Module 4:

Introduction to High-Performance Computing with an RNA-sequencing Pipeline

Michelle Franc Ragsac (She/Hers/Siya)
4th Year BISB Student, Emma Farley Lab
mragsac@eng.ucsd.edu

September 15, 2021

Learning Goals of Module 4

1. Background on High Performance Computing (HPC)
    a. Motivations for using a compute cluster for biological data analysis
    b. Introduction to the Triton Shared Compute Cluster (TSCC) at UCSD

2. **Practical Case Study:** RNA-Sequencing
    a. Rationale for using RNA-sequencing experiments to study gene expression
    b. Roadmap of a standard RNA-sequencing analysis pipeline

3. Running through a RNA-Seq analysis pipeline on TSCC
    a. Understanding **FASTQ** File Data Quality with **`fastqc`**
    b. Mapping Illumina short reads to the genome with **STAR**
    c. Sorting and Indexing aligned sequencing reads with **`samtools`**
    d. Generating a gene expression counts table with **`featureCounts`**
    e. Compiling Next-Generation Sequencing (NGS) file logs with **`multiqc`**
    f. (If there's time) Differential analysis of gene expression with **DESeq2**

If you haven’t worked with me to set up your TSCC account yet, I have a few hidden slides for installation instructions for those that still need to **generate SSH keys** on their computer!

Local Computer Setup

Generating SSH Keys

First off, what is SSH and why do I need SSH keys?

SSH (Secure SHell) protocol is a method to securely remote login from one computer to another while protecting the security of communication and integrity with encryption.

Almost all UNIX-based systems include the **ssh** command, which is used to start the **SSH** client program to initiate secure connections to an **SSH** server on a remote machine.

SSH keys are cryptographic keys using a public key cryptosystem--functionally, they act as passwords to control who can access certain systems!

We will be using **SSH keys** to securely log into TSCC training accounts that have been assigned to you! Without **SSH keys**, you will *not* be able to access your account.

You can learn more about SSH from SSH Academy (https://www.ssh.com/academy).

How exactly do SSH keys work?

your computer

① Generate the SSH keys on your computer

generates two files in a hidden `~/.ssh/` folder:

public key `id_rsa.pub`

a lock you can place in places you want to access from your local computer

② Share your public key with *remote* (e.g., TSCC) accounts you want to access without a password

SSH server

private key `id_rsa`

*a **master key** you can use to access places where your public key (lock) is located*

******do not share the private key, keep this file on your computer***

How do I generate my own SSH keys on my computer?

SSH keys are relatively easy to generate and *you only have to generate them once for a computer*!

After the **SSH keys** are generated, the public key can be used with any account you want to access from that computer.

You can generate **SSH keys** with the command:

ssh-keygen

Example output of ssh-keygen command:

```
klar (11:39) ~>ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ylo/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/ylo/.ssh/id_rsa.
Your public key has been saved in /home/ylo/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:Up6KjbnEV4Hgfo75YM393QdQsK3Z0aTNBz0DoirrW+c ylo@klar
The key's randomart image is:
+---[RSA 2048]----+
|     .       ..oo..|
|    . . .  . .o.X.|
|     . . o.  ..+ B|
|    .   o.o  .+ ..|
|     ..o.S   o..   |
|    . %o=      .   |
|     @.B...      . |
|    o.=. o. . .  .|
|     .oo  E. . .. |
+----[SHA256]-----+
klar (11:40) ~>
```

https://www.ssh.com/ssh/keygen/

After generating my keys, how do I set up my TSCC account?

The **Triton Shared Compute Cluster (TSCC)** is a high-performance computing (HPC) cluster based at UCSD that multiple research labs analyze their data on! We were able to get training accounts for bootcamp so you can get some experience on the cluster :-)

To access your assigned training account, please email me (mragsac@eng.ucsd.edu) your **SSH public key file** that you generated with the `ssh-keygen` command!

If you're having trouble finding the file, the **public** (`id_rsa.pub`) and **private** (`id_rsa`) keys can usually be found in the hidden folder `~/.ssh/` !

