CRI Bioinformatics Workshop

Bulk-RNAseq Pipeline

Hands-on Training

Center for Research Informatics – Bioinformatics Core

Genomics and proteomics data analysis

BiCF applies appropriate and state-of-the-arts statistical and bioinformatic methodologies to analyze genomics data generated from standard and emerging assays.

Consulting, grant writing and training

BiCF provides consulting services for experimental design or data analysis; grant writing assistance, including bioinformatics development, cost analysis, and documentation of tools to complete the research.

Data management system development

BiCF offers enterprise solutions for project and study management, for data production , sharing and integration .

OUR AWESOME TEAM

Mengjie Chen, PhD

Faculty Director

Wenjun Kang, MS

Technical Director

Yan Li, PhD

Associate Director

Bioinformaticians

Jason Shapiro, PhD

Diana Vera Cruz, PhD

Evan Wu

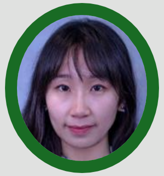
Katie Aracena, PhD

David Tieri, PhD

Houxiang Zhu, PhD

Qiaoshan Lin, PhD

Yildiz Koca, PhD



Zhongyu Li, MS

Geetha Priyanka, MS

Contact us: bioinformatics@bsd.uchicago.edu Submit a project request: https://biocore.cri.uchicago.edu/

Interacting with the Bioinformatics and Biostatistics Cores

Theodore Karrison, PhD
Technical Director, Biostatistics Core Facility

Biostatistics Core Facility

Members

Scientific Director:	Donald Hedeker, PhD
Technical Director:	Theodore Karrison, PhD
Faculty Affiliate:	Jim Dignam, PhD
Biostatisticians:	Yan Che, MS
	John Cursio, PhD
	Guimin Gao, PhD
	Mihai Giurcanu, PhD
	Sang Mee Lee, PhD
	Eric Polley, PhD
	Mei Polley, PhD

Role and Responsibilities

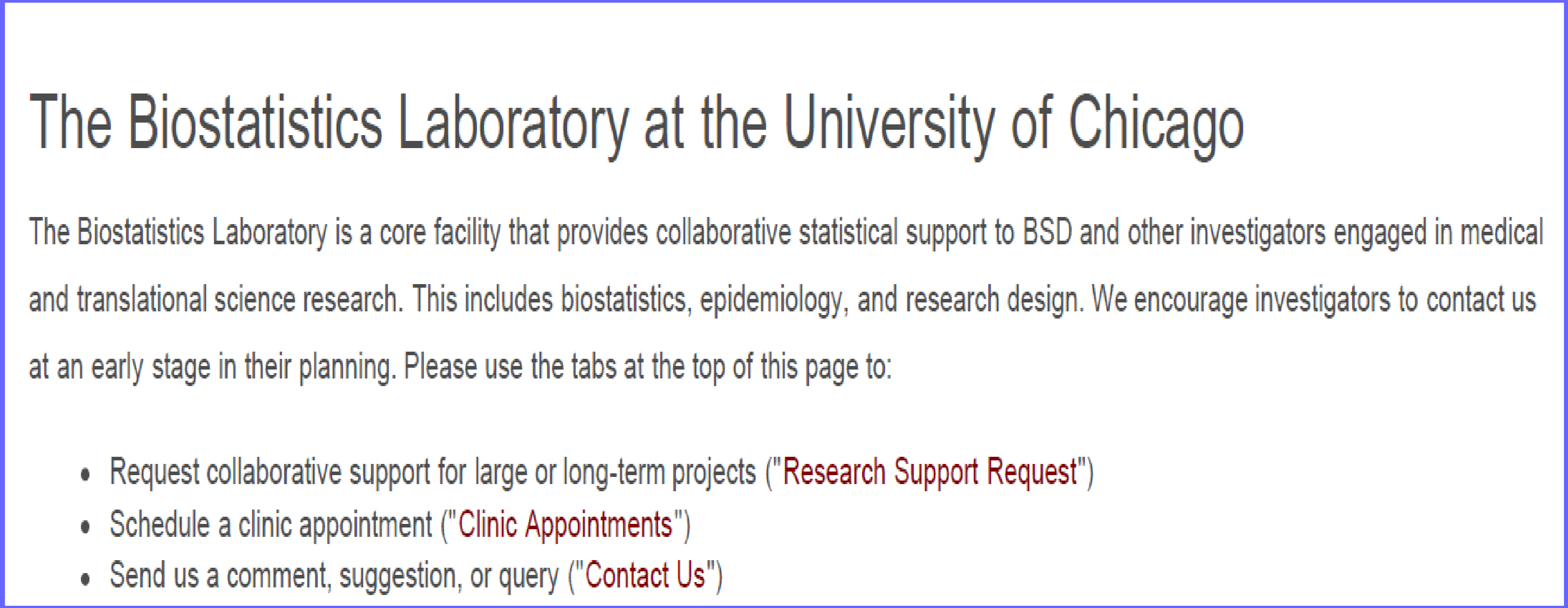
- Collaborate with UCCCC investigators in the formulation of study designs and data analysis plans, including sample-size and power calculations
- Collaborate on the design, analysis, and reporting of investigator-initiated clinical trials
- Work with the UCCCC clinical trial data managers to facilitate effective use of clinical trial data collection resources (OnCORE, REDCap, Velos—to be phase out)
- Report outcomes of completed clinical trials into ClinicalTrials.gov
- Collaborate in the development of grant proposals (co-investigator)
- Perform statistical analyses and assist in the interpretation of study findings, summarization of results, and preparation of manuscripts for publication

Role and Responsibilities (cont.)

- Operate Biostatistics Clinic for short-term consultations
- Review protocols for the Protocol Review Monitoring Committee (PRMC)
- Perform statistical methodological research on problems arising from collaborative work
- *Interact with Bioinformatics Core to support UCCCC investigators conducting research involving genomics, proteomics, and other “omics” data*

Accessing the Biostatistics Core

For most collaborations, requests and tracking are performed via our *Biotime* website. Access can be obtained via the following link: https://biotime.uchicago.edu/



The Biostatistics Laboratory at the University of Chicago

The Biostatistics Laboratory is a core facility that provides collaborative statistical support to BSD and other investigators engaged in medical and translational science research. This includes biostatistics, epidemiology, and research design. We encourage investigators to contact us at an early stage in their planning. Please use the tabs at the top of this page to:

- Request collaborative support for large or long-term projects ("Research Support Request")
- Schedule a clinic appointment ("Clinic Appointments")
- Send us a comment, suggestion, or query ("Contact Us")

Biostatistics Clinic:

The Biostatistics Clinic is supported by CTSA funds. Investigators can sign up for a one-hour time slot for free, short-term statistical consulting advice.

How will the Cores Interact to Better Serve Investigators?

- When investigators come to the BCF and we determine that bioinformatics expertise is needed, the investigator will be referred to the BIC (if not already contacted)

- Conversely, investigators coming first to the BIC will be referred to the BCF if needed

- These referrals will be tracked within each Core's database for reporting purposes

- When BCF and BIC are to work together with an investigator, and initial three-way meeting will take place to determine best study design, analytic approaches, and division of labor

- BIC holds five workshops annually. The BCF will attend a fall and spring workshop to "advertise" and promote the benefits of joint collaboration

- BIC operates a weekly "walk-in" clinic (Tuesday from 12:30-3:30). A representative(s) from the BCF will be present at this clinic on the first Tuesday of each month

The BIC has strong capabilities, both in bioinformatics and biostatistics. What can the BCF add?

- Expertise in Study Design

  -- Formulation of hypotheses

  -- Power and sample-size calculations

  -- Causal inference (experimental [RCTs] vs. observational studies, bias, confounding)

  -- Efficiency (paired vs. parallel designs, blocking, matching, covariate adjustment)

- Variance Components

  -- Biological replicates, repeated measurements, technical replicates: For example, each patient is measured at multiple time points and each assay is replicated twice. Let i, j, k denote k-th replicate assay for the i-th patient at j-th time point

$$y_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk}$$

$$V(y_{ijk}) = \sigma_a^2 + \sigma_b^2 + \sigma_\varepsilon^2$$

Typically, $\sigma_a^2 > \sigma_b^2 > \sigma_\varepsilon^2$

- Multiplicity

  -- Often a major issue in bioinformatics research

- Longitudinal Data Analysis

  -- Miixed effects regression modelling

- Time-to-Event Analysis

  -- Determine whether a genomic marker(s) is associated with length of survival

Summary

Hopefully, this brief presentation will help you, as an investigator, determine when joint collaboration with the Bioinformatics and Biostatistics Cores will be beneficial to your research.

Mengjie Chen and I will be more than happy to provide additional guidance and answer questions:

Theodore Karrison: tkarrison@health.bsd.uchicago.edu

Mengjie Chen: mchen12@bsd.uchicago.edu

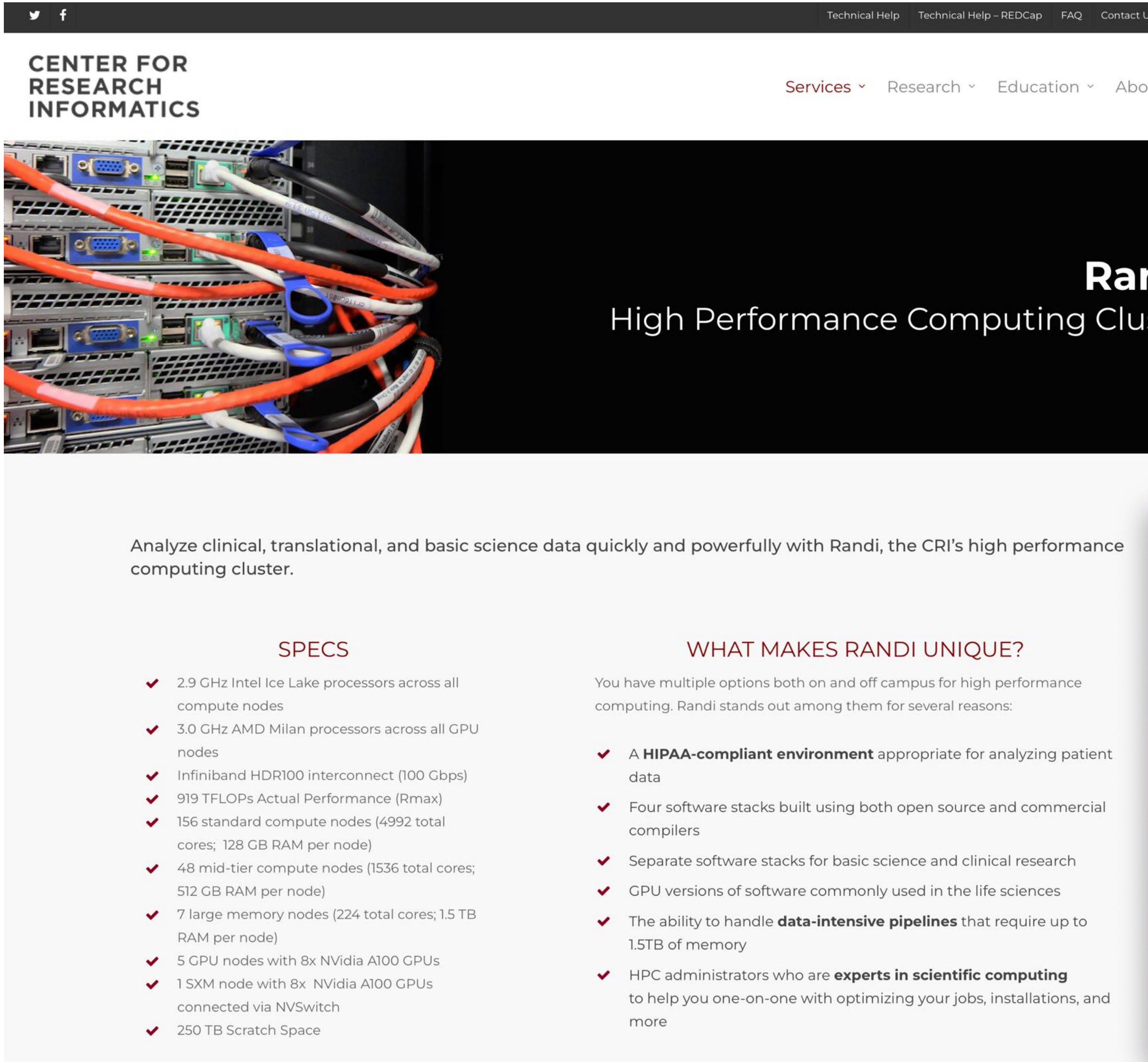
Bioinformatics & Biostatistics Core
Joint Office Hours

Consultation Service Topics:

- NGS Sequencing: Bulk, single-cell, and spatial data
- Multi-Omics Integration: Strategies for combining datasets
- Experimental Design & Power Analysis: Planning robust studies
- Grant Application Support: Guidance to strengthen proposals
- Bioinformatics & Biostatistics Collaboration: Integrated feedback on analysis and interpretation

Office Hours:
Bioinformatics Core: **Every Tuesday, 12:30 PM – 3:30 PM**
Biostatistics Core: First Tuesday of each month, **12:30 PM – 3:30 PM**
Nov 4, Dec 2, Jan 6, Feb 3, Mar 3, Apr 7

Location: **Medical Campus Peck Pavillion, N161** https://cri.uchicago.edu/hpc/



CENTER FOR RESEARCH INFORMATICS

Randi
High Performance Computing Cluster

Analyze clinical, translational, and basic science data quickly and powerfully with Randi, the CRI's high performance computing cluster.

