Bulk RNA-seq analysis with the Cancer Genomics Cloud, powered by Seven Bridges

Phillip Webster, Genomics Scientist
March 8, 2022

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Class Overview

- Lecture 1 - Overview

  Hands-on demo: your first project on the CGC
- **Lecture 2 - Bulk RNA-seq workflow(s) and sequence alignment**

  Hands-on demo: your CGC workflow
- Lecture 3 - Differential expression and visualizations

  Hands-on demo: your CGC visualization
- Lecture 4 - Single-cell RNA-seq overview and workflow

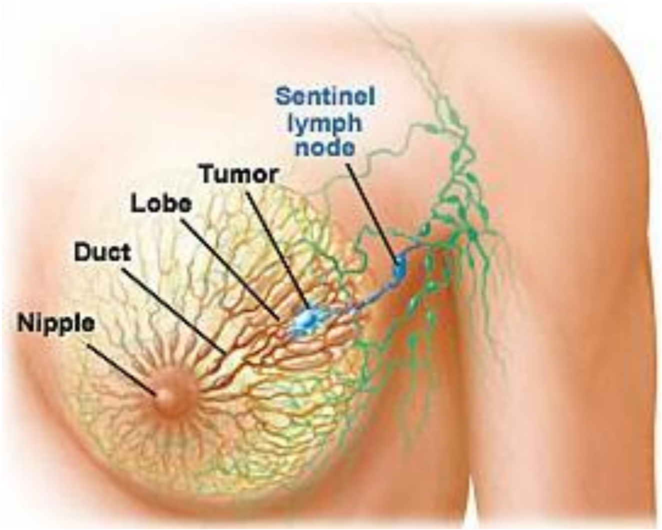
  Hands-on demo: Seurat

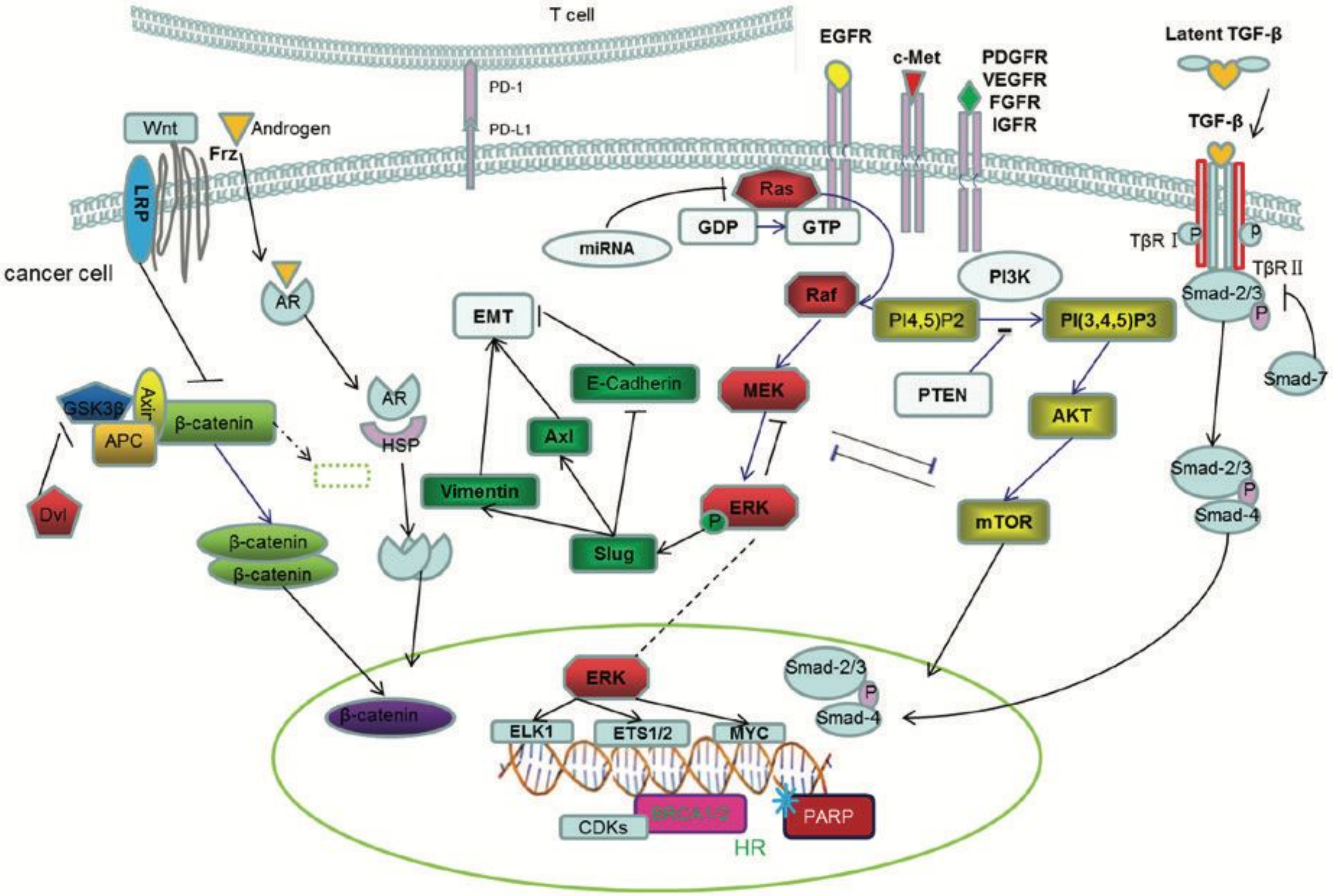
Lecture Overview

- Bulk RNA-seq workflow(s).
- Overview of CWL and workflow editor
- Description of tools and steps to create a workflow
- Hands-on demo: your CGC RNA-seq workflow