I hope the hidden resources helped!
Let’s get back to the main presentation :-)

Background on High Performance Computing (HPC)

What's a compute cluster and why do I want to use one?

- Biological datasets nowadays can get quite large, so more power is needed to perform analyses due to high memory and storage requirements!

- **Servers** contain much more memory, storage, and compute capacity than local devices (e.g., laptop computers)
- **Clusters** are multiple servers that are physically linked together

https://www.seattletimes.com/explore/at-home/drowning-in-paperwork-how-to-get-it-organized/

What is the Triton Shared Compute Cluster (TSCC)?

- **TSCC** is a research cluster housed at the **San Diego Supercomputer Center (SDSC)**
- Utilizes a "*condo* cluster" system built by contributing researchers
  1. Labs can purchase **compute nodes** (think computer with lots of RAM and minimal storage)
  2. These **nodes** then become a part of the cluster's compute power for all members to use
- Also hosts "*hotel* services" so researchers can use **TSCC** via a pay-as-you-go model

https://www.elementaconsulting.com/projects/uc-san-diego-supercomputer-center/

You can learn more about TSCC from the User Guide (https://www.sdsc.edu/support/user_guides/tscc.html)

Using the Triton Shared Compute Cluster (TSCC) for Biological Research

How can I access TSCC? What do I do to log into the cluster?

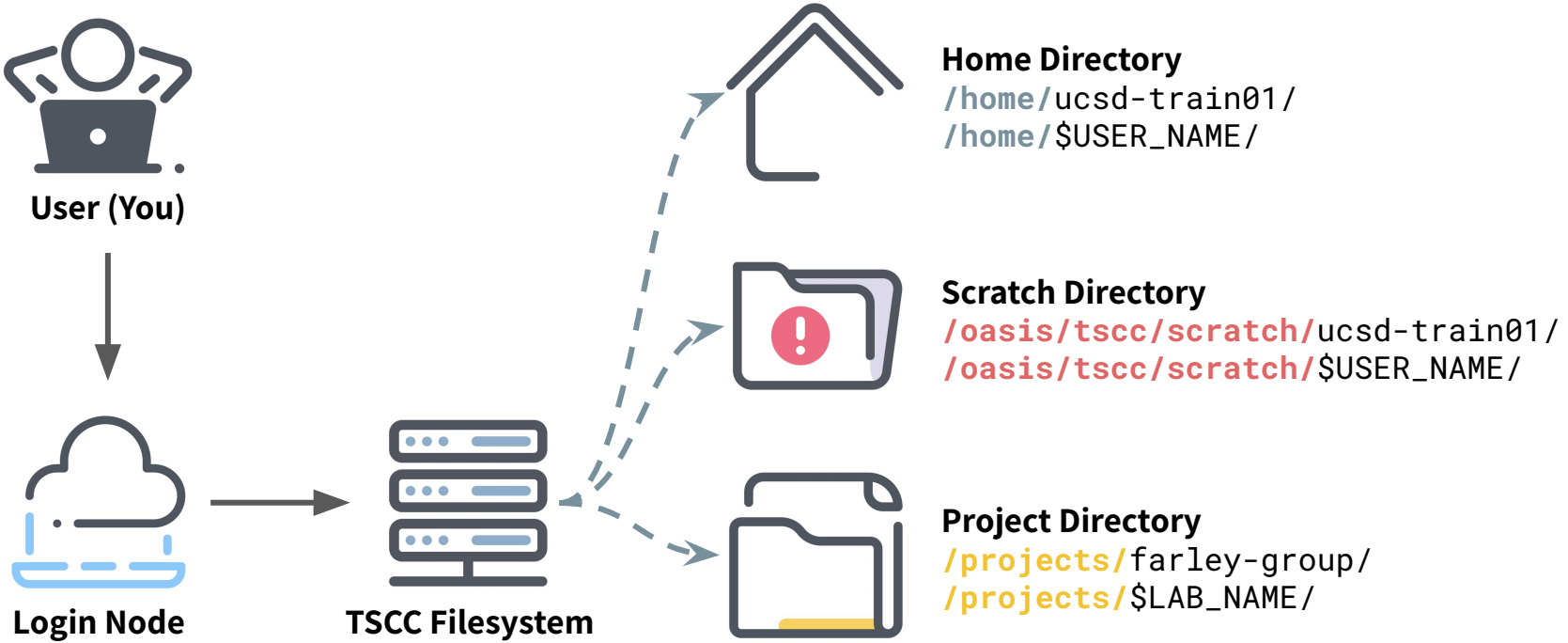
User (You)

