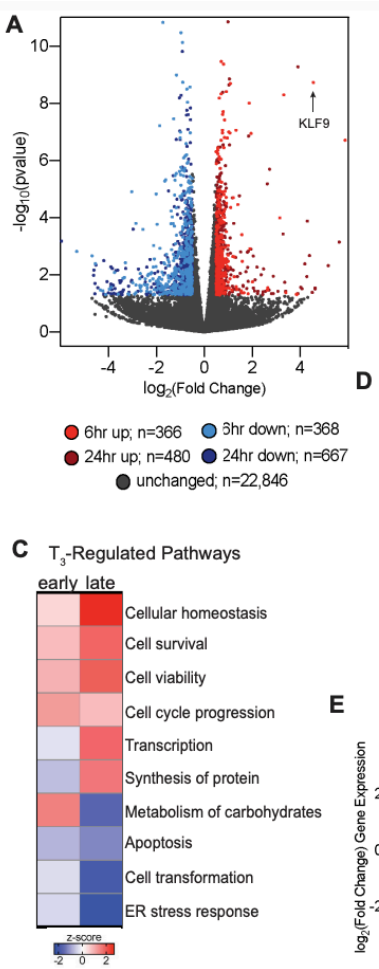
What can be done with the processed data?

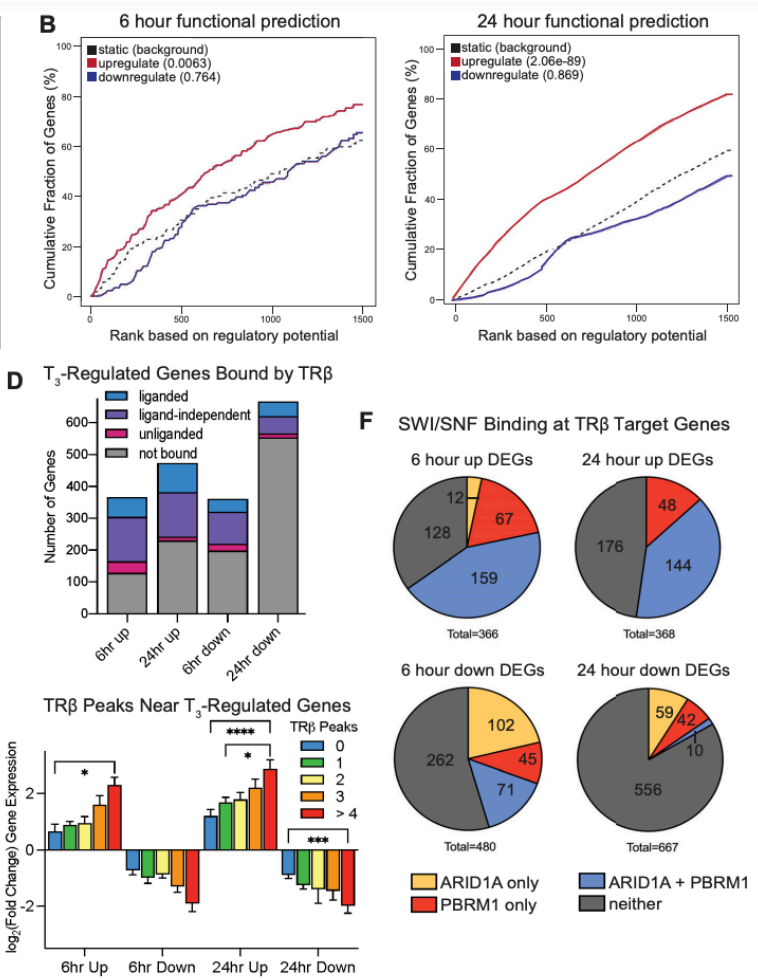


What can be done with the processed data?









Module 3: Hands-On Workshop

Group A

Is this your first time analyzing an RNA-Seq dataset?

*Participants **without** a laptop or are new to RNA-Seq are encouraged to join Group A*

Group B

Do you have previous experience with R/Rstudio?