Saved By The CGC

Grad Phil circa 2017
Easily Excited
Hopeful





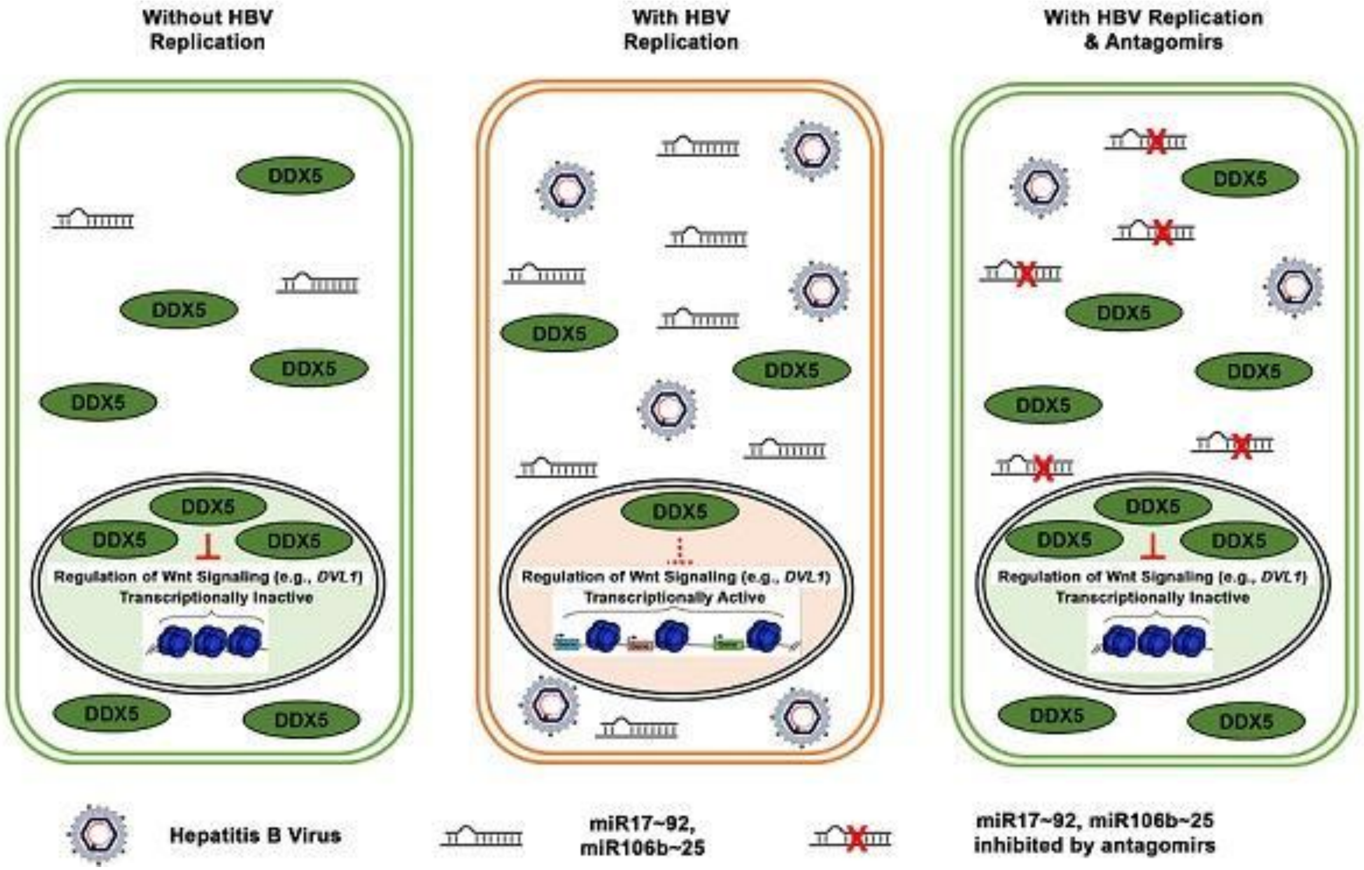
CANCER GENOMICS CLOUD
SEVEN BRIDGES

pipettegazette.uthscsa.edu

www.atlantictraining.com/blog/10-incredible-facts-breasts/

Nan wu., *et al. Oncology Letters.* 2018

Overview of data we will use for bulk RNA-seq analysis

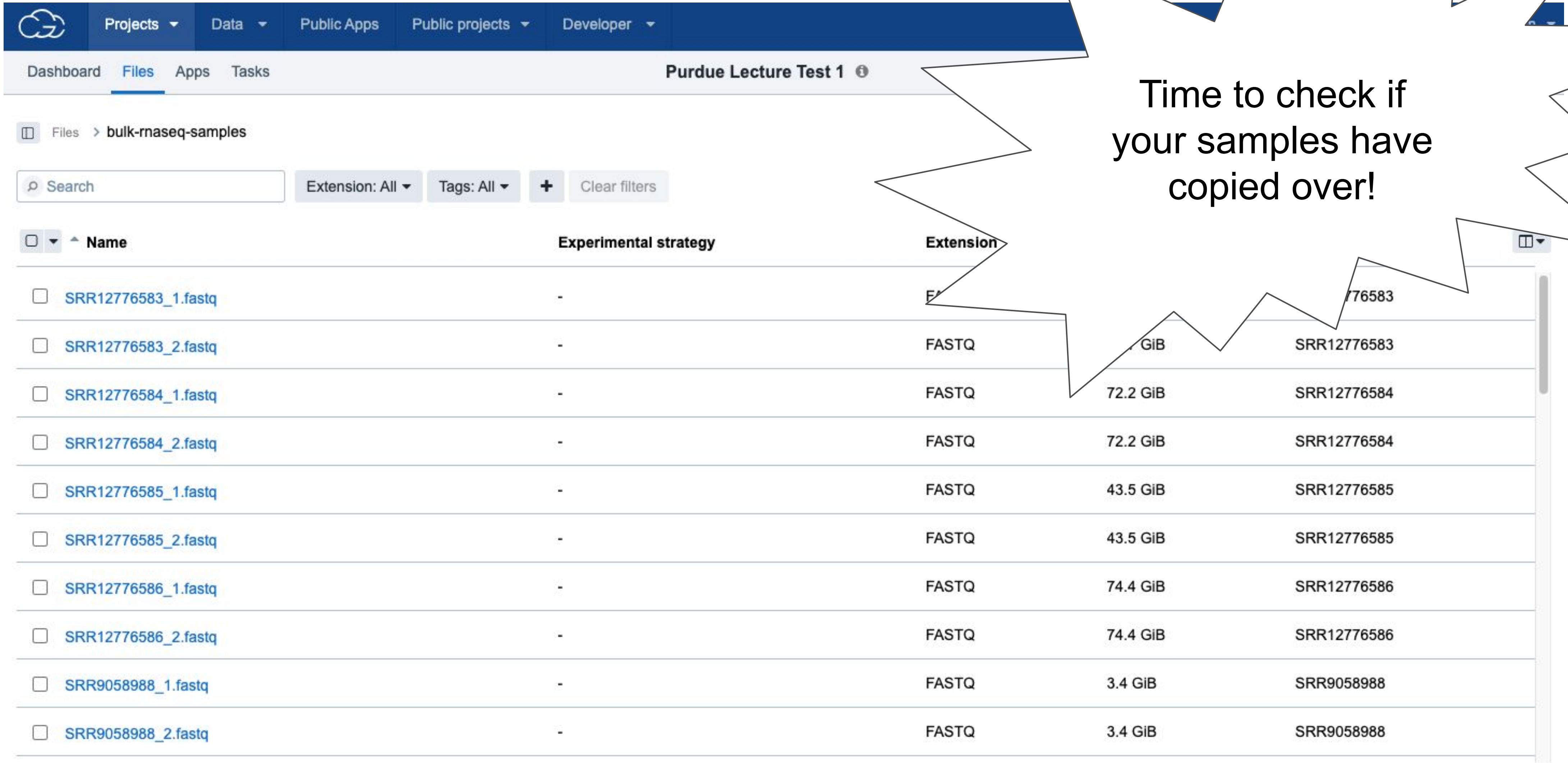


Mani SKK, et al. Restoration of RNA helicase DDX5 suppresses hepatitis B virus (HBV) biosynthesis and Wnt signaling in HBV-related hepatocellular carcinoma. *Theranostics* 2020; 10(24):10957-10972. doi:10.7150/thno.49629. https://www.thno.org/v10p10957.htm

Data availability

All sequencing data are available through the NCBI Gene Expression Omnibus (GEO) database (accession number **GSE131257**).