Login Node

There are 4 **login nodes** on TSCC:

```
ssh $USER_NAME@tscc-login1.sdsc.edu
ssh $USER_NAME@tscc-login2.sdsc.edu
ssh $USER_NAME@tscc-login11.sdsc.edu
ssh $USER_NAME@tscc-login12.sdsc.edu
```

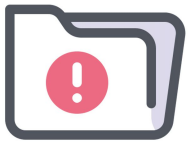
Where and how can I access files on TSCC?



Is there a difference between HOME, SCRATCH, & PROJECT?

Home Directory
`/home/ucsd-train01/`
`/home/$USER_NAME/`

- Each user has their own permanent home folder
- Very minimal space (**100 GB**)



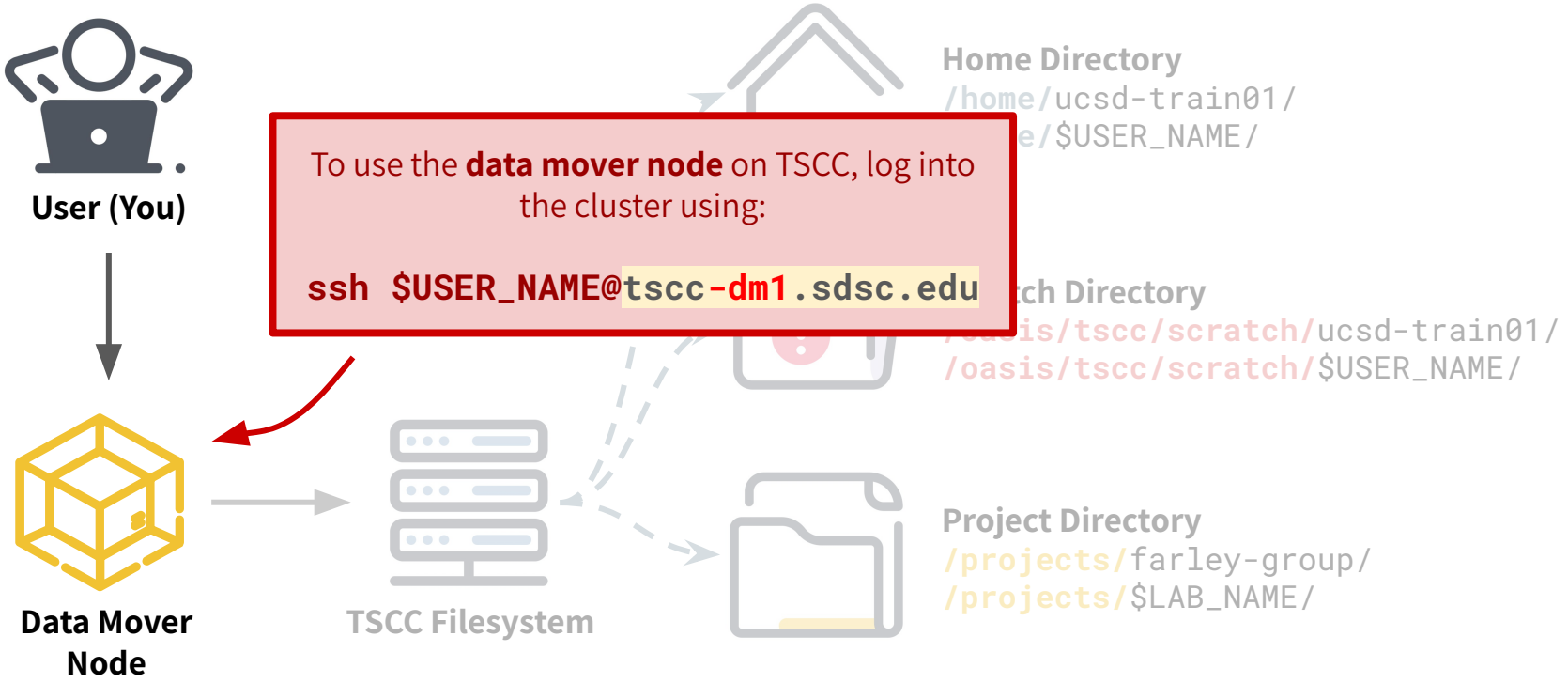
Scratch Directory
`/oasis/tscc/scratch/ucsd-train01/`
`/oasis/tscc/scratch/$USER_NAME/`