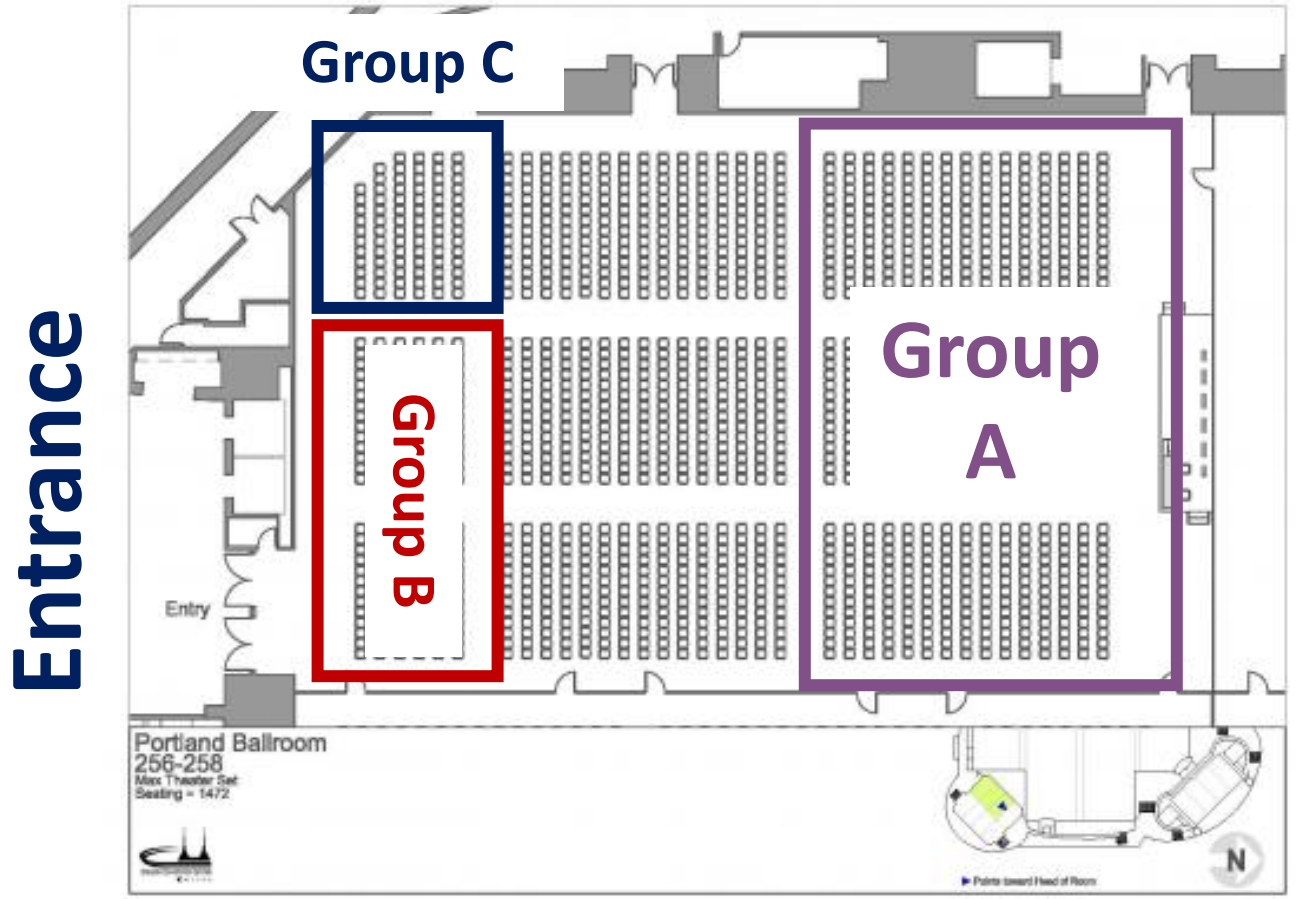
*Participants **with** a laptop are encouraged to join Group B*

Group C

Do you have questions about an RNA-Seq dataset you generated or would like to analyze?

Informal discussion about NGS

Hands-On Workshop



Please take a few minutes to join a group!