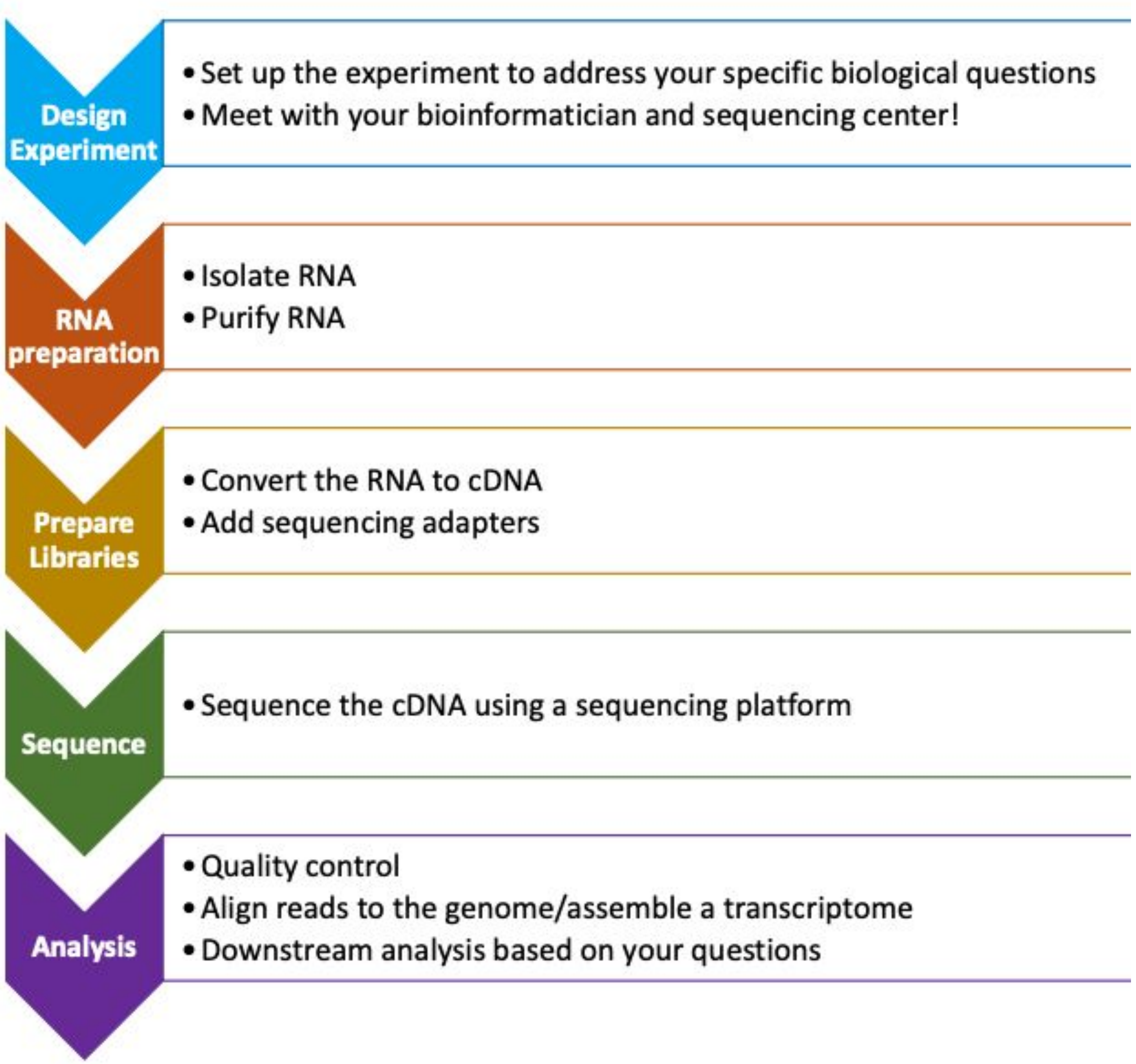
RNAseq data transferred from SRA to CGC



Name	Experimental strategy	Extension		
☐ SRR12776583_1.fastq	-	F		76583
☐ SRR12776583_2.fastq	-	FASTQ	GiB	SRR12776583
☐ SRR12776584_1.fastq	-	FASTQ	72.2 GiB	SRR12776584
☐ SRR12776584_2.fastq	-	FASTQ	72.2 GiB	SRR12776584
☐ SRR12776585_1.fastq	-	FASTQ	43.5 GiB	SRR12776585
☐ SRR12776585_2.fastq	-	FASTQ	43.5 GiB	SRR12776585
☐ SRR12776586_1.fastq	-	FASTQ	74.4 GiB	SRR12776586
☐ SRR12776586_2.fastq	-	FASTQ	74.4 GiB	SRR12776586
☐ SRR9058988_1.fastq	-	FASTQ	3.4 GiB	SRR9058988
☐ SRR9058988_2.fastq	-	FASTQ	3.4 GiB	SRR9058988

Time to check if your samples have copied over!

RNA-seq workflow overview



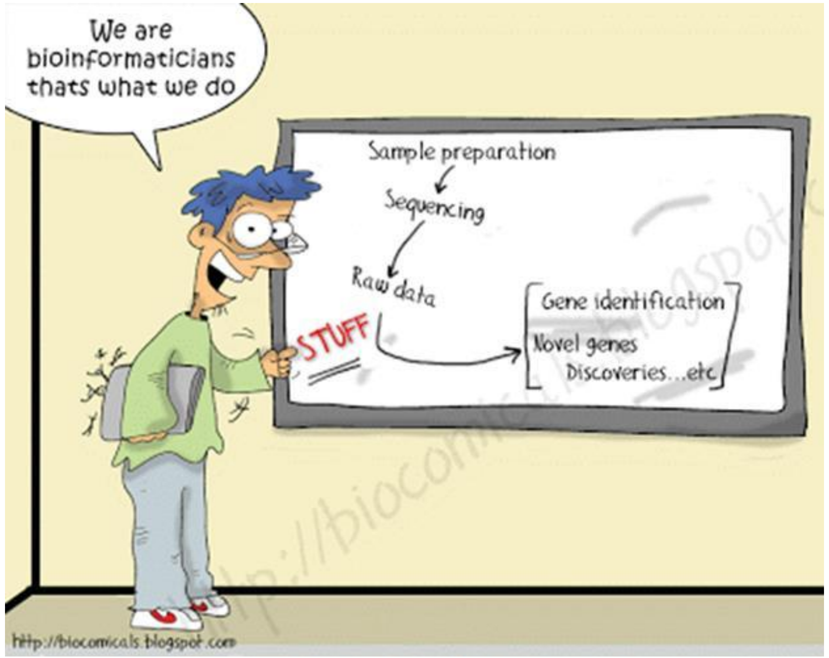
Design Experiment
- Set up the experiment to address your specific biological questions
- Meet with your bioinformatician and sequencing center!