- Each user has their own temporary folder
- Lots of space (**25 TB**), *but* untouched files get purged **every 3 months**

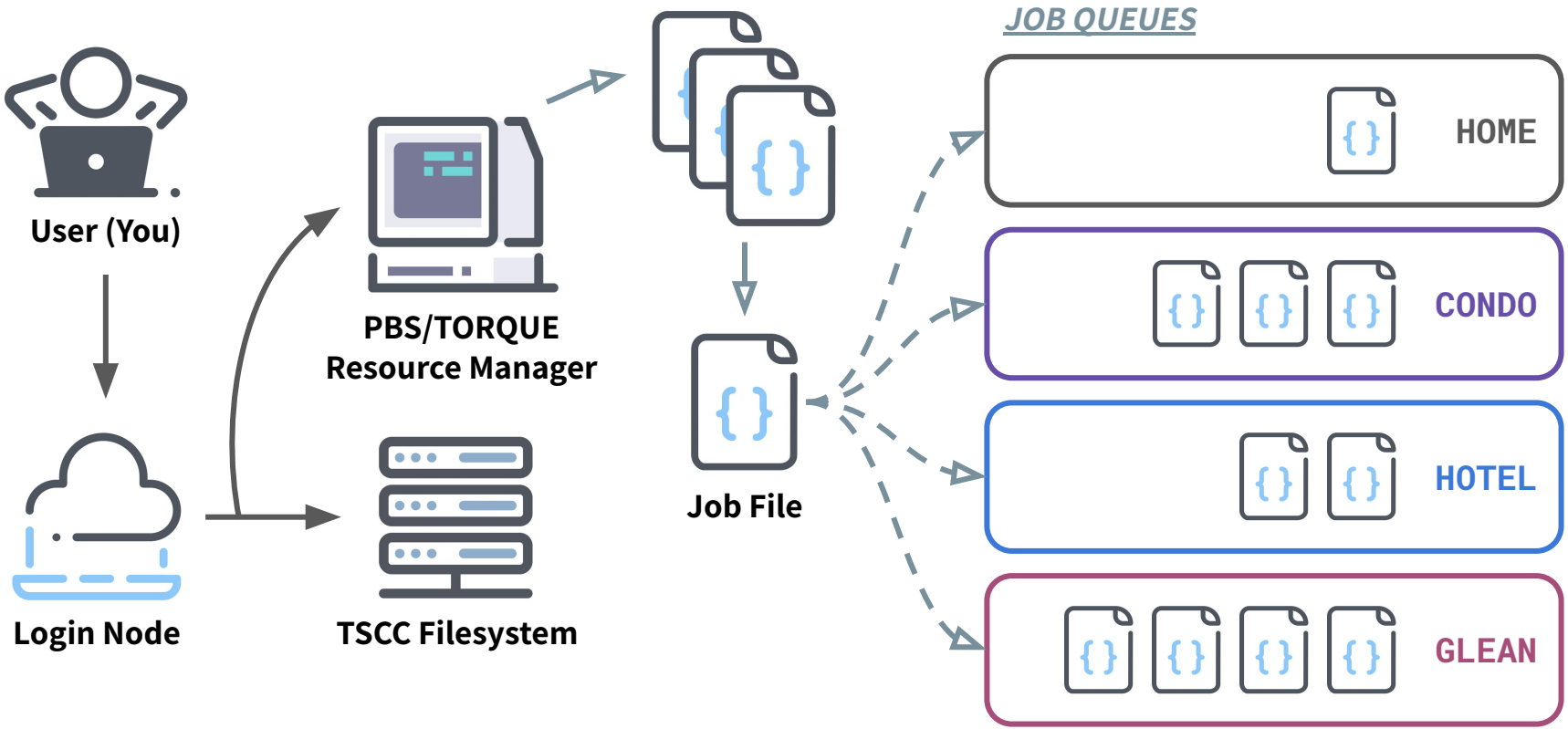
Project Directory
`/projects/farley-group/`
`/projects/$LAB_NAME/`