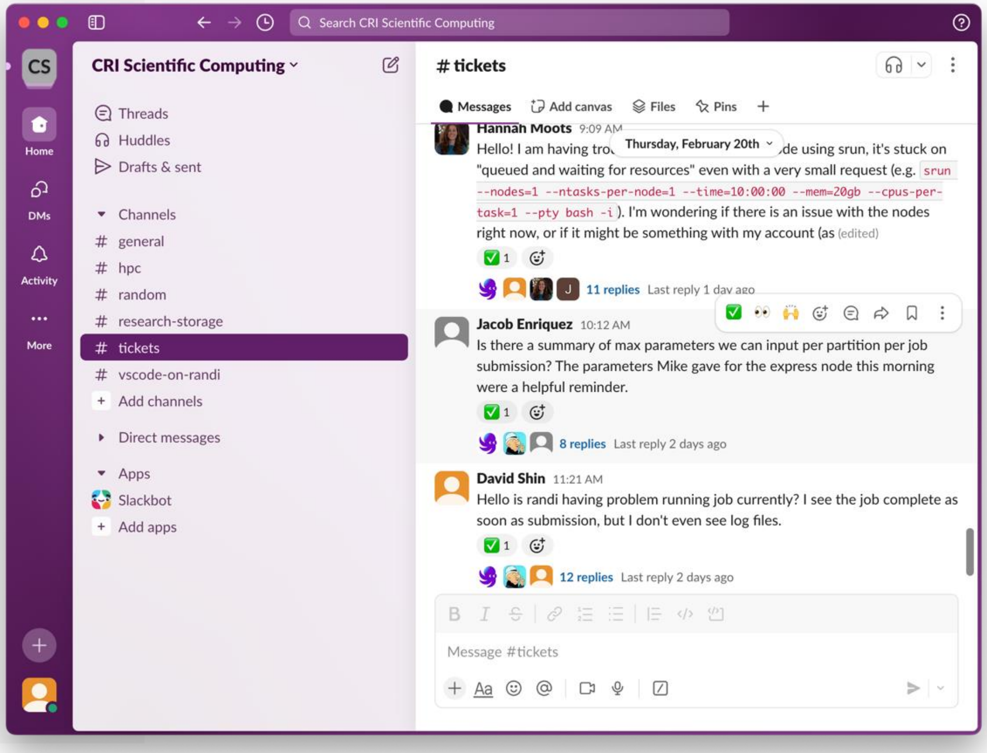
SPECS

- ✓ 2.9 GHz Intel Ice Lake processors across all compute nodes
- ✓ 3.0 GHz AMD Milan processors across all GPU nodes
- ✓ Infiniband HDR100 interconnect (100 Gbps)
- ✓ 919 TFLOPs Actual Performance (Rmax)
- ✓ 156 standard compute nodes (4992 total cores; 128 GB RAM per node)
- ✓ 48 mid-tier compute nodes (1536 total cores; 512 GB RAM per node)
- ✓ 7 large memory nodes (224 total cores; 1.5 TB RAM per node)
- ✓ 5 GPU nodes with 8x NVidia A100 GPUs
- ✓ 1 SXM node with 8x NVidia A100 GPUs connected via NVSwitch
- ✓ 250 TB Scratch Space

WHAT MAKES RANDI UNIQUE?

You have multiple options both on and off campus for high performance computing. Randi stands out among them for several reasons:

- ✓ A **HIPAA-compliant environment** appropriate for analyzing patient data
- ✓ Four software stacks built using both open source and commercial compilers
- ✓ Separate software stacks for basic science and clinical research
- ✓ GPU versions of software commonly used in the life sciences
- ✓ The ability to handle **data-intensive pipelines** that require up to 1.5TB of memory
- ✓ HPC administrators who are **experts in scientific computing** to help you one-on-one with optimizing your jobs, installations, and more



tickets

Hannah Moots 9:09 AM
Hello! I am having tro[...]de using srun, it's stuck on "queued and waiting for resources" even with a very small request (e.g. `srun --nodes=1 --ntasks-per-node=1 --time=10:00:00 --mem=20gb --cpus-per-task=1 --pty bash -i`). I'm wondering if there is an issue with the nodes right now, or if it might be something with my account (as (edited)
✓ 1
11 replies Last reply 1 day ago

Jacob Enriquez 10:12 AM
Is there a summary of max parameters we can input per partition per job submission? The parameters Mike gave for the express node this morning were a helpful reminder.
✓ 1
8 replies Last reply 2 days ago

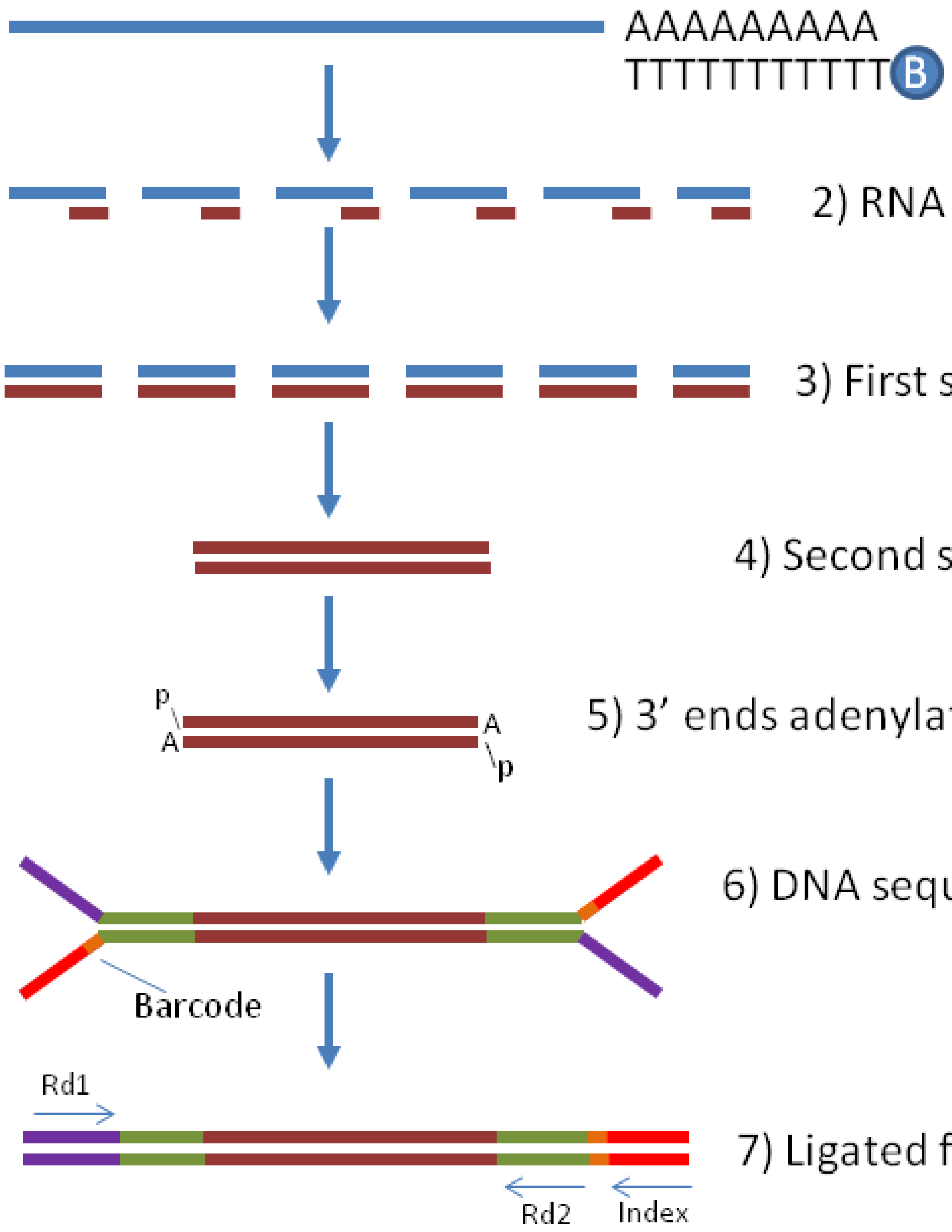
David Shin 11:21 AM
Hello is randi having problem running job currently? I see the job complete as soon as submission, but I don't even see log files.
✓ 1
12 replies Last reply 2 days ago

Objectives

- Learn to run Nextflow RNA-Seq pipeline on Randi HPC
- Learn to run our in-house app for differential expression analysis

What is Bulk RNA-Seq?

- Bulk RNA sequencing (**bulk RNA-Seq**) measures **gene expression levels** in a sample by sequencing the **total RNA from a mixture of cells**. Unlike **single-cell RNA-Seq**, bulk RNA-Seq provides an **average expression profile** across all cells in a sample.



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

https://www.labome.com/method/RNA-seq.html

Biological Questions Bulk RNA-Seq Can Answer

Our focus today

✅ **Differential Gene Expression (DGE)** → Which genes are **upregulated/downregulated** between conditions?

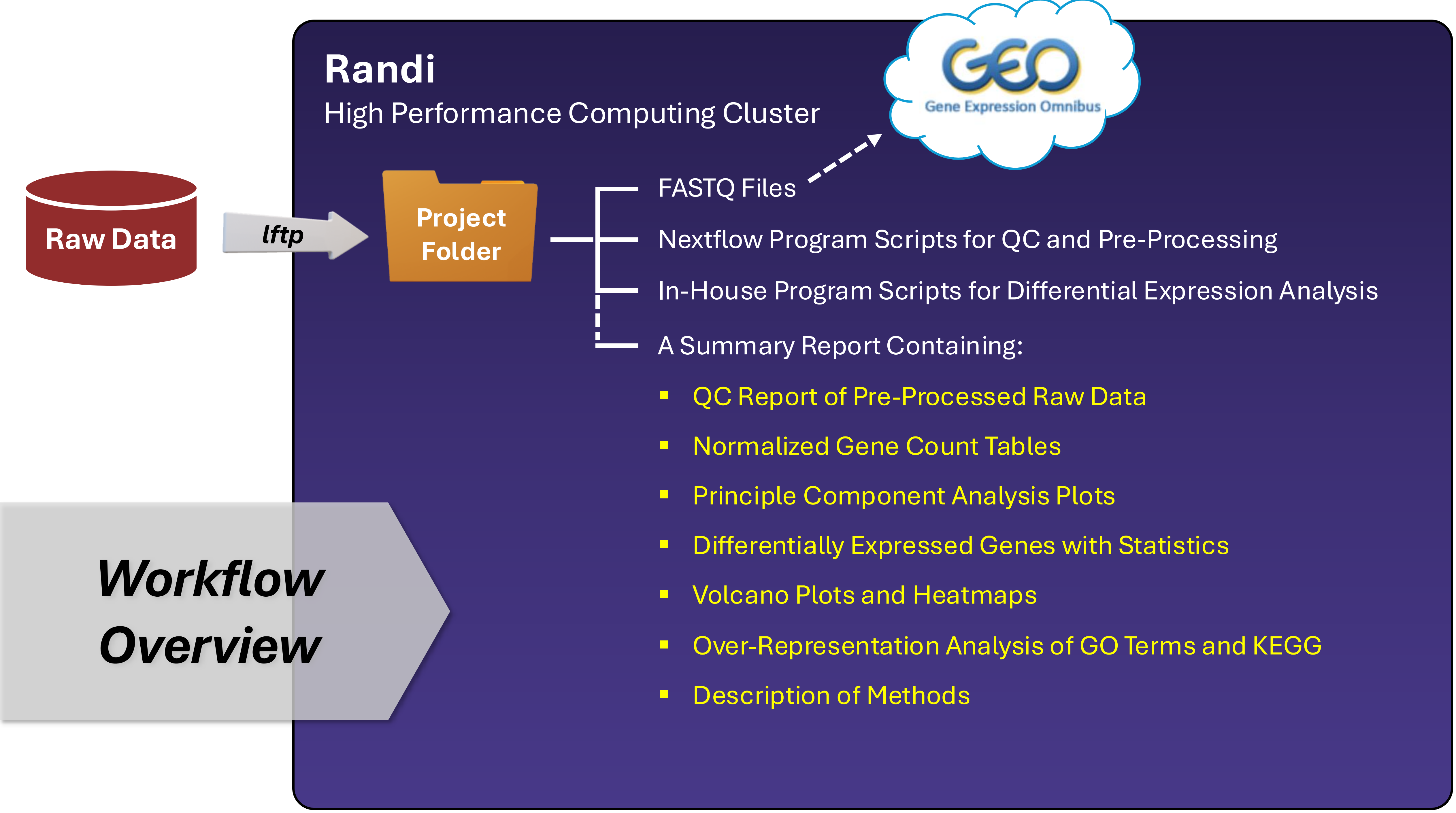
✅ **Pathway & Functional Enrichment** → What **biological processes** are affected? (e.g., gene-set over-representation analysis, GSEA)

✅ **Alternative Splicing & Isoform Analysis** → Are there changes in **splicing patterns**?

✅ **Mutation & Fusion Detection** → Are there **SNPs, RNA editing sites, or fusion transcripts**?

✅ **Cell-Type Deconvolution** → What cell types contribute to gene expression changes?

Workflow Overview



Raw Data

lftp

Randi

High Performance Computing Cluster

Project Folder

- FASTQ Files
- Nextflow Program Scripts for QC and Pre-Processing
- In-House Program Scripts for Differential Expression Analysis
- A Summary Report Containing:
  - QC Report of Pre-Processed Raw Data
  - Normalized Gene Count Tables
  - Principle Component Analysis Plots
  - Differentially Expressed Genes with Statistics
  - Volcano Plots and Heatmaps
  - Over-Representation Analysis of GO Terms and KEGG
  - Description of Methods

Agenda & Key Activities

Section 1

- Introduction to the Nextflow RNAseq Pipeline
- Hands-on Practice on Running Nextflow on the Randi Server
- Interpretation of Nextflow Outputs

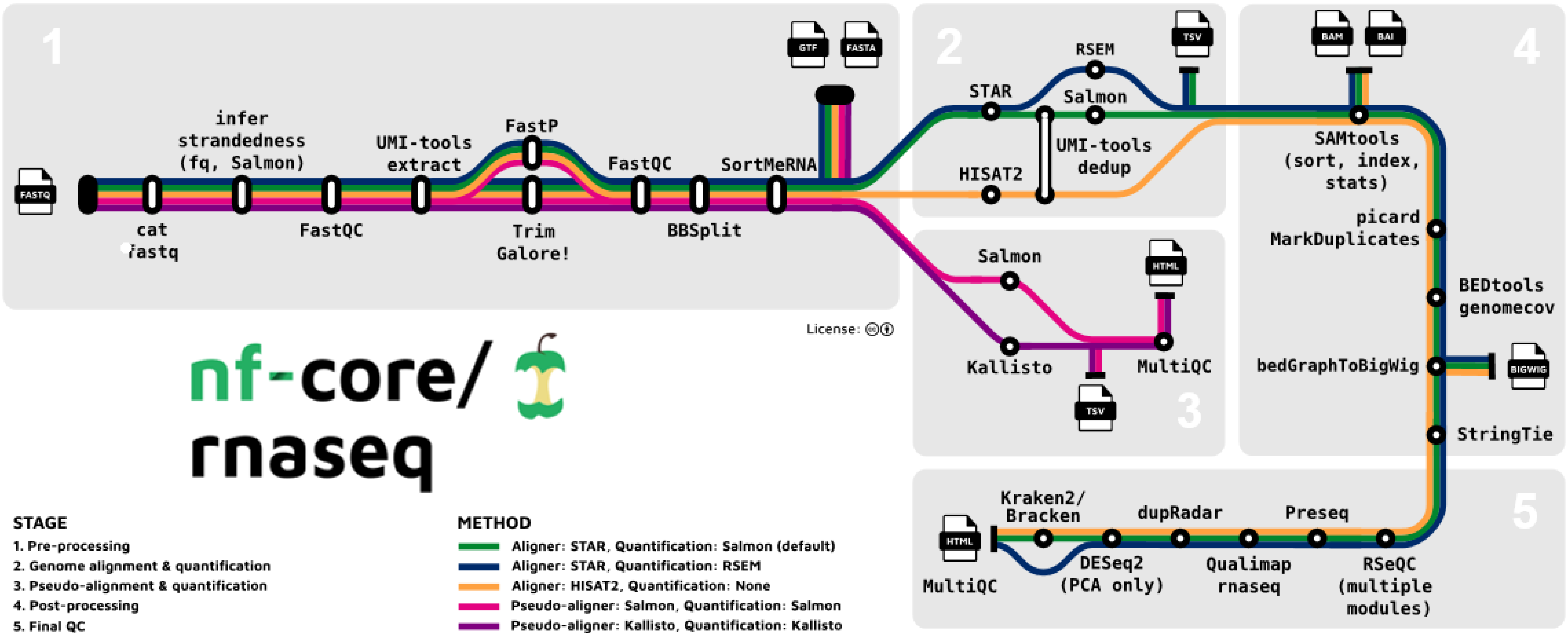
TEN-MINUTES BREAK

Section 2

- Introduction to the DE Analysis Principles
- Demo of the DE Analysis App on Randi
- Hands-on Practice on Running the DE Analysis App
- Interpretation of the DE Analysis Results