RNA preparation
- Isolate RNA
- Purify RNA

Prepare Libraries
- Convert the RNA to cDNA
- Add sequencing adapters

Sequence
- Sequence the cDNA using a sequencing platform

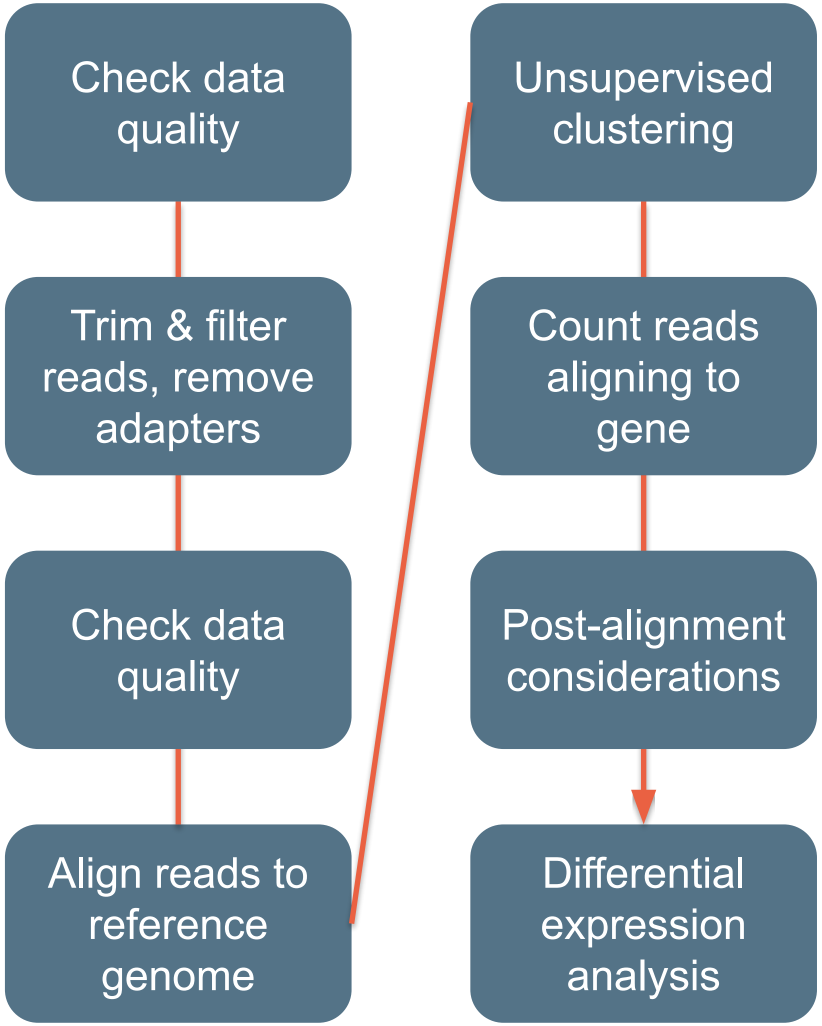
Analysis
- Quality control
- Align reads to the genome/assemble a transcriptome
- Downstream analysis based on your questions



We are here!

Slide adapted from Dr. Nadia Atallah, Purdue University

Standard Differential Expression Analysis



Check data quality

Trim & filter reads, remove adapters

Check data quality

Align reads to reference genome

Unsupervised clustering

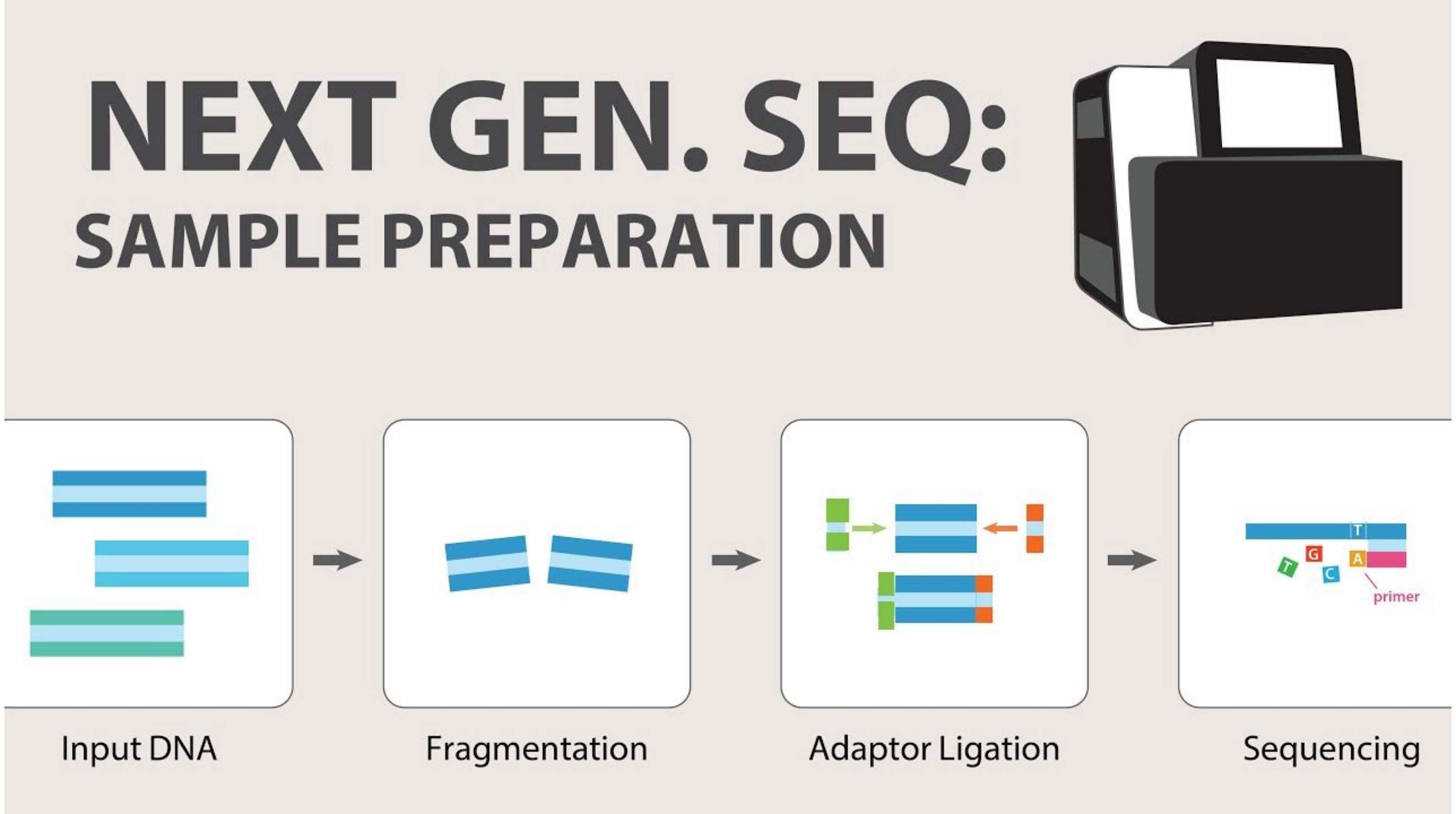
Count reads aligning to gene

Post-alignment considerations

Differential expression analysis

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Standard Differential Expression Analysis - Sample prep



Standard Differential Expression Analysis - Alignment vs Pseudo-alignment

Alignment-based
(STAR, HISAT2, etc.)