- Long-term archival storage space purchased by labs (e.g., **`ps-farleylab`**)
- Shared amongst members of a lab

How can I move my files off TSCC?



How can I actually run things on TSCC?



What is the difference between HOME, CONDO, HOTEL, & GLEAN?

HOME

- Purchased nodes reserved for members of a particular user group (e.g., members of the Farley Lab)
- **Unlimited time limit**

CONDO

- Allows contributors to TSCC to run their compute jobs on nodes greater than those they have purchased
- **8-hour time limit**

HOTEL

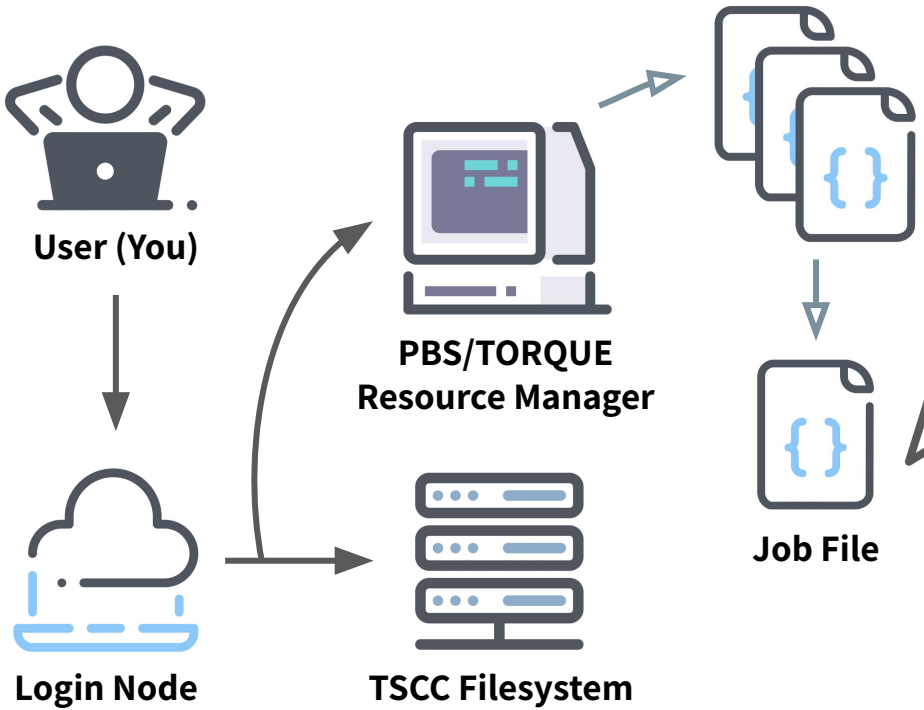
- Supports all non-contributors to TSCC
- **168-hour time limit**

This is the only node we're allowed to use our training accounts for !!

GLEAN

- Allows users to run jobs free-of-charge on available idle nodes within the **CONDO** queue
- **1-hour time limit**

What are cluster job files? What do they usually contain?