Section 1

Nextflow RNAseq Pipeline



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

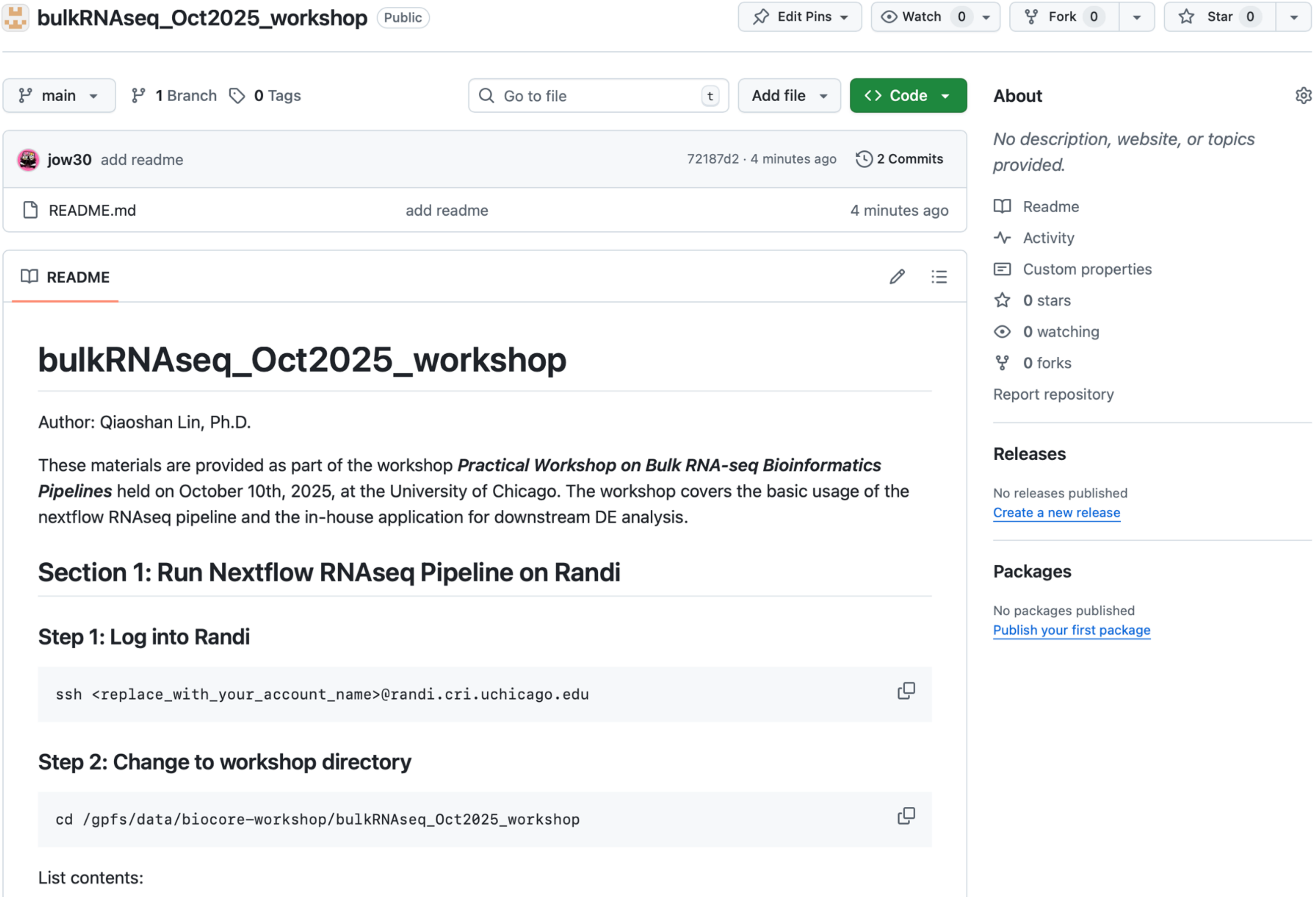
- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟥 Pseudo-aligner: Salmon, Quantification: Salmon
- 🟪 Pseudo-aligner: Kallisto, Quantification: Kallisto

1-3 hours

https://nf-co.re/rnaseq/3.18.0/

Run Nextflow RNAseq Pipeline on Randi

https://github.com/CRI-Biocore/bulkRNAseq_Oct2025_workshop



No description, website, or topics provided.

bulkRNAseq_Oct2025_workshop

Author: Qiaoshan Lin, Ph.D.

These materials are provided as part of the workshop ***Practical Workshop on Bulk RNA-seq Bioinformatics Pipelines*** held on October 10th, 2025, at the University of Chicago. The workshop covers the basic usage of the nextflow RNAseq pipeline and the in-house application for downstream DE analysis.

Section 1: Run Nextflow RNAseq Pipeline on Randi

Step 1: Log into Randi

```
ssh <replace_with_your_account_name>@randi.cri.uchicago.edu
```

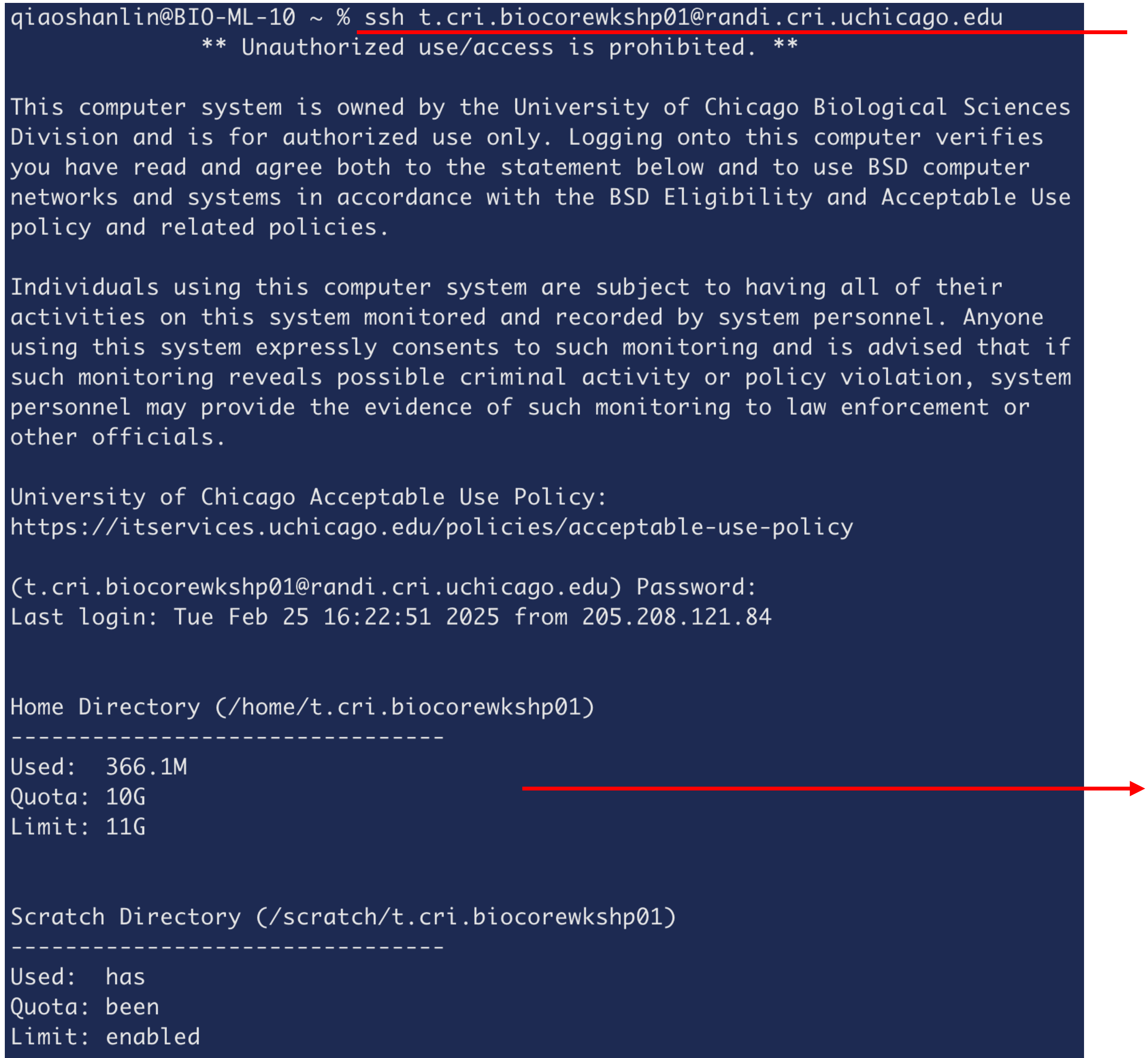
Step 2: Change to workshop directory

```
cd /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop
```

List contents:

Run Nextflow RNAseq Pipeline on Randi

Step 1: Log into Randi   Terminal / iTerm (MacOS)   PuTTY / Xshell (Windows)



```
qiaoshanlin@BIO-ML-10 ~ % ssh t.cri.biocorewkshp01@randi.cri.uchicago.edu
              ** Unauthorized use/access is prohibited. **

This computer system is owned by the University of Chicago Biological Sciences
Division and is for authorized use only. Logging onto this computer verifies
you have read and agree both to the statement below and to use BSD computer
networks and systems in accordance with the BSD Eligibility and Acceptable Use
policy and related policies.

Individuals using this computer system are subject to having all of their
activities on this system monitored and recorded by system personnel. Anyone
using this system expressly consents to such monitoring and is advised that if
such monitoring reveals possible criminal activity or policy violation, system
personnel may provide the evidence of such monitoring to law enforcement or
other officials.

University of Chicago Acceptable Use Policy:
https://itservices.uchicago.edu/policies/acceptable-use-policy

(t.cri.biocorewkshp01@randi.cri.uchicago.edu) Password:
Last login: Tue Feb 25 16:22:51 2025 from 205.208.121.84


Home Directory (/home/t.cri.biocorewkshp01)
--------------------------------
Used:  366.1M
Quota: 10G
Limit: 11G


Scratch Directory (/scratch/t.cri.biocorewkshp01)
--------------------------------
Used:  has
Quota: been
Limit: enabled
```

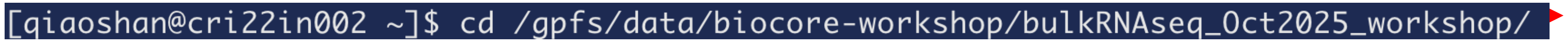
ssh your_account_name@randi.cri.uchicago.edu

Enter password and you will be logged in

Every account has a 11G limit in the home directory

Run Nextflow RNAseq Pipeline on Randi

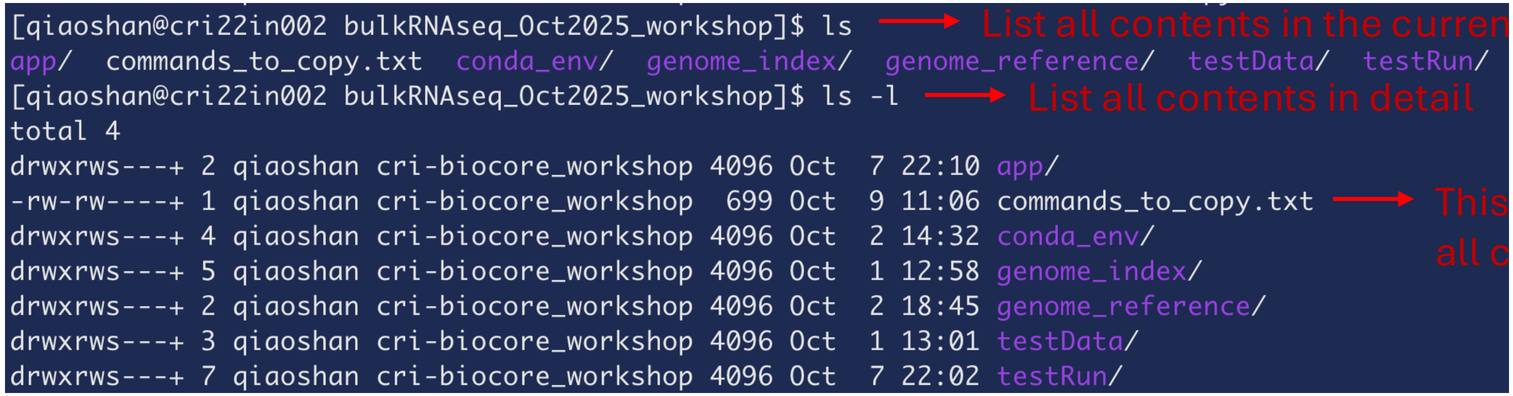
Step 2: Change to workshop directory



```
[qiaoshan@cri22in002 ~]$ cd /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/
```