Pseudoalignment-based
(salmon, kallisto)

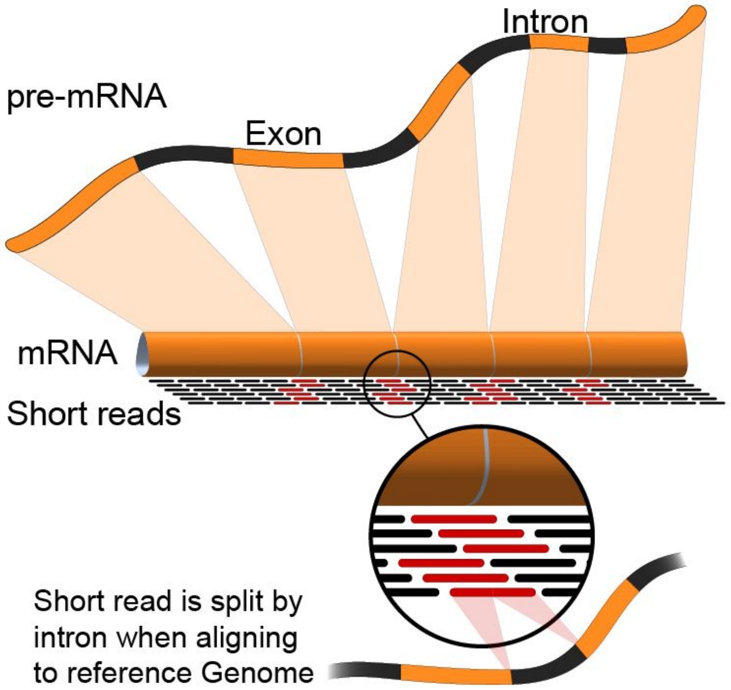


Image from https://hpc.ilri.cgiar.org/beca/bioinfo/rna-seq.html

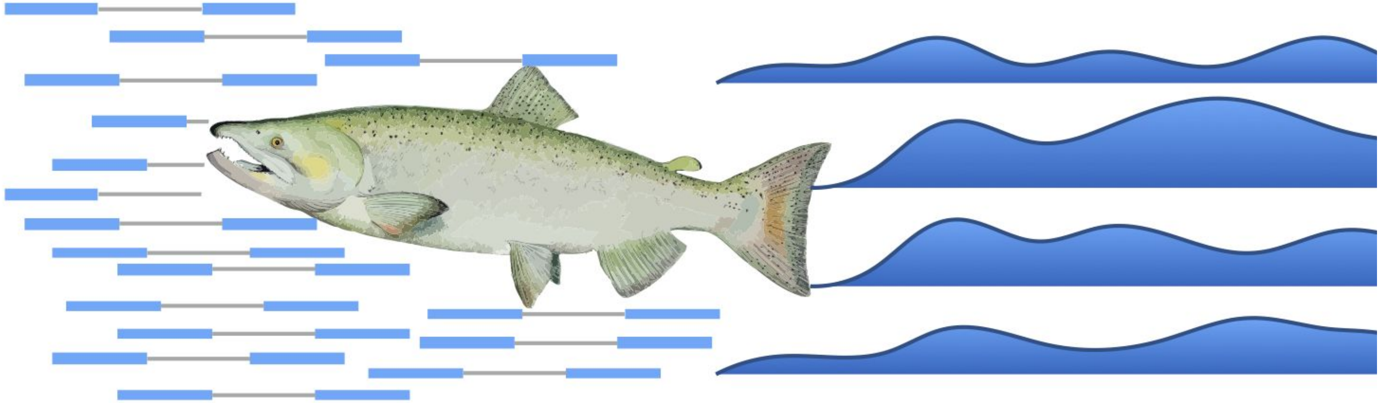
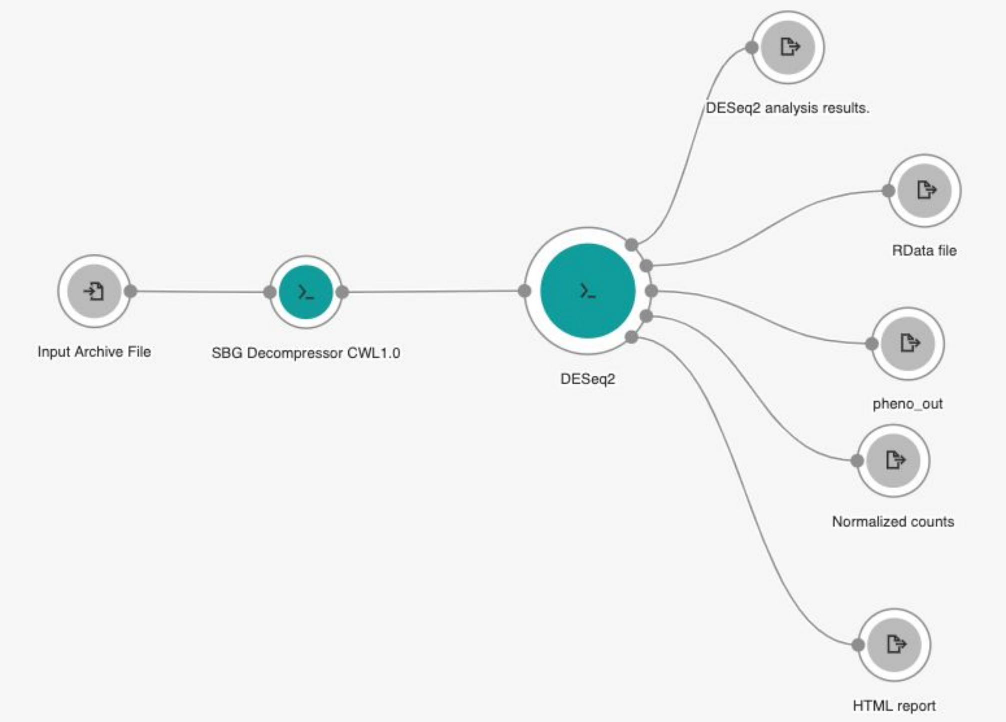
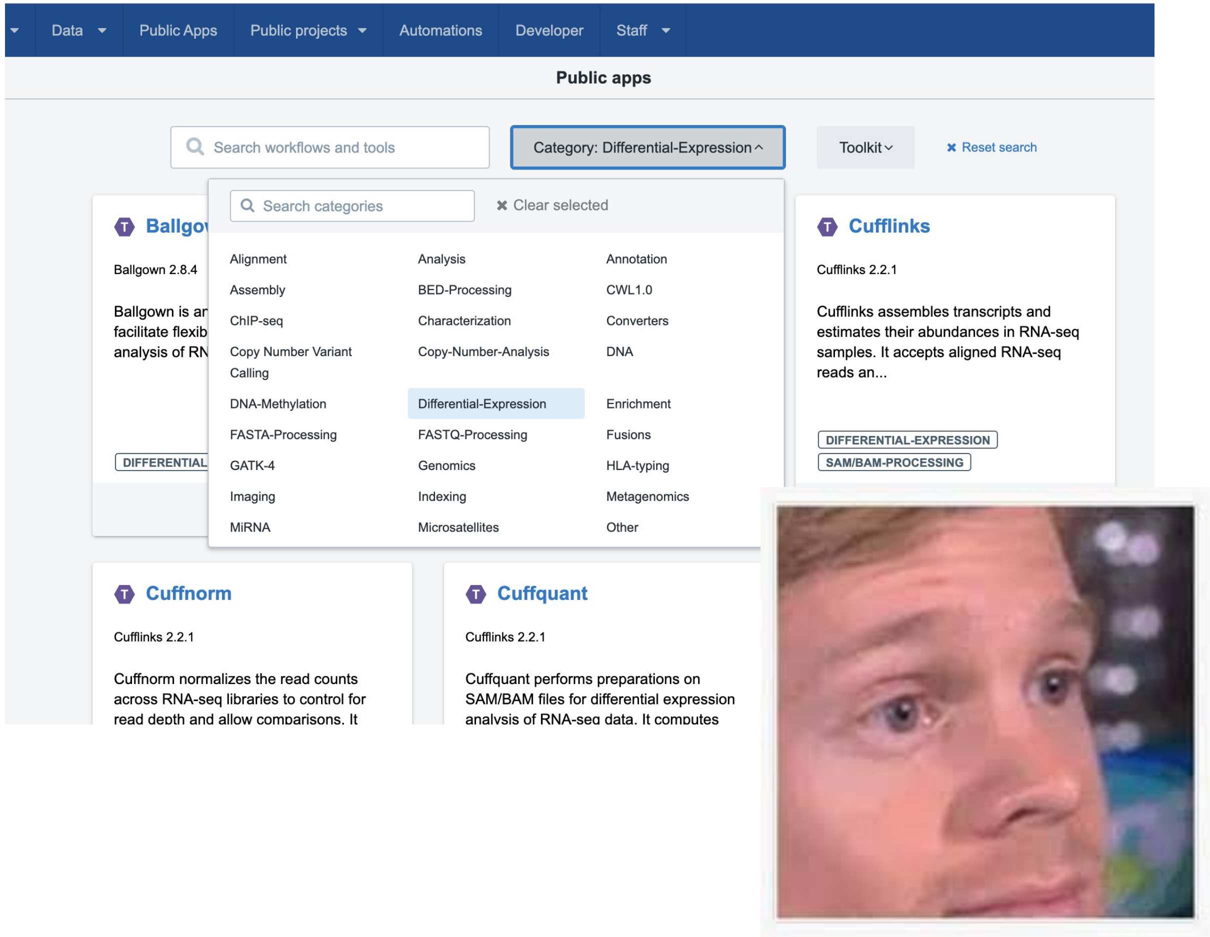