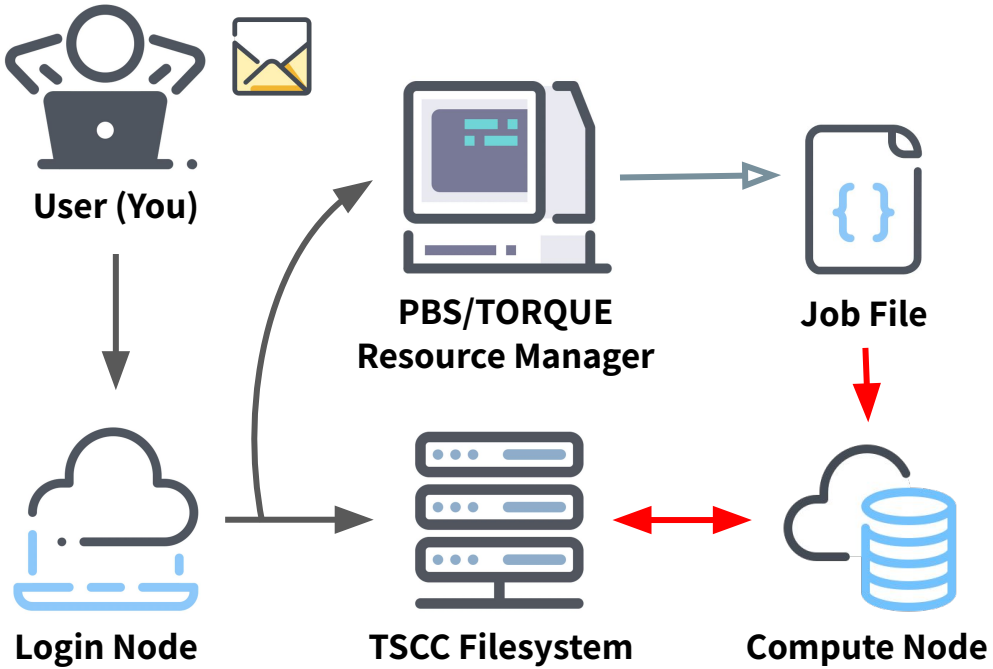


Contents of a Typical Job File:

- Name of the Job
- Queue for the Job
- Amount of Walltime Required
- Requested Amount of Computational Resources
- Commands for the Job

- Name of the Job's Log/Error File
- *Who is paying for the Job?*
- *Where should TSCC send status updates about the Job?*

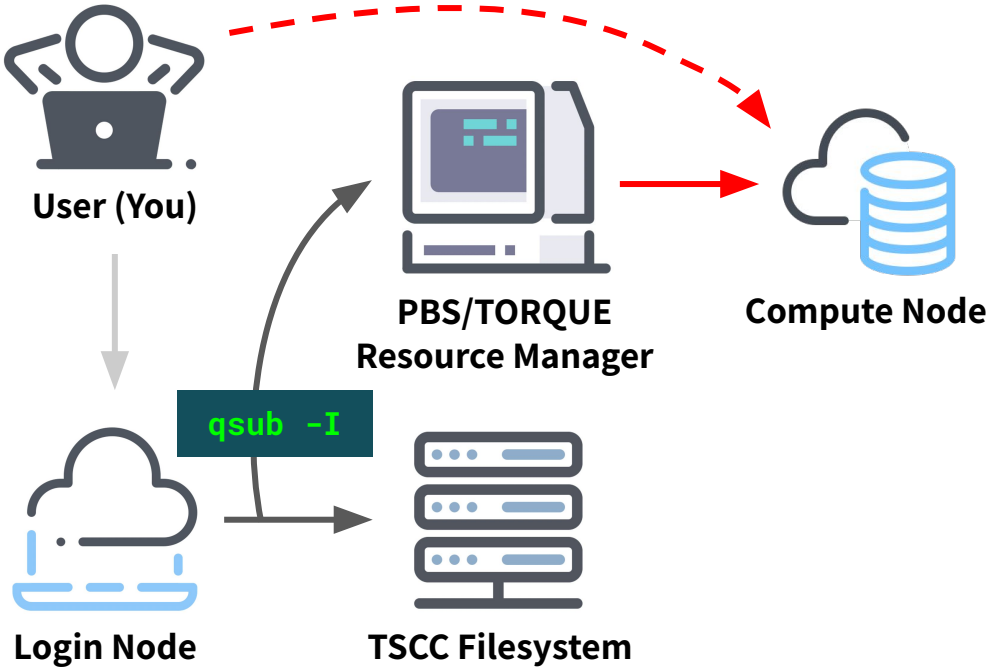
What is a **non-interactive** compute job?