Change to the directory of **BulkRNAseq workshop**

Please find all workshop materials inside this folder



```
[qiaoshan@cri22in002 bulkRNAseq_Oct2025_workshop]$ ls
app/  commands_to_copy.txt  conda_env/  genome_index/  genome_reference/  testData/  testRun/
[qiaoshan@cri22in002 bulkRNAseq_Oct2025_workshop]$ ls -l
total 4
drwxrws---+ 2 qiaoshan cri-biocore_workshop 4096 Oct  7 22:10 app/
-rw-rw----+ 1 qiaoshan cri-biocore_workshop  699 Oct  9 11:06 commands_to_copy.txt
drwxrws---+ 4 qiaoshan cri-biocore_workshop 4096 Oct  2 14:32 conda_env/
drwxrws---+ 5 qiaoshan cri-biocore_workshop 4096 Oct  1 12:58 genome_index/
drwxrws---+ 2 qiaoshan cri-biocore_workshop 4096 Oct  2 18:45 genome_reference/
drwxrws---+ 3 qiaoshan cri-biocore_workshop 4096 Oct  1 13:01 testData/
drwxrws---+ 7 qiaoshan cri-biocore_workshop 4096 Oct  7 22:02 testRun/
```

List all contents in the current directory

List all contents in detail

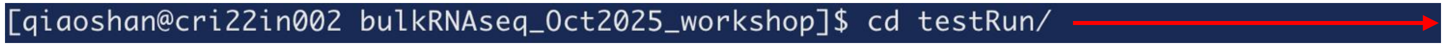
This is the file where you can copy and paste all commands to use for this workshop

What is in all these folders?

app: source codes of the downstream in-house app for DE analysis
conda_env: conda environment files (including all compiled R packages)
genome_index: preprocessed reference genome that allows fast searching, alignment, or quantification of sequencing reads against it.
genome_reference: genome reference files
testData: raw fastq files
testRun: where you will keep all your scripts and results

Run Nextflow RNAseq Pipeline on Randi

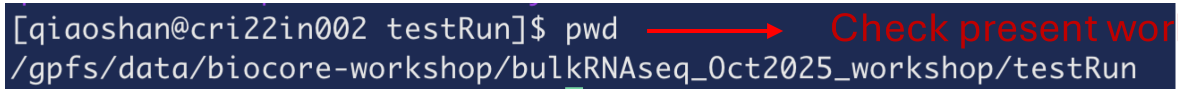
Step 3: Build your working directory



```
[qiaoshan@cri22in002 bulkRNAseq_Oct2025_workshop]$ cd testRun/
```

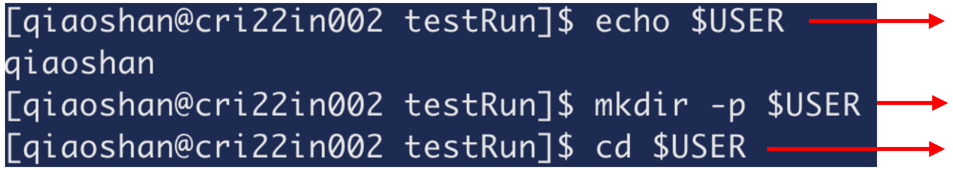
Enter the testRun folder

Note: Please do NOT change anything outside the testRun folder



```
[qiaoshan@cri22in002 testRun]$ pwd
/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun
```

Check present working directory to confirm you'r inside the testRun folder



```
[qiaoshan@cri22in002 testRun]$ echo $USER
qiaoshan
[qiaoshan@cri22in002 testRun]$ mkdir -p $USER
[qiaoshan@cri22in002 testRun]$ cd $USER
```

Check your username

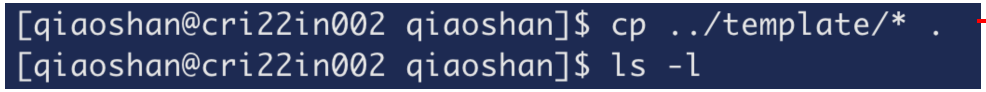
Make a folder inside testRun with your username

Enter your folder (This is your working directory)

Note: Please do NOT change or write anything outside your working directory

Run Nextflow RNAseq Pipeline on Randi

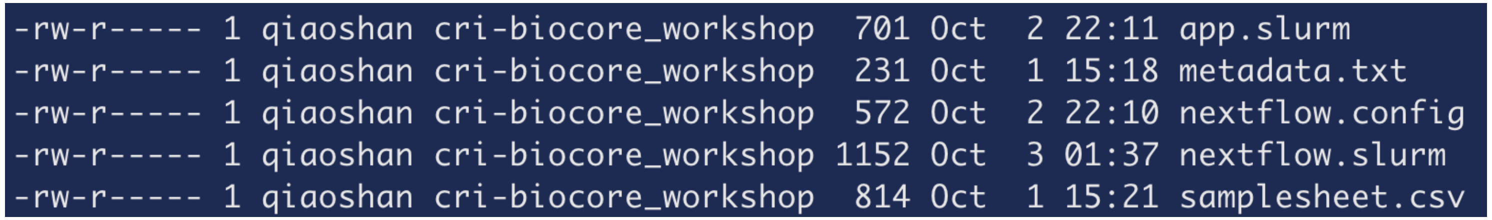
Step 4: Prepare scripts and configurations



```
[qiaoshan@cri22in002 qiaoshan]$ cp ../template/* .
[qiaoshan@cri22in002 qiaoshan]$ ls -l
```

Copy all scripts from template

There should be five files:



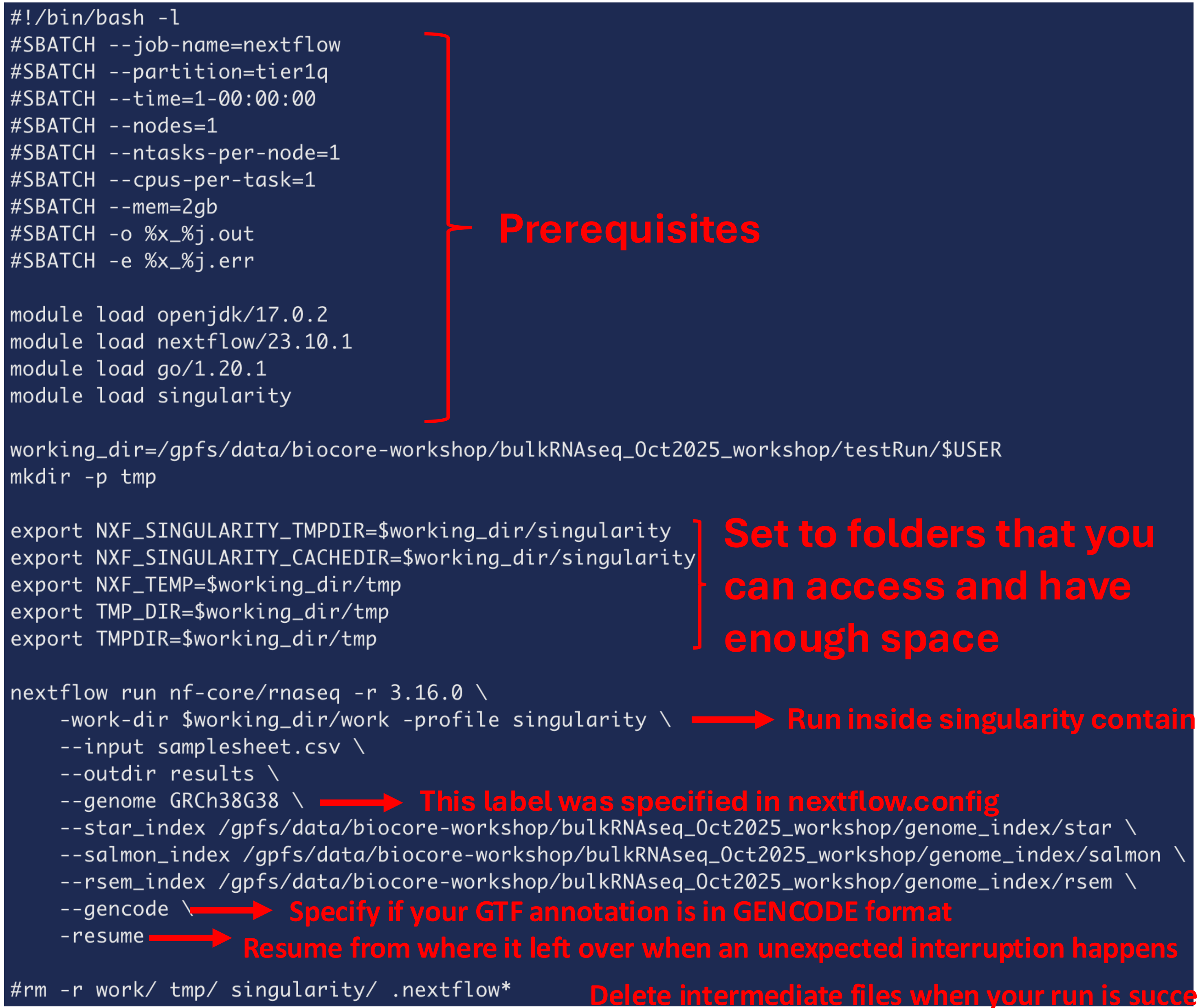
```
-rw-r----- 1 qiaoshan cri-biocore_workshop  701 Oct  2 22:11 app.slurm
-rw-r----- 1 qiaoshan cri-biocore_workshop  231 Oct  1 15:18 metadata.txt
-rw-r----- 1 qiaoshan cri-biocore_workshop  572 Oct  2 22:10 nextflow.config
-rw-r----- 1 qiaoshan cri-biocore_workshop 1152 Oct  3 01:37 nextflow.slurm
-rw-r----- 1 qiaoshan cri-biocore_workshop  814 Oct  1 15:21 samplesheet.csv
```

What does each file do?

samplesheet.csv: This is a table containing raw fastq file locations for each sample so that Nextflow knows where to read the files.
nextflow.config: This is a configuration file for Nextflow to run on Randi.
nextflow.slurm: This is the job script to submit to run the Nextflow pipeline.

metadata.txt: This is a table containing sample information like condition, phenotype, batch, etc.
app.slurm: This is the job script to submit to run the downstream in-house application for DE analysis.

Key Parameters



```
#!/bin/bash -l
#SBATCH --job-name=nextflow
#SBATCH --partition=tier1q
#SBATCH --time=1-00:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=2gb
#SBATCH -o %x_%j.out
#SBATCH -e %x_%j.err

module load openjdk/17.0.2
module load nextflow/23.10.1
module load go/1.20.1
module load singularity

working_dir=/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/$USER
mkdir -p tmp

export NXF_SINGULARITY_TMPDIR=$working_dir/singularity
export NXF_SINGULARITY_CACHEDIR=$working_dir/singularity
export NXF_TEMP=$working_dir/tmp
export TMP_DIR=$working_dir/tmp
export TMPDIR=$working_dir/tmp

nextflow run nf-core/rnaseq -r 3.16.0 \
    -work-dir $working_dir/work -profile singularity \
    --input samplesheet.csv \
    --outdir results \
    --genome GRCh38G38 \
    --star_index /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/genome_index/star \
    --salmon_index /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/genome_index/salmon \
    --rsem_index /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/genome_index/rsem \
    --gencode \
    -resume

#rm -r work/ tmp/ singularity/ .nextflow*
```

Prerequisites

Set to folders that you can access and have enough space

Run inside singularity container

This label was specified in nextflow.config

Use pre-built index to save time

Specify if your GTF annotation is in GENCODE format

Resume from where it left over when an unexpected interruption happens

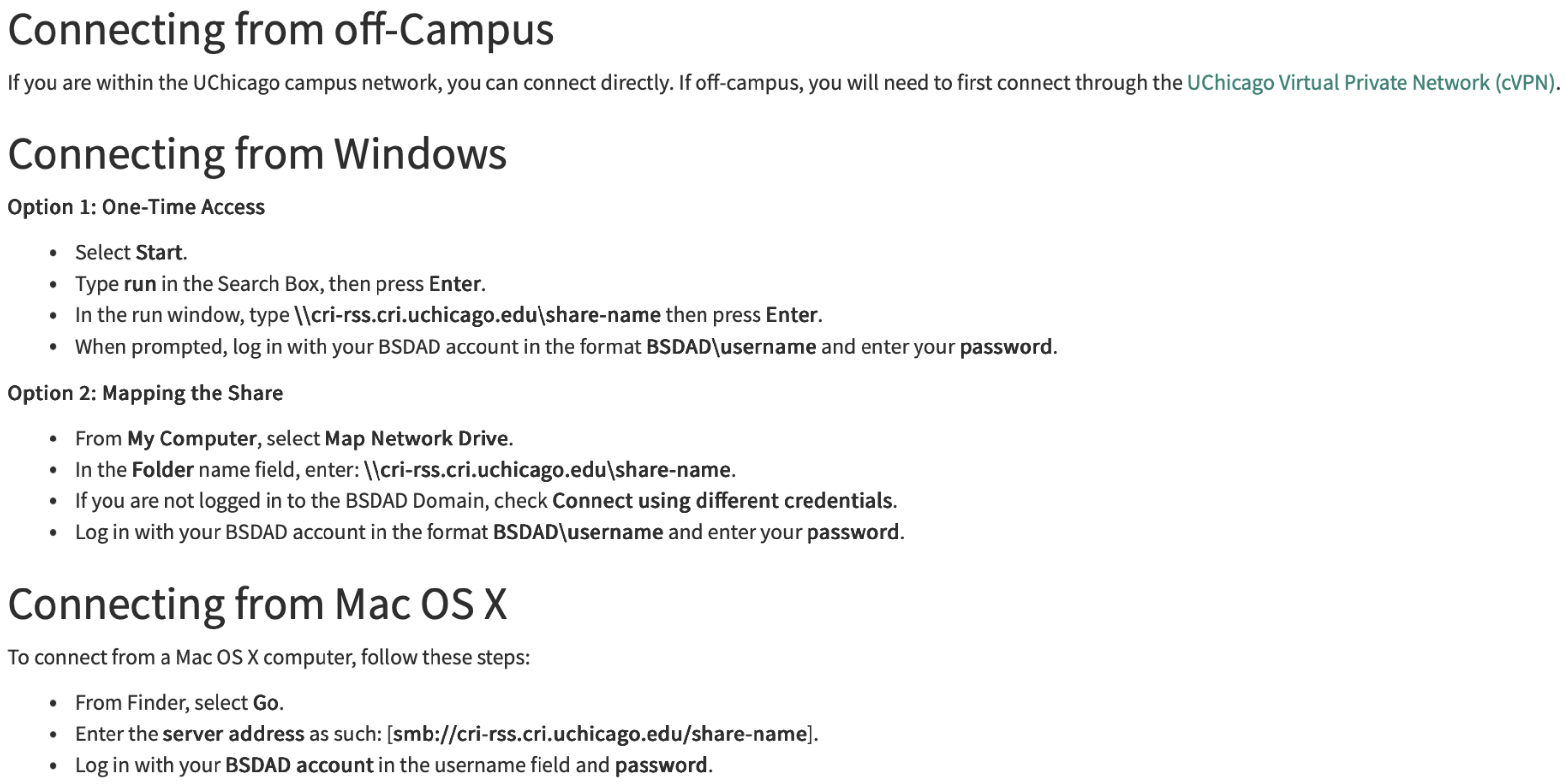
Delete intermediate files when your run is successfully completed.