Image from https://combine-lab.github.io/salmon/

Standard Differential Expression Analysis - CWL & Tools





Ballgo

Ballgown 2.8.4

Ballgown is ar
facilitate flexib
analysis of RN

Cufflinks

Cufflinks 2.2.1

Cufflinks assembles transcripts and estimates their abundances in RNA-seq samples. It accepts aligned RNA-seq reads an...

Cuffnorm

Cufflinks 2.2.1

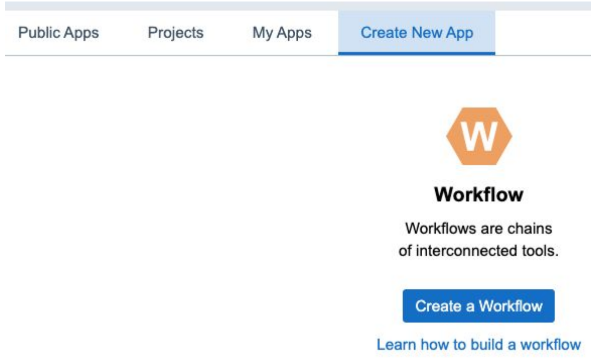
Cuffnorm normalizes the read counts across RNA-seq libraries to control for read depth and allow comparisons. It

Cuffquant

Cufflinks 2.2.1

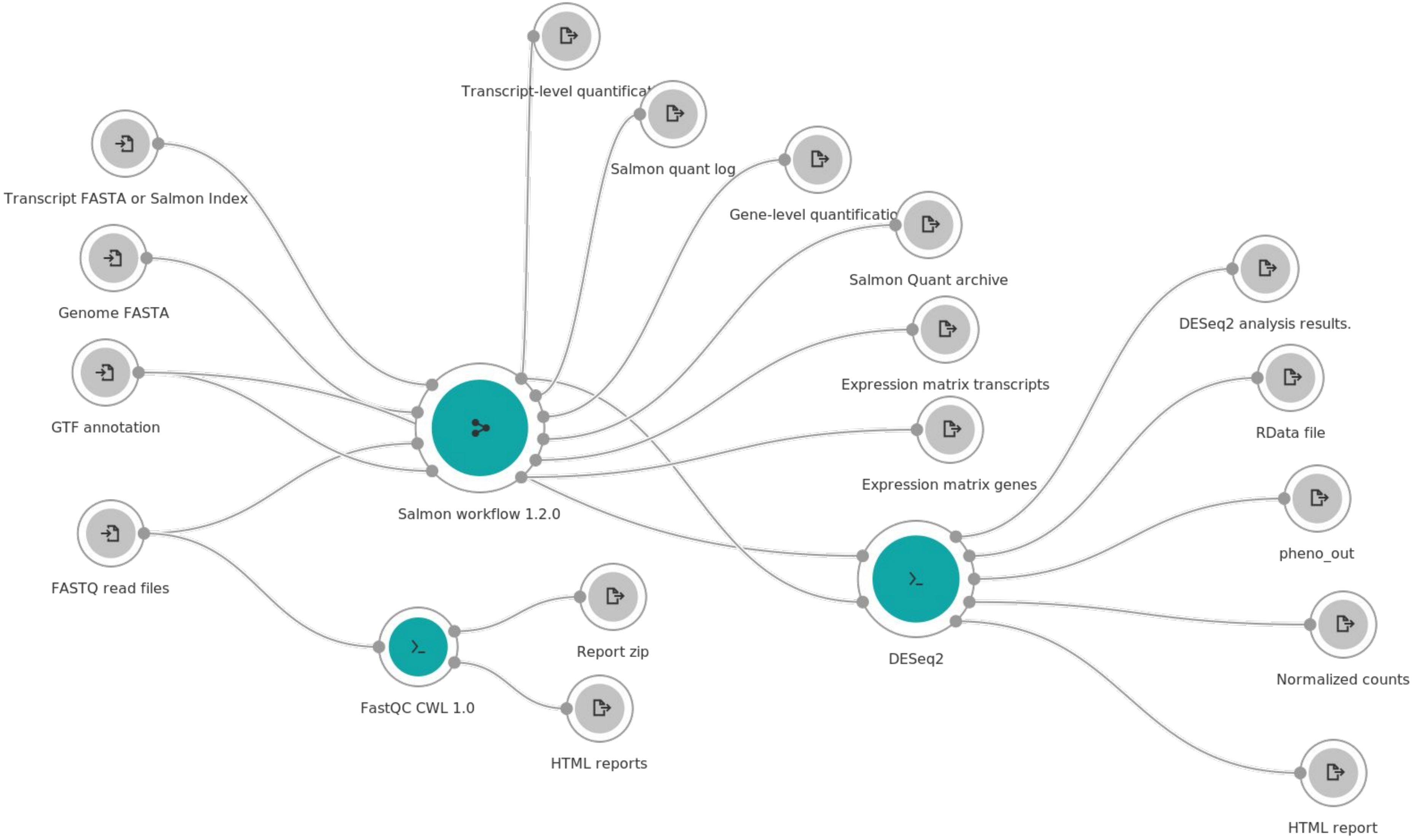
Cuffquant performs preparations on SAM/BAM files for differential expression analysis of RNA-seq data. It computes