Non-Interactive Compute Jobs

1. **Submit a Job File with required parameters** for the computational task
2. After the job gets off the queue, the job is allocated a node with the required resources and **runs the instructions in the Job File**
3. When the job is completed, the user is notified via Email (*if it was set*)

What is an **interactive** compute job?



Interactive Compute Jobs

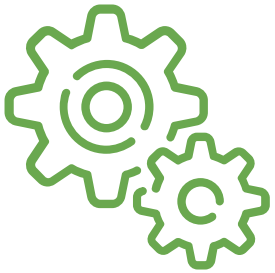
1. **Submit a command on the CLI with the parameters** for the job
2. After the job gets off the queue, the job is allocated a node with the required resources and **the user is moved from the login node to the assigned node to run commands**
3. When the job is completed, the user can exit the node

Let’s use the terminal
to **start an interactive job** to use
after the next section!

Open up your terminal application
to follow along with the class :-)

?

Does anybody have any questions after submitting an interactive job?



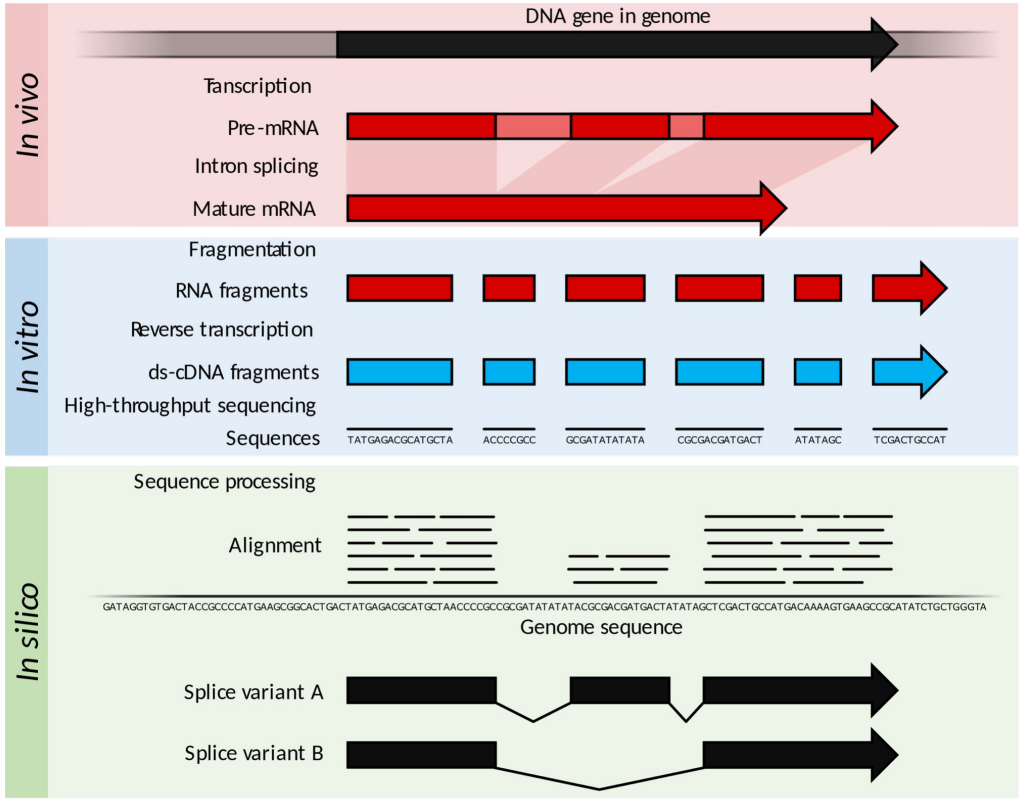
RNA-Sequencing*

Using Next-Generation Sequencing (NGS) to study gene expression in biological samples

** I won't be going deeply into RNA-seq because it should've been covered in Module 2!*

What is RNA-sequencing and what are the applications?