Run Nextflow RNAseq Pipeline on Randi

Step 4: Prepare scripts and configurations

How to edit the sample files, parameters, and metadata files?

https://uchicago.service-now.com/kb_view.do?sys_kb_id=27ffbebb97fc025423087cf11153af5b



Connecting from off-Campus

If you are within the UChicago campus network, you can connect directly. If off-campus, you will need to first connect through the UChicago Virtual Private Network (cVPN).

Connecting from Windows

Option 1: One-Time Access

- Select **Start**.
- Type **run** in the Search Box, then press **Enter**.
- In the run window, type **\\cri-rss.cri.uchicago.edu\share-name** then press **Enter**.
- When prompted, log in with your BSDAD account in the format **BSDAD\username** and enter your **password**.

Option 2: Mapping the Share

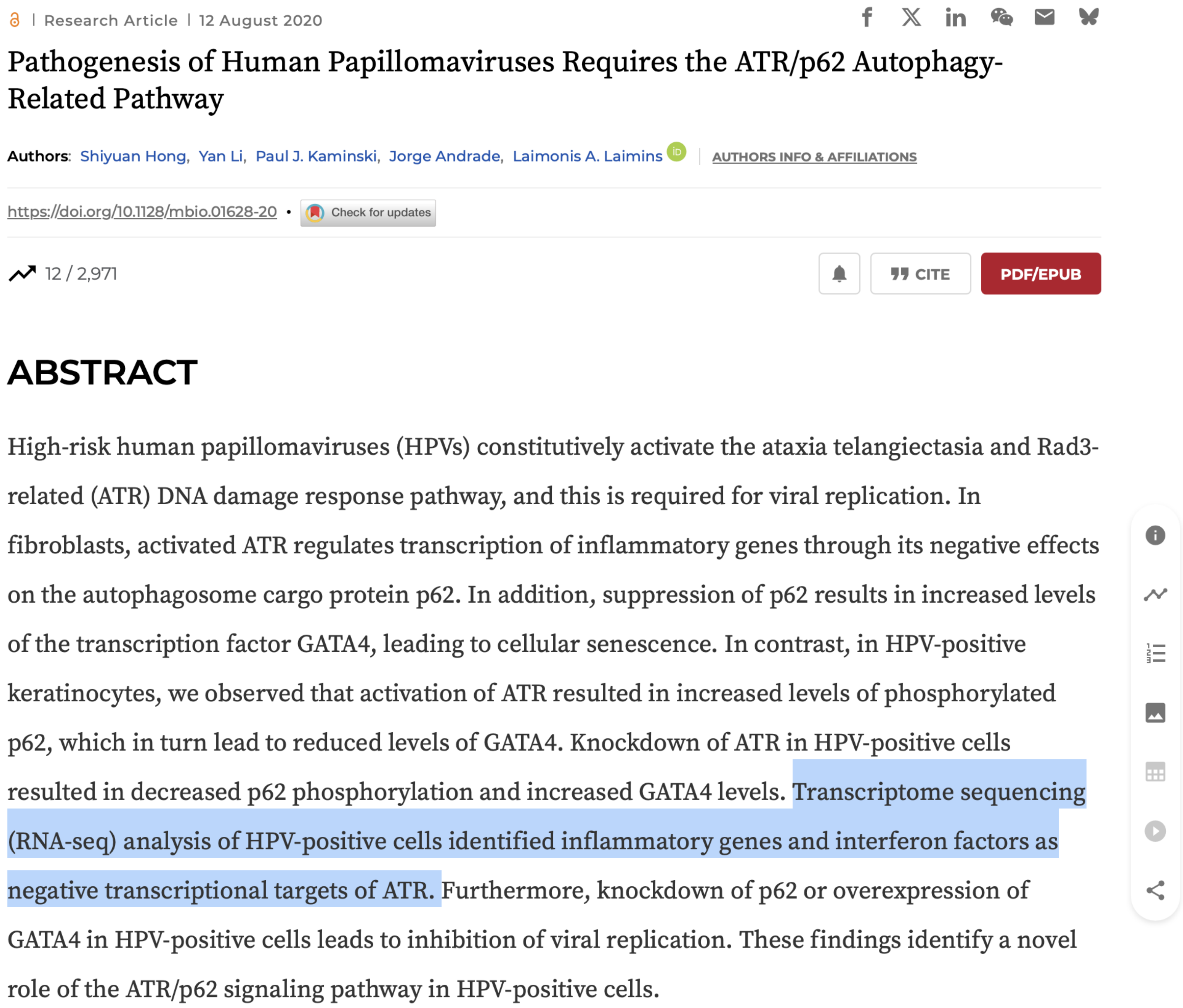
- From **My Computer**, select **Map Network Drive**.
- In the **Folder** name field, enter: **\\cri-rss.cri.uchicago.edu\share-name**.
- If you are not logged in to the BSDAD Domain, check **Connect using different credentials**.
- Log in with your BSDAD account in the format **BSDAD\username** and enter your **password**.

Connecting from Mac OS X

To connect from a Mac OS X computer, follow these steps:

- From Finder, select **Go**.
- Enter the **server address** as such: [**smb://cri-rss.cri.uchicago.edu/share-name**].
- Log in with your **BSDAD account** in the username field and **password**.

Test Data (GSE142482)



Research Article | 12 August 2020
12 / 2,971
CITE
PDF/EPUB
AUTHORS INFO & AFFILIATIONS
Check for updates

Pathogenesis of Human Papillomaviruses Requires the ATR/p62 Autophagy-Related Pathway

Authors: Shiyuan Hong, Yan Li, Paul J. Kaminski, Jorge Andrade, Laimonis A. Laimins

https://doi.org/10.1128/mbio.01628-20

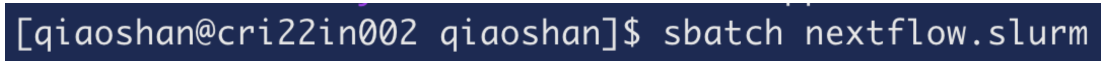
ABSTRACT

High-risk human papillomaviruses (HPVs) constitutively activate the ataxia telangiectasia and Rad3-related (ATR) DNA damage response pathway, and this is required for viral replication. In fibroblasts, activated ATR regulates transcription of inflammatory genes through its negative effects on the autophagosome cargo protein p62. In addition, suppression of p62 results in increased levels of the transcription factor GATA4, leading to cellular senescence. In contrast, in HPV-positive keratinocytes, we observed that activation of ATR resulted in increased levels of phosphorylated p62, which in turn lead to reduced levels of GATA4. Knockdown of ATR in HPV-positive cells resulted in decreased p62 phosphorylation and increased GATA4 levels. Transcriptome sequencing (RNA-seq) analysis of HPV-positive cells identified inflammatory genes and interferon factors as negative transcriptional targets of ATR. Furthermore, knockdown of p62 or overexpression of GATA4 in HPV-positive cells leads to inhibition of viral replication. These findings identify a novel role of the ATR/p62 signaling pathway in HPV-positive cells.

Hong SLi Y, Kaminski PJ, Andrade J, Laimins LA.2020.Pathogenesis of Human Papillomaviruses Requires the ATR/p62 Autophagy-Related Pathway. mBio11:10.1128/mbio.01628-20.https://doi.org/10.1128/mbio.01628-20

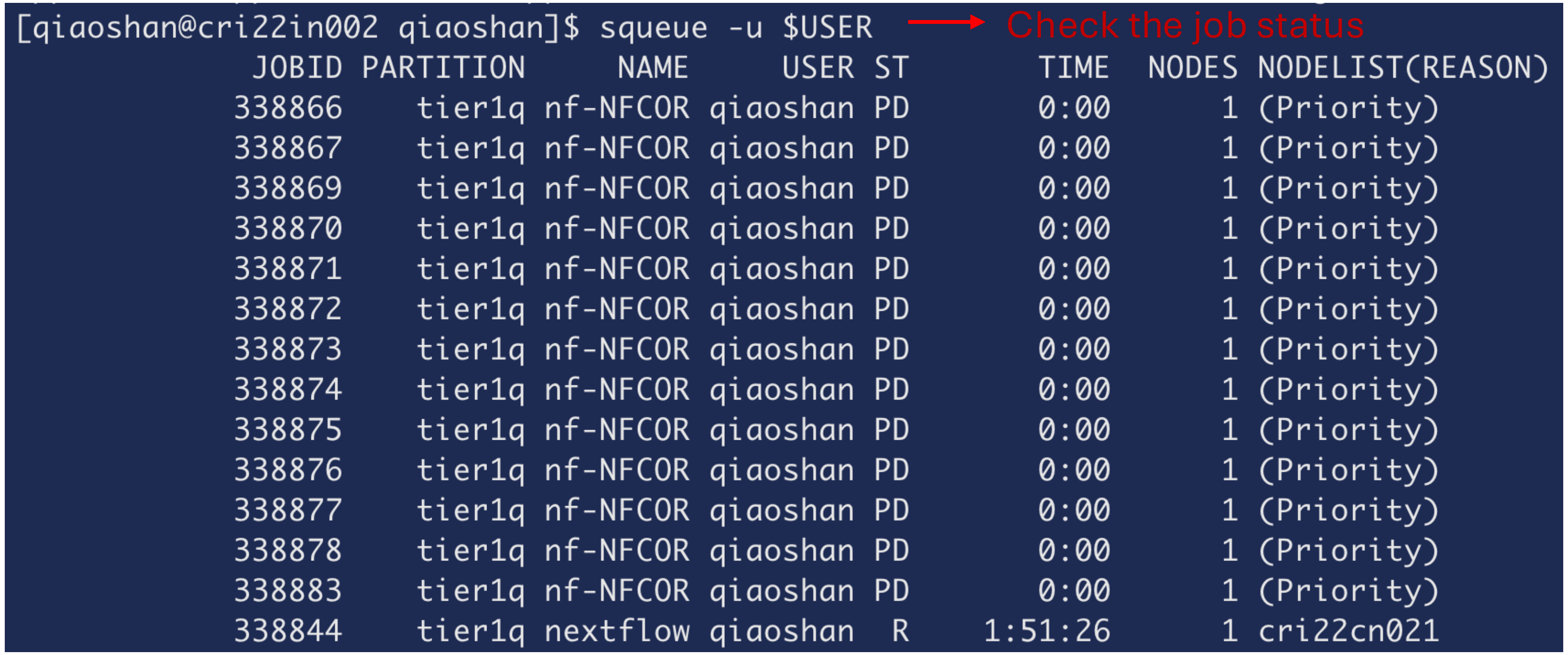
Run Nextflow RNAseq Pipeline on Randi

Step 5: Submit the job



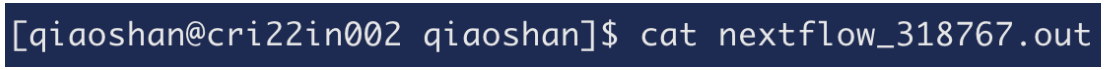
```
[qiaoshan@cri22in002 qiaoshan]$ sbatch nextflow.slurm
```

Submit the job to run Nextflow pipeline



```
[qiaoshan@cri22in002 qiaoshan]$ squeue -u $USER
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            338866    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338867    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338869    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338870    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338871    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338872    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338873    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338874    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338875    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338876    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338877    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338878    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338883    tier1q nf-NFCOR qiaoshan PD       0:00      1 (Priority)
            338844    tier1q nextflow qiaoshan  R    1:51:26      1 cri22cn021
```

Check the job status

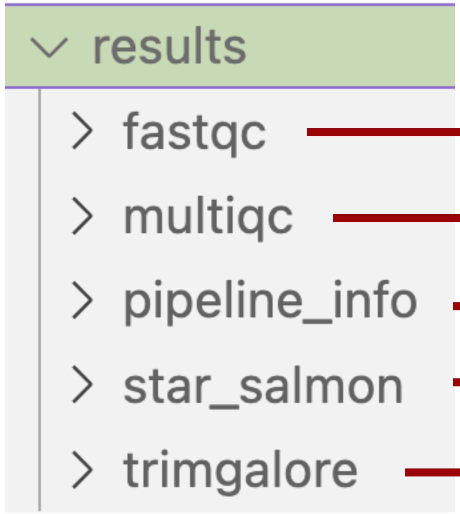


```
[qiaoshan@cri22in002 qiaoshan]$ cat nextflow_318767.out
```

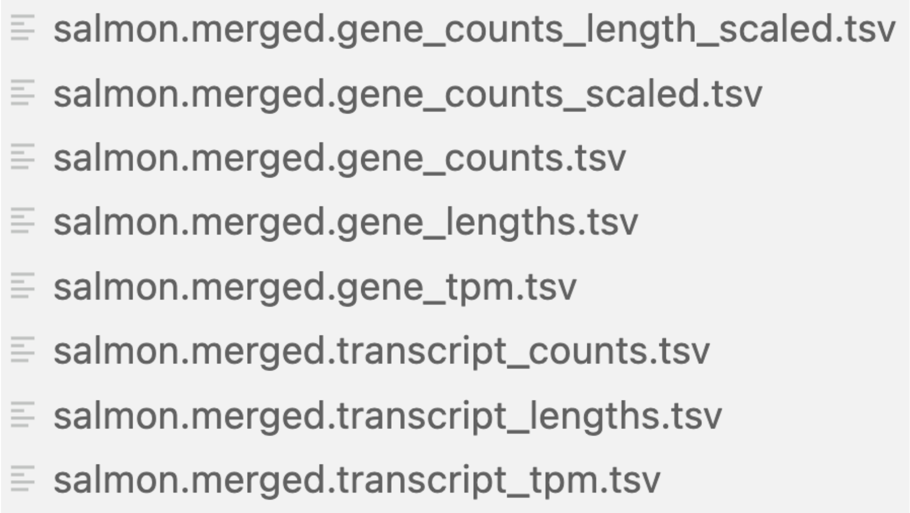
Check the job status details

Note: The job could take several hours to finish, depending on node availability on the server.