What tools will we use? How I create a workflow?



Workflow

Workflows are chains of interconnected tools.



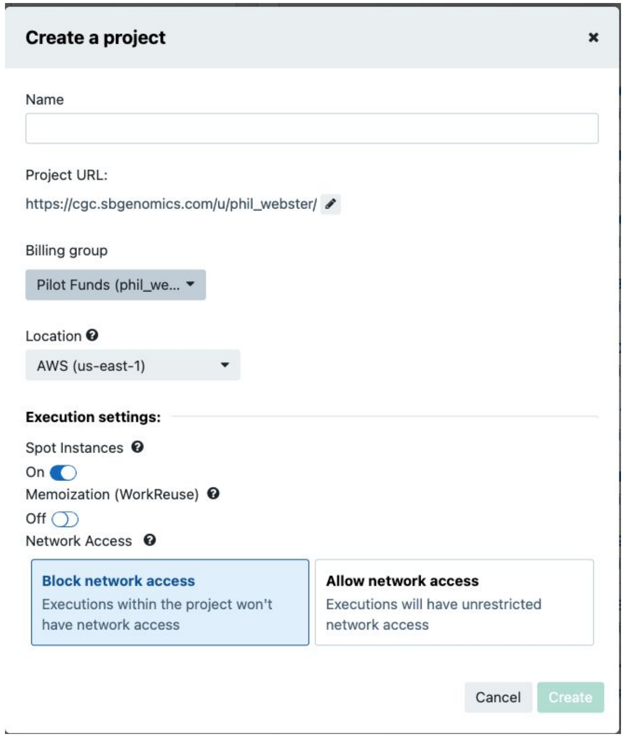
RNA-seq User Flow

Create a Project

Select/access data

Select/create tools

Create and run analysis



Create a project

Name

Project URL:
https://cgc.sbgenomics.com/u/phil_webster/

Billing group
Pilot Funds (phil_we...

Location
AWS (us-east-1)

Execution settings:

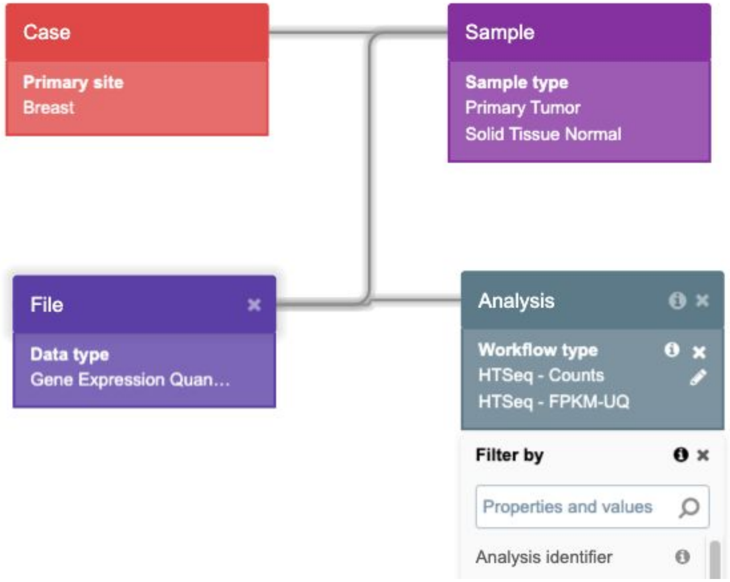
Spot Instances
On

Memoization (WorkReuse)
Off

Network Access

Block network access
Executions within the project won't have network access

Allow network access
Executions will have unrestricted network access



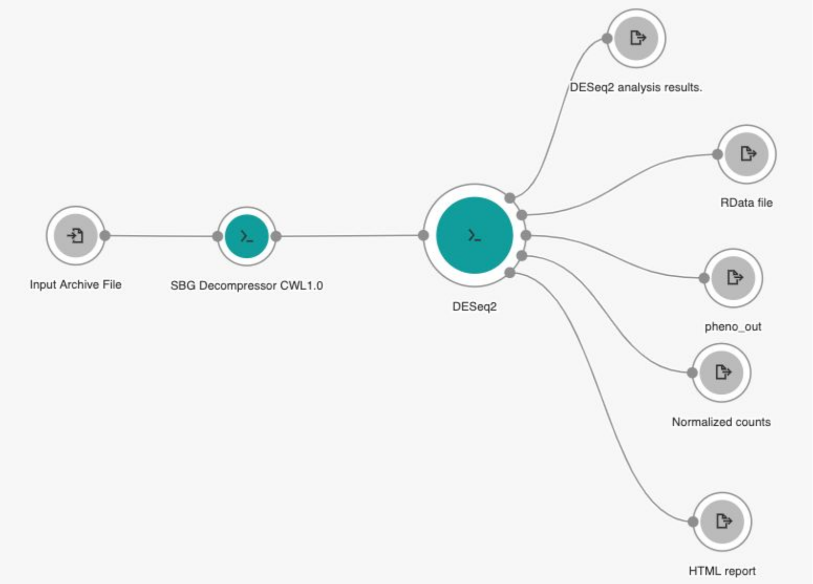
Case
Primary site
Breast

Sample
Sample type
Primary Tumor
Solid Tissue Normal

File
Data type
Gene Expression Quan...