- The goal of RNA-seq is to **reveal the presence and quantity of RNA** within a biological sample at a given moment (i.e., transcriptome)
- One of the largest applications of RNA-seq is to look at **differential expression** between conditions
  - Disease vs. Non-Disease
  - Differential isoform expression

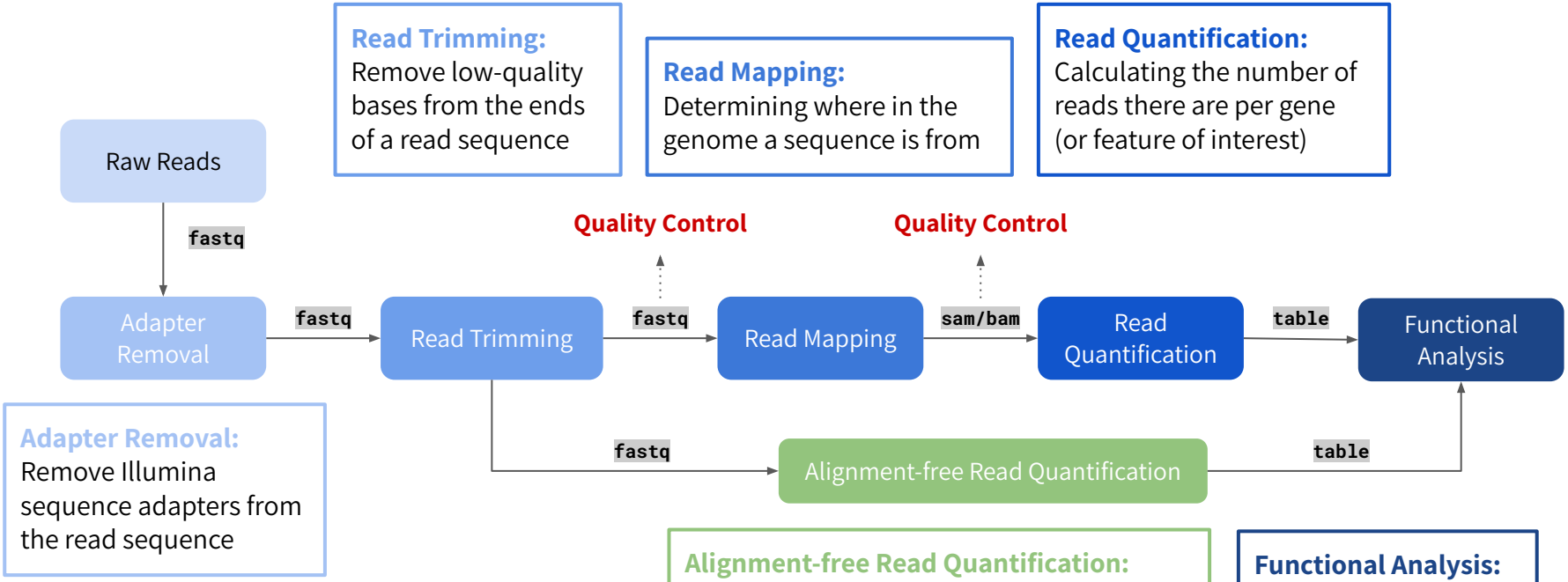


https://en.wikipedia.org/wiki/RNA-Seq

Running through a RNA-Sequencing Pipeline on TSCC

Let’s get some practical experience with the cluster using your assigned training accounts!

What is a standard RNA-sequencing analysis workflow?



Adapted from Clarence Mah's BISB Bootcamp 2020 slides!

A *brief sampling* of common RNA-seq analysis pipeline tools

Adapter Removal, Read Trimming

- Trimmomatic
- cutAdapt

Read Mapping

- STAR
- Bowtie & Bowtie2
- HISAT2
- TopHat

Alignment-free Read Quantification

- kallisto
- salmon

Quality Control

- FastQC
- SAMtools
- MultiQC

Read Quantification

- featureCounts
- Htseq-count
- RSEM

Functional Analysis

- DESeq2
- edgeR

Let’s use the terminal to run an RNA-sequencing analysis pipeline on TSCC!

Follow along with the class on your computer :-)

?

Does anybody have any questions after the demo?