Interpretation of Nextflow Outputs



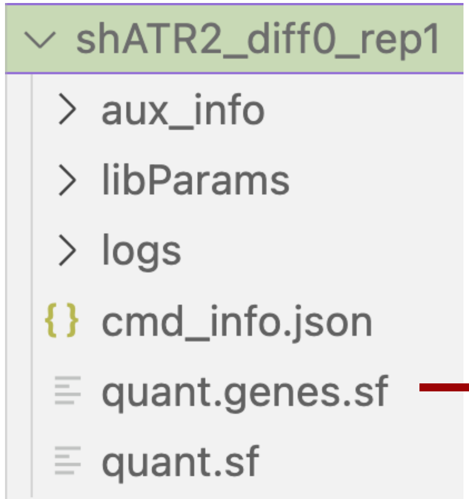
results
- fastqc → FastQC reports of raw reads and trimmed reads
- multiqc → *First to check* One report with QC of each step integrated in one place
- pipeline_info → Commands for each step & Environment setup
- star_salmon → Filtered bam files & **Gene counts**
- trimgalore → Read trimming reports



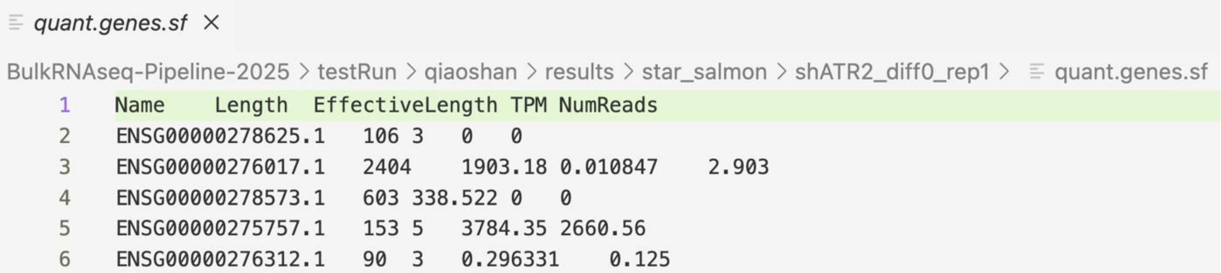
- salmon.merged.gene_counts_length_scaled.tsv
- salmon.merged.gene_counts_scaled.tsv
- salmon.merged.gene_counts.tsv
- salmon.merged.gene_lengths.tsv
- salmon.merged.gene_tpm.tsv
- salmon.merged.transcript_counts.tsv
- salmon.merged.transcript_lengths.tsv
- salmon.merged.transcript_tpm.tsv

Each sample has a salmon output folder with *.sf files.

I recommend using *.sf files for the downstream to take advantage of the bias correction by Salmon



shATR2_diff0_rep1
- aux_info
- libParams
- logs
- cmd_info.json
- quant.genes.sf
- quant.sf



quant.genes.sf ×

BulkRNAseq-Pipeline-2025 > testRun > qiaoshan > results > star_salmon > shATR2_diff0_rep1 > quant.genes.sf

```
Name     Length  EffectiveLength TPM NumReads
ENSG00000278625.1    106 3    0    0
ENSG00000276017.1    2404     1903.18 0.010847     2.903
ENSG00000278573.1    603 338.522 0    0
ENSG00000275757.1    153 5    3784.35 2660.56
ENSG00000276312.1    90   3    0.296331     0.125
```

Agenda & Key Activities

Section 1

- Introduction to the Nextflow RNAseq Pipeline
- Hands-on Practice on Running Nextflow on the Randi Server
- Interpretation of Nextflow Outputs

TEN-MINUTES BREAK

Section 2

- Introduction to the DE Analysis Principles
- Demo of the DE Analysis App on Randi
- Hands-on Practice on Running the DE Analysis App
- Interpretation of the DE Analysis Results

Section 2

DE Analysis

Key Considerations in DE Analysis

⚠ **Batch Effects** → Use PCA/MDS to check for unwanted variation

⚠ **Replicates Matter** → More replicates = higher statistical power (highly recommend >=3 rep)

⚠ **Read Depth** → Sufficient sequencing depth (>25M for human/mouse)

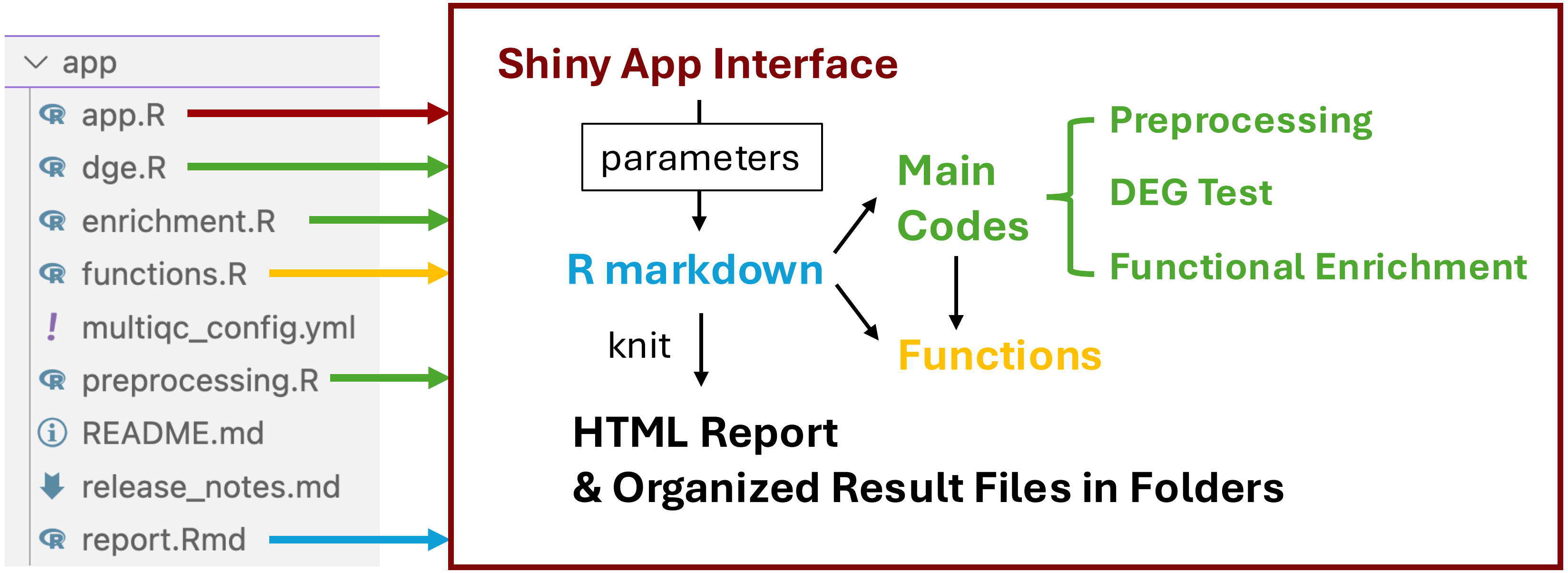
⚠ **Data Distribution** → RNA-seq data is often over-dispersed (use a model like Negative Binomial)

⚠ **Normalization** → Correct for library size & sequencing depth

⚠ **Multiple Testing Correction** → Use adjusted p-values to control false positives

In-House Downstream Analysis

Architecture



app
- app.R
- dge.R
- enrichment.R
- functions.R
- multiqc_config.yml
- preprocessing.R
- README.md
- release_notes.md
- report.Rmd

Shiny App Interface

parameters

R markdown

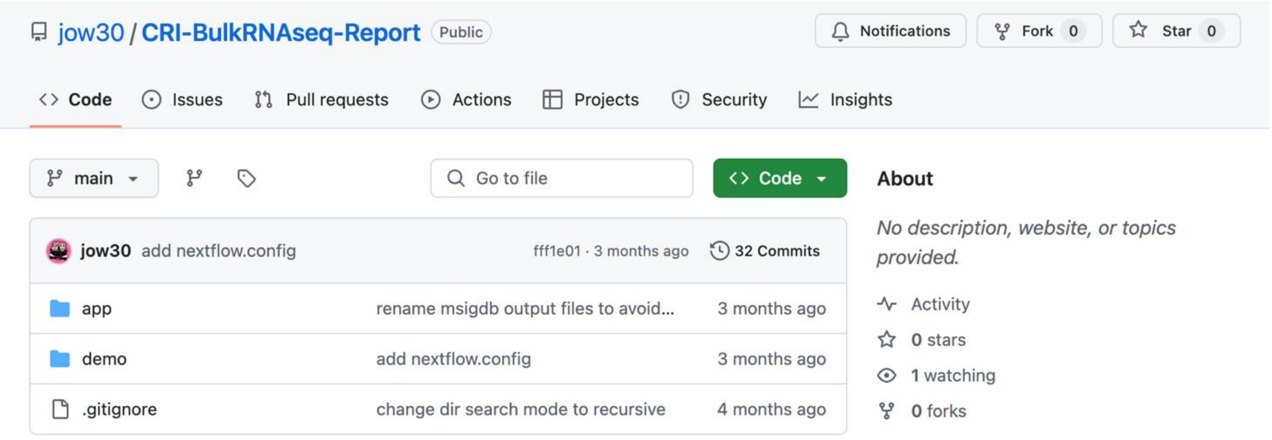
Main Codes

- Preprocessing
- DEG Test
- Functional Enrichment

Functions

knit

HTML Report
& Organized Result Files in Folders



jow30 / CRI-BulkRNAseq-Report Public

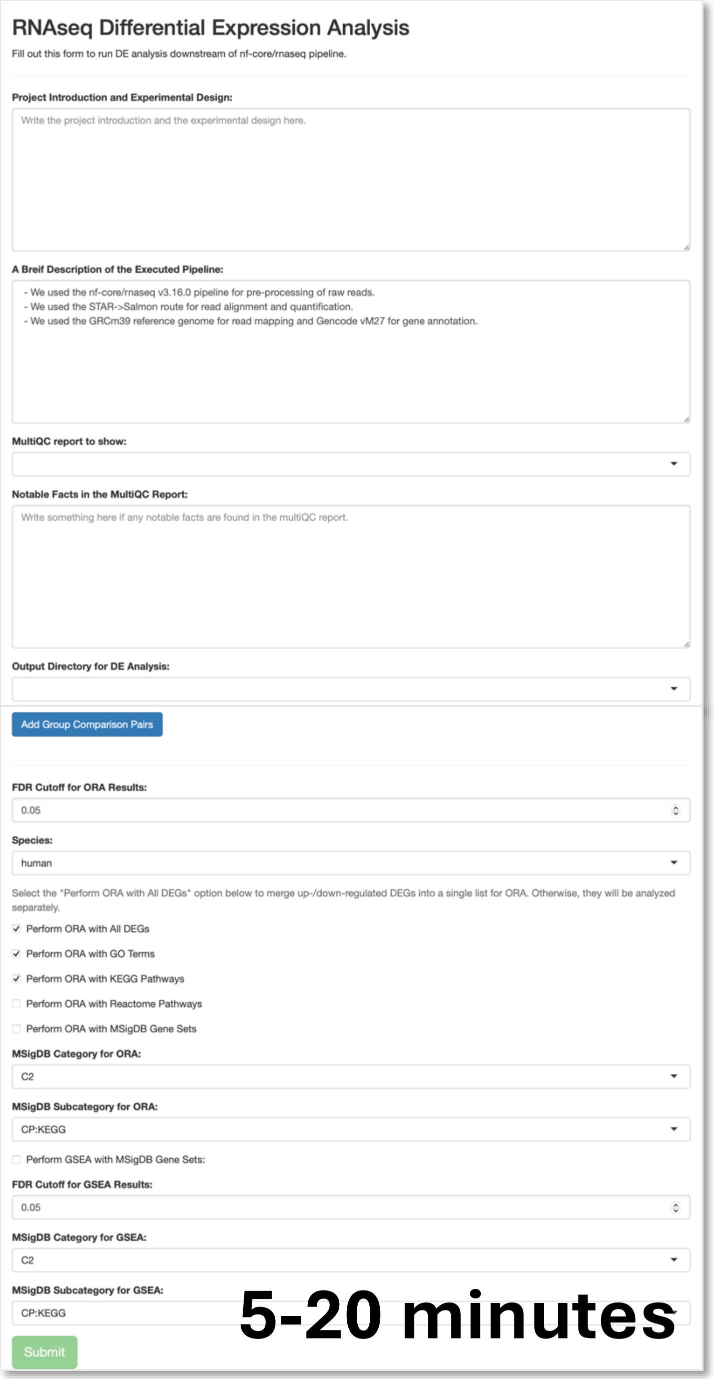
jow30 add nextflow.config fff1e01 · 3 months ago 32 Commits

app	rename msigdb output files to avoid...	3 months ago
demo	add nextflow.config	3 months ago
.gitignore	change dir search mode to recursive	4 months ago

About

No description, website, or topics provided.

- Activity
- 0 stars
- 1 watching
- 0 forks



RNAseq Differential Expression Analysis

Fill out this form to run DE analysis downstream of nf-core/rnaseq pipeline.

Project Introduction and Experimental Design:

Write the project introduction and the experimental design here.

A Breif Description of the Executed Pipeline:

- We used the nf-core/rnaseq v3.16.0 pipeline for pre-processing of raw reads.
- We used the STAR->Salmon route for read alignment and quantification.
- We used the GRCm39 reference genome for read mapping and Gencode vM27 for gene annotation.

MultiQC report to show:

Notable Facts in the MultiQC Report:

Write something here if any notable facts are found in the multiQC report.

Output Directory for DE Analysis:

FDR Cutoff for ORA Results: 0.05

Species: human

Select the "Perform ORA with All DEGs" option below to merge up-/down-regulated DEGs into a single list for ORA. Otherwise, they will be analyzed separately.