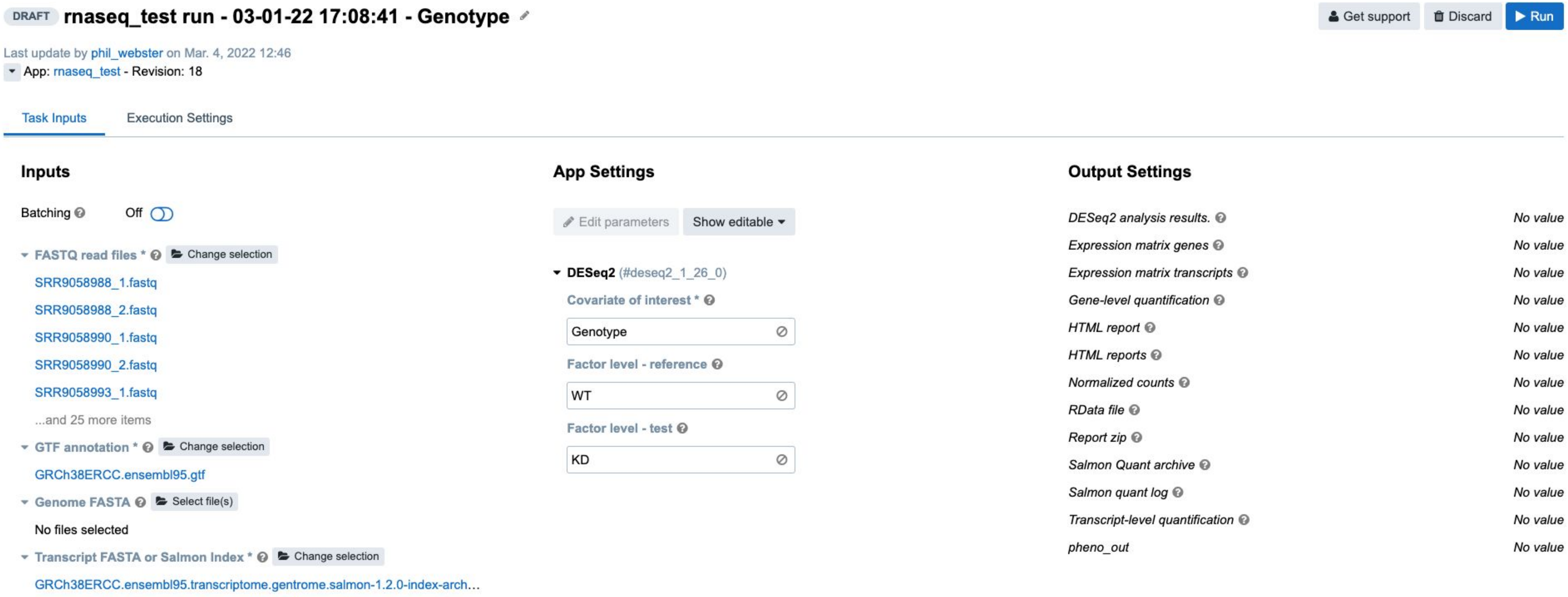
Analysis
Workflow type
HTSeq - Counts
HTSeq - FPKM-UQ

Filter by
Properties and values
Analysis identifier



RNA-seq User Flow

Create a Project → Select/access data → Select/create tools → Create and run analysis



DRAFT **rnaseq_test run - 03-01-22 17:08:41 - Genotype**

Last update by phil_webster on Mar. 4, 2022 12:46

App: rnaseq_test - Revision: 18

Inputs

Batching Off

FASTQ read files *

SRR9058988_1.fastq
SRR9058988_2.fastq
SRR9058990_1.fastq
SRR9058990_2.fastq
SRR9058993_1.fastq
...and 25 more items

GTF annotation *

GRCh38ERCC.ensembl95.gtf

Genome FASTA

No files selected

Transcript FASTA or Salmon Index *

GRCh38ERCC.ensembl95.transcriptome.gentrome.salmon-1.2.0-index-arch...

App Settings

DESeq2 (#deseq2_1_26_0)

Covariate of interest *: Genotype

Factor level - reference: WT

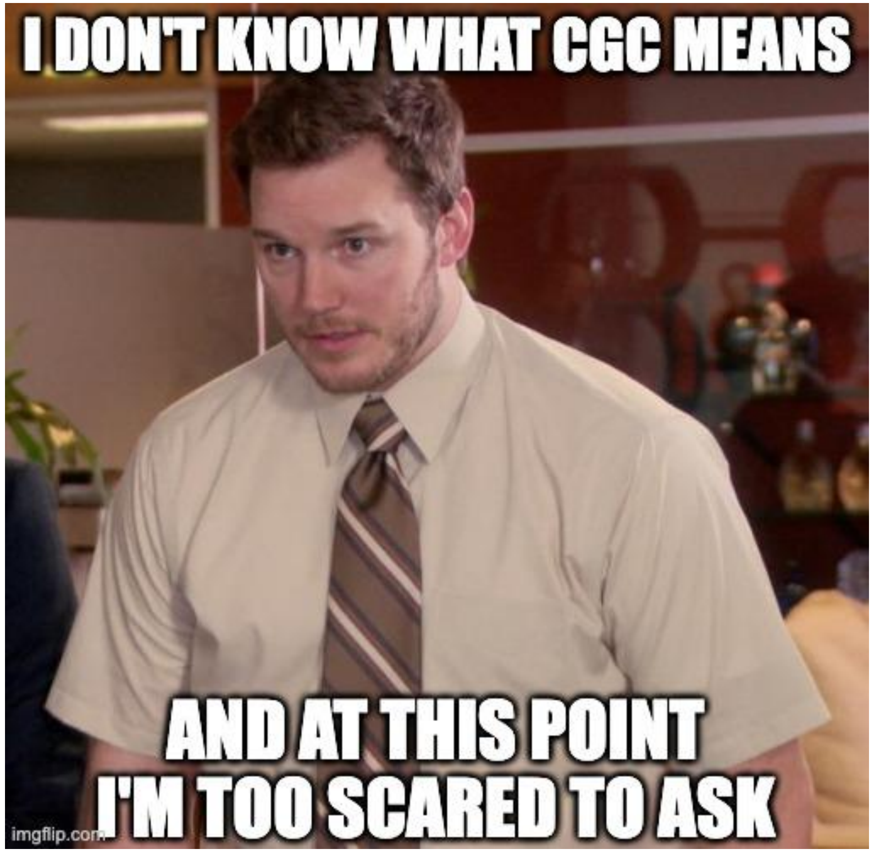
Factor level - test: KD

Output Settings

DESeq2 analysis results.	*No value*
Expression matrix genes	*No value*
Expression matrix transcripts	*No value*
Gene-level quantification	*No value*
HTML report	*No value*
HTML reports	*No value*
Normalized counts	*No value*
RData file	*No value*
Report zip	*No value*
Salmon Quant archive	*No value*
Salmon quant log	*No value*
Transcript-level quantification	*No value*
pheno_out	*No value*

RNA-seq live demo

Ask Questions!



Have questions? Contact us via email or attend office hours

Attend Office Hours every week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday

Manisha Ray

manisha.ray@sevenbridges.com

Zelia Worman

zelia.worman@sevenbridges.com

Phil Webster

phil.webster@sevenbridges.com

CANCER GENOMICS CLOUD
SEVEN BRIDGES

https://www.cancergenomicscloud.org

Q&A and Discussion

CANCER GENOMICS CLOUD