☑ Perform ORA with All DEGs

☑ Perform ORA with GO Terms

☑ Perform ORA with KEGG Pathways

□ Perform ORA with Reactome Pathways

□ Perform ORA with MSigDB Gene Sets

MSigDB Category for ORA: C2

MSigDB Subcategory for ORA: CP:KEGG

□ Perform GSEA with MSigDB Gene Sets:

FDR Cutoff for GSEA Results: 0.05

MSigDB Category for GSEA: C2

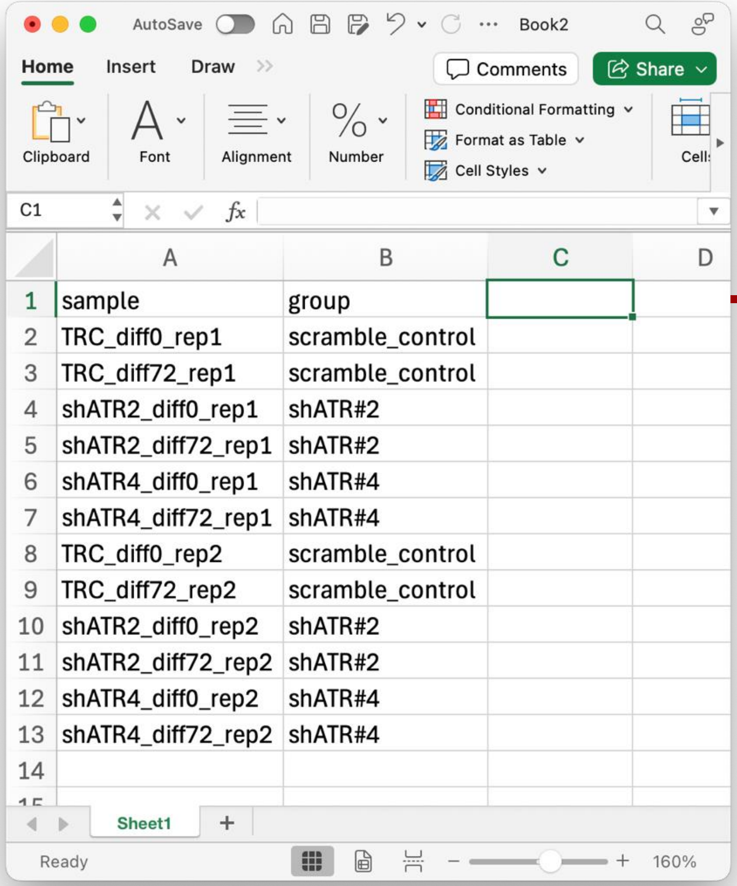
MSigDB Subcategory for GSEA: CP:KEGG

5-20 minutes

In-House Downstream Analysis

Step 1: Setup experimental groups

1. Edit the table in Excel



	A	B	C	D
1	sample	group		
2	TRC_diff0_rep1	scramble_control		
3	TRC_diff72_rep1	scramble_control		
4	shATR2_diff0_rep1	shATR#2		
5	shATR2_diff72_rep1	shATR#2		
6	shATR4_diff0_rep1	shATR#4		
7	shATR4_diff72_rep1	shATR#4		
8	TRC_diff0_rep2	scramble_control		
9	TRC_diff72_rep2	scramble_control		
10	shATR2_diff0_rep2	shATR#2		
11	shATR2_diff72_rep2	shATR#2		
12	shATR4_diff0_rep2	shATR#4		
13	shATR4_diff72_rep2	shATR#4		
14				

2. Save the table in the txt or csv format

TXT
metadata

3. Upload to Randi working directory

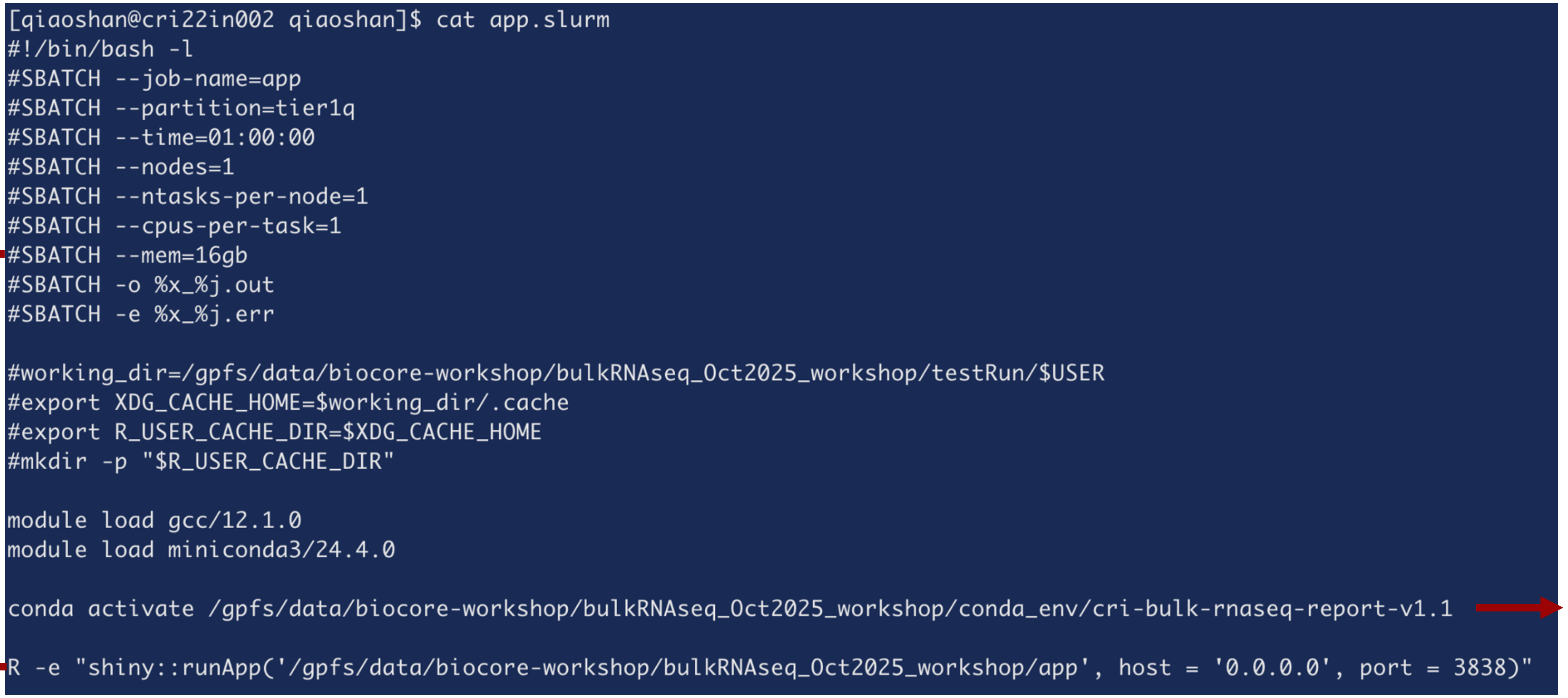
Note: metadata.txt has been set up for you already in this workshop so you don't need to edit it to run the test data.

The header must contain *sample* and *group*. You can add as many experimental factors as you want to columns.

If a batch effect needs to be corrected, add a *batch* column so that the batch-effect-removal option can be enabled.

In-House Downstream Analysis

Step 2: Run the slurm script



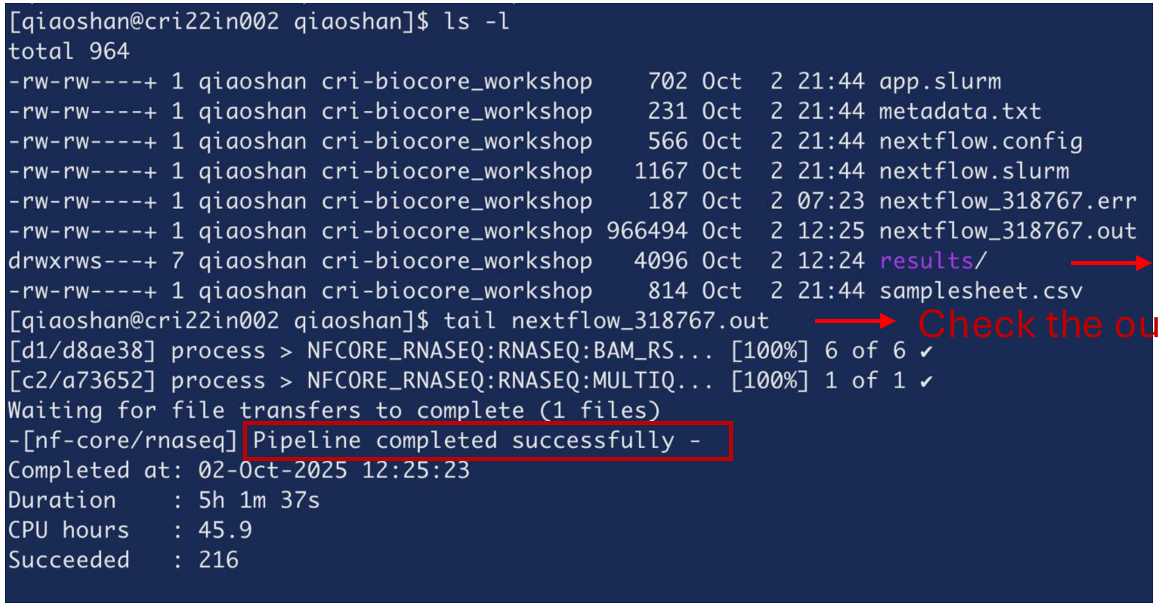
```
[qiaoshan@cri22in002 qiaoshan]$ cat app.slurm
#!/bin/bash -l
#SBATCH --job-name=app
#SBATCH --partition=tier1q
#SBATCH --time=01:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=16gb
#SBATCH -o %x_%j.out
#SBATCH -e %x_%j.err

#working_dir=/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/$USER
#export XDG_CACHE_HOME=$working_dir/.cache
#export R_USER_CACHE_DIR=$XDG_CACHE_HOME
#mkdir -p "$R_USER_CACHE_DIR"

module load gcc/12.1.0
module load miniconda3/24.4.0

conda activate /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/conda_env/cri-bulk-rnaseq-report-v1.1

R -e "shiny::runApp('/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/app', host = '0.0.0.0', port = 3838)"
```

Increase mem if data is large

Enter the conda environment

Run R shiny app

Run Nextflow RNAseq Pipeline on Randi

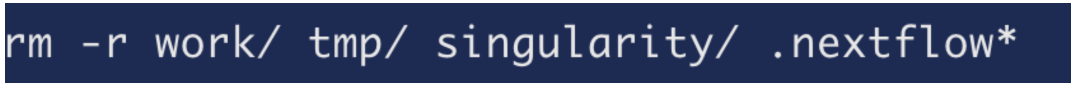


```
[qiaoshan@cri22in002 qiaoshan]$ ls -l
total 964
-rw-rw----+ 1 qiaoshan cri-biocore_workshop    702 Oct  2 21:44 app.slurm
-rw-rw----+ 1 qiaoshan cri-biocore_workshop    231 Oct  2 21:44 metadata.txt
-rw-rw----+ 1 qiaoshan cri-biocore_workshop    566 Oct  2 21:44 nextflow.config
-rw-rw----+ 1 qiaoshan cri-biocore_workshop   1167 Oct  2 21:44 nextflow.slurm
-rw-rw----+ 1 qiaoshan cri-biocore_workshop    187 Oct  2 07:23 nextflow_318767.err
-rw-rw----+ 1 qiaoshan cri-biocore_workshop 966494 Oct  2 12:25 nextflow_318767.out
drwxrws---+ 7 qiaoshan cri-biocore_workshop   4096 Oct  2 12:24 results/
-rw-rw----+ 1 qiaoshan cri-biocore_workshop    814 Oct  2 21:44 samplesheet.csv
[qiaoshan@cri22in002 qiaoshan]$ tail nextflow_318767.out
[d1/d8ae38] process > NFCORE_RNASEQ:RNASEQ:BAM_RS... [100%] 6 of 6 ✔
[c2/a73652] process > NFCORE_RNASEQ:RNASEQ:MULTIQ... [100%] 1 of 1 ✔
Waiting for file transfers to complete (1 files)
-[nf-core/rnaseq] Pipeline completed successfully -
Completed at: 02-Oct-2025 12:25:23
Duration    : 5h 1m 37s
CPU hours   : 45.9
Succeeded   : 216
```

Nextflow outputs

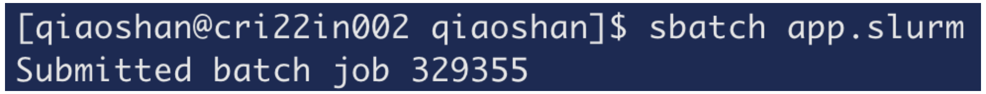
Check the output messages in the .out file

Once you confirm the pipeline has been completed successfully, use the following command to clean up the intermediate files

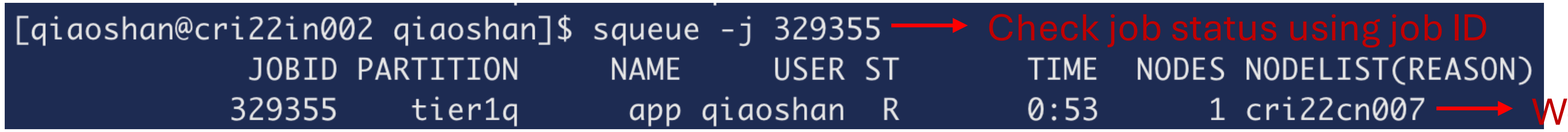


```
rm -r work/ tmp/ singularity/ .nextflow*
```

Then submit the DE analysis script



```
[qiaoshan@cri22in002 qiaoshan]$ sbatch app.slurm
Submitted batch job 329355
```



```
[qiaoshan@cri22in002 qiaoshan]$ squeue -j 329355
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            329355    tier1q      app qiaoshan  R       0:53      1 cri22cn007
```

Check job status using job ID

We will need this running node name

In-House Downstream Analysis

Step 3: Fill out the form

1. Open a local terminal and run:

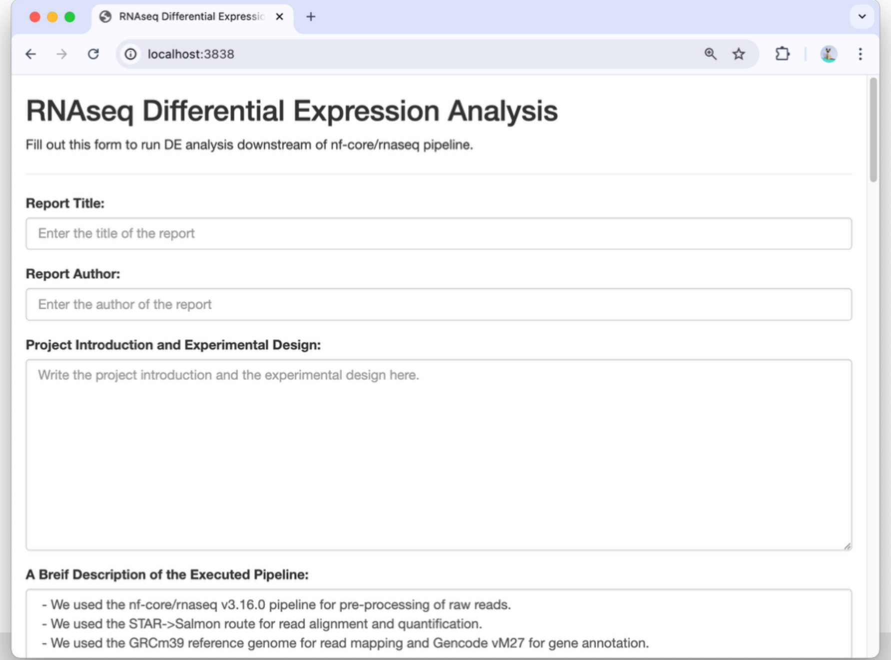
```
ssh -N -f -L 3838:shinyApp_running_node:3838 username@randi.cri.uchicago.edu
```

   1. -L XXXX:cri22cnYYY:3838 → **Forward local port XXXX to remote port 3838**.
   2. -N → **No interactive shell**.
   3. -f → **Run in the background** (not supported in Windows OpenSSH).

2. Open http://localhost:3838/ in a browser

3. Follow the hints to fill out the form and submit

Note: the port might be conflicted when multiple users are running on the same node. If there shows an error, try to change the port number and restart the job.



RNAseq Differential Expression Analysis

Fill out this form to run DE analysis downstream of nf-core/rnaseq pipeline.

Report Title:

Enter the title of the report

Report Author:

Enter the author of the report

Project Introduction and Experimental Design:

Write the project introduction and the experimental design here.

A Breif Description of the Executed Pipeline:

- We used the nf-core/rnaseq v3.16.0 pipeline for pre-processing of raw reads.
- We used the STAR->Salmon route for read alignment and quantification.
- We used the GRCm39 reference genome for read mapping and Gencode vM27 for gene annotation.

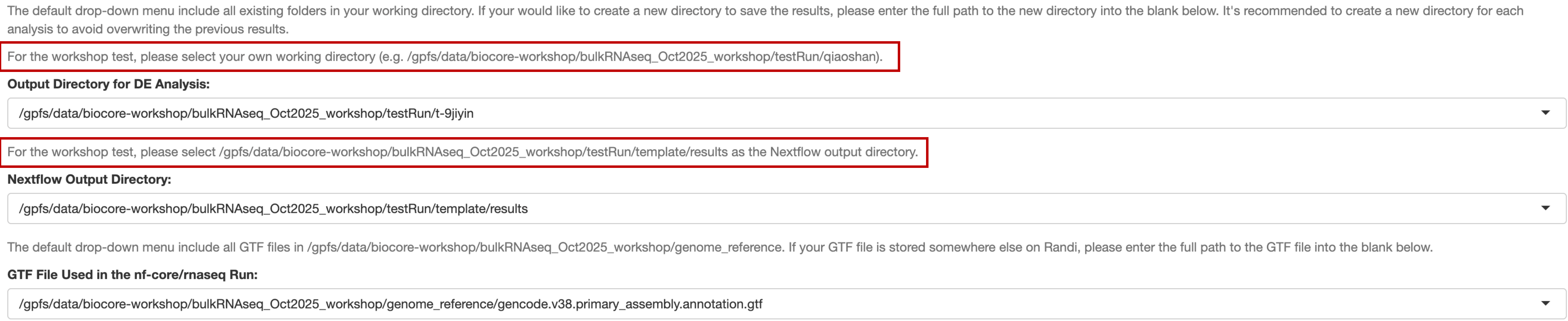
If the above doesn't work on your local computer, try:

```
ssh -N -J username@randi.cri.uchicago.edu -L 3838:0.0.0.0:3838 username@shinyApp_running_node
```

In-House Downstream Analysis

Step 3: Fill out the form

It’s important to provide the correct directory paths to the app; otherwise, it won’t find the target files.



The default drop-down menu include all existing folders in your working directory. If your would like to create a new directory to save the results, please enter the full path to the new directory into the blank below. It's recommended to create a new directory for each analysis to avoid overwriting the previous results.

For the workshop test, please select your own working directory (e.g. /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/qiaoshan).

Output Directory for DE Analysis:

/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/t-9jiyin

For the workshop test, please select /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/template/results as the Nextflow output directory.

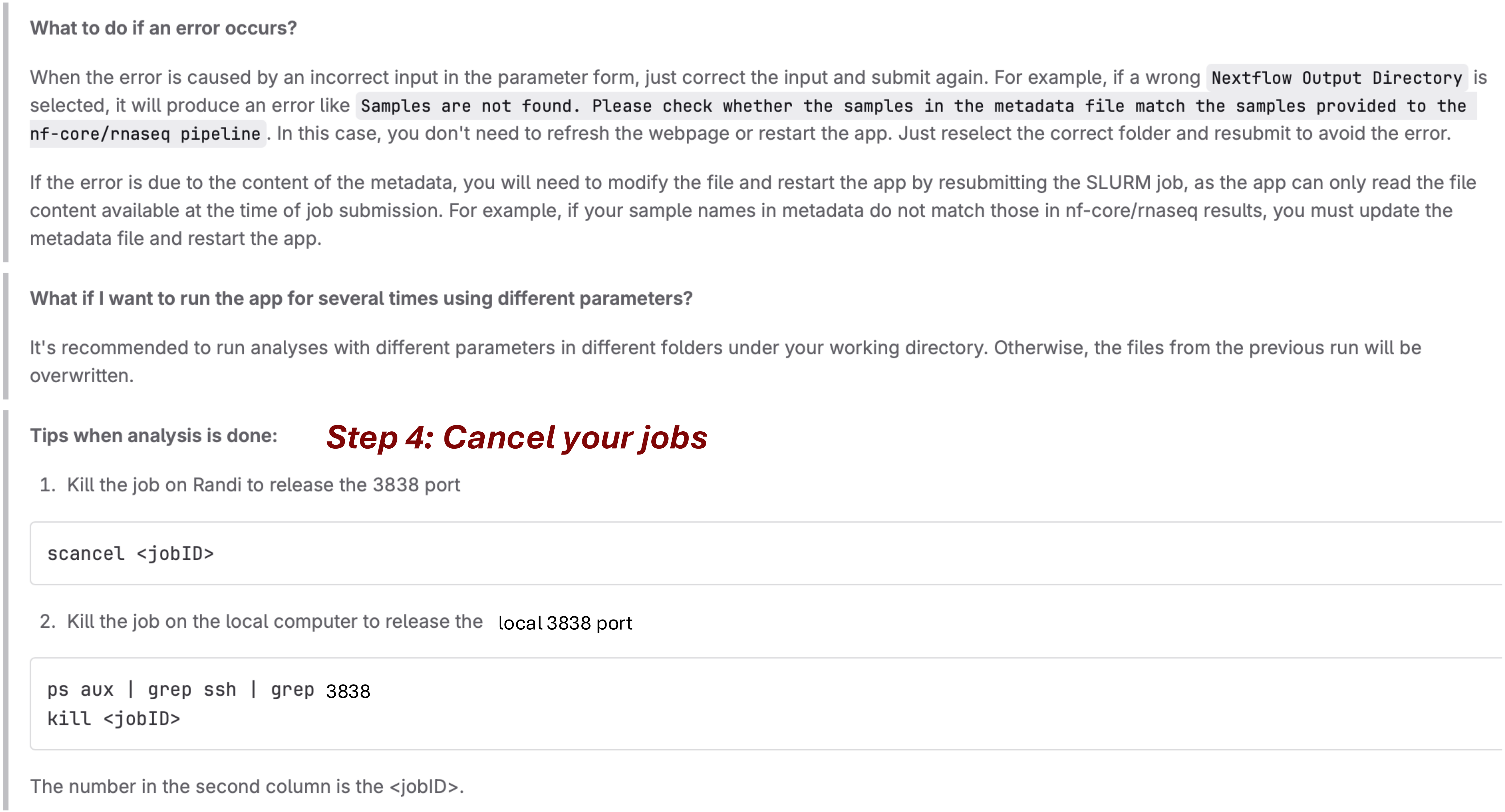
Nextflow Output Directory:

/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/testRun/template/results

The default drop-down menu include all GTF files in /gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/genome_reference. If your GTF file is stored somewhere else on Randi, please enter the full path to the GTF file into the blank below.

GTF File Used in the nf-core/rnaseq Run:

/gpfs/data/biocore-workshop/bulkRNAseq_Oct2025_workshop/genome_reference/gencode.v38.primary_assembly.annotation.gtf



What to do if an error occurs?

When the error is caused by an incorrect input in the parameter form, just correct the input and submit again. For example, if a wrong `Nextflow Output Directory` is selected, it will produce an error like `Samples are not found. Please check whether the samples in the metadata file match the samples provided to the nf-core/rnaseq pipeline`. In this case, you don't need to refresh the webpage or restart the app. Just reselect the correct folder and resubmit to avoid the error.

If the error is due to the content of the metadata, you will need to modify the file and restart the app by resubmitting the SLURM job, as the app can only read the file content available at the time of job submission. For example, if your sample names in metadata do not match those in nf-core/rnaseq results, you must update the metadata file and restart the app.

What if I want to run the app for several times using different parameters?

It's recommended to run analyses with different parameters in different folders under your working directory. Otherwise, the files from the previous run will be overwritten.

Tips when analysis is done:

Step 4: Cancel your jobs

1. Kill the job on Randi to release the 3838 port

```
scancel <jobID>
```

2. Kill the job on the local computer to release the local 3838 port

```
ps aux | grep ssh | grep 3838
kill <jobID>
```

The number in the second column is the <jobID>.

Interpretation of DE Analysis Results

PCA Plot

- A dimension reduction approach
- To show similar samples clustering together
- To reveal batch effects
- To infer which gene is the most valuable for clustering the data (eigenvector)

Volcano Plot

- To highlight significant DEGs

Heatmap

- To visualize and validate expression changes across samples

Over Representation Analysis (ORA)

- Which gene set is enriched with DEGs (whether a gene set contains more DEGs than expected by chance)

Gene Set Enrichment Analysis (GSEA)

- Identifies pathways that are globally up- or down-regulated (by using continuous gene rankings instead of requiring an arbitrary DEG cutoff)

Reminders

- Please remember to delete your folder after practice to save space for other people.

- We will hold the bulkRNAseq_Oct2025_workshop folder for you until Nov. 7th. All data will be removed to make room for our next workshop.

Upcoming Workshop: Functional Analysis

Yan Li, PhD

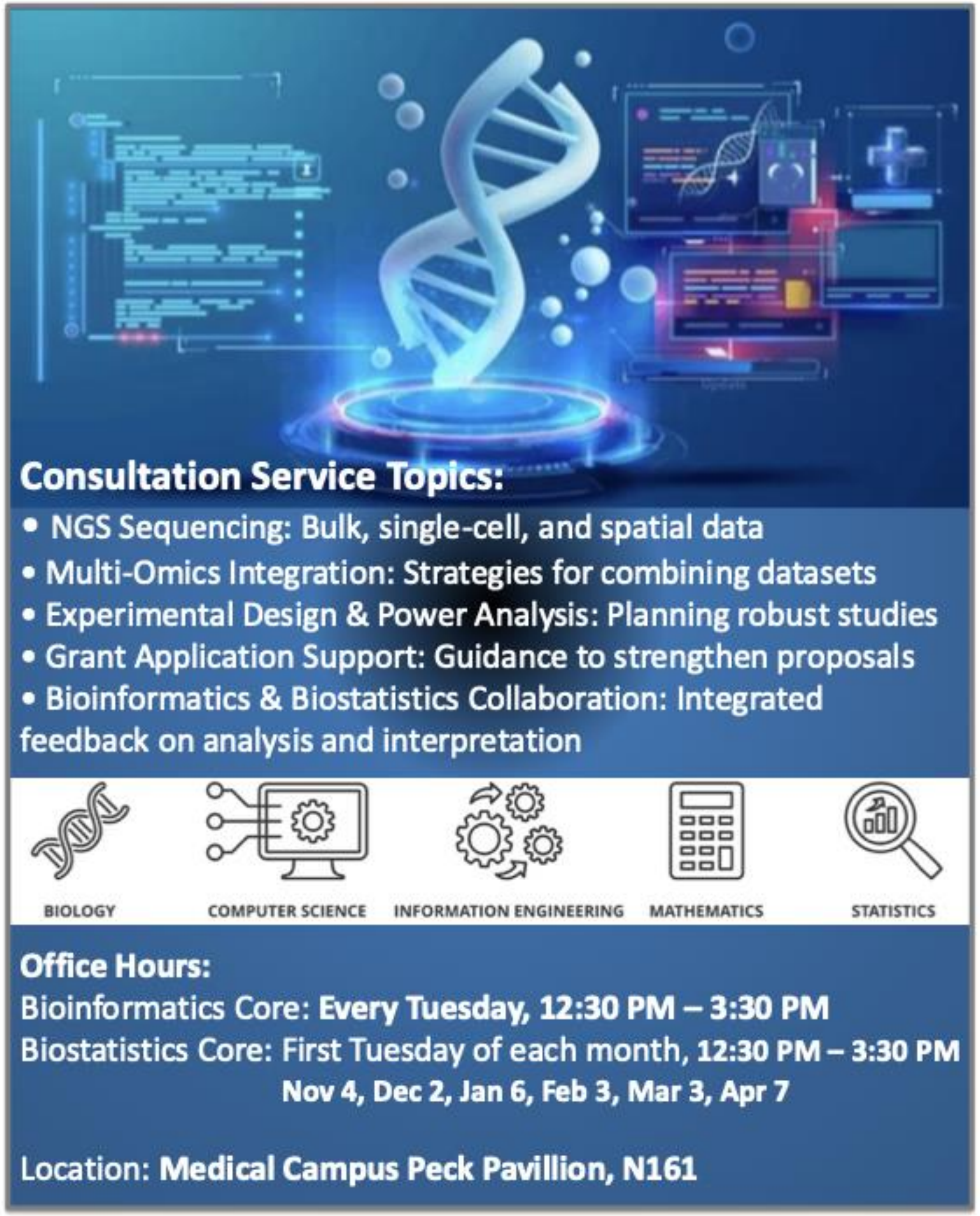
Associate
Director

Date: November 19th 1-3pm

Thanks for attending! 💐

Q & A

Bioinformatics & Biostatistics Core Joint Office Hours



BIOLOGY
COMPUTER SCIENCE
INFORMATION ENGINEERING
MATHEMATICS
STATISTICS

Consultation Service Topics:

- NGS Sequencing: Bulk, single-cell, and spatial data
- Multi-Omics Integration: Strategies for combining datasets
- Experimental Design & Power Analysis: Planning robust studies
- Grant Application Support: Guidance to strengthen proposals
- Bioinformatics & Biostatistics Collaboration: Integrated feedback on analysis and interpretation

Office Hours:

Bioinformatics Core: **Every Tuesday, 12:30 PM – 3:30 PM**

Biostatistics Core: First Tuesday of each month, **12:30 PM – 3:30 PM**

Nov 4, Dec 2, Jan 6, Feb 3, Mar 3, Apr 7

Location: **Medical Campus Peck Pavillion, N161**

Please give us some feedback!

https://mycri.cri.uchicago.edu/educations/trainings/75/survey/