**Objective: Carry out RNA-seq analysis from human tissue.**

**The RNA-seq assignment introduces you into a user friendly interface software that is free and aids in the analyses of multiple types of sequencing data. For this class you will contrast the transcriptome of two types of human tissue.**

**RNA-seq assignment 1 will introduce you to the software, then you will upload the transcriptome file and annotations file, map your Fastqs files to the transcriptome, manipulate the results file to create a matrix and extract information.**

**Advice: Please start this assignment early! The software is free and can take a very long time to finish running.**

Ensure that you have a good and stable internet connection and enough space on your computer for this assignment (at least 100 Mb).

You do not need to download any software! You are going to be running everything on Galaxy (usegalaxy.org)

**Step 1:** **Create your account:**

Go to <https://usegalaxy.org/> click on users and register your own account. Check your email and follow directions to validate your account.

**Step 2: Upload transcriptome and annotations file:**

After logging in at galaxy select ‘get data’ option in the toolbar on the left. Select ‘upload file’ and on the bottom select ‘paste/fetch data’, paste the link below and click ‘start’:

<ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_32/gencode.v32.pc_transcripts.fa.gz>

Repeat the above steps for this link:

<ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_32/gencode.v32.basic.annotation.gtf.gz>

**\*\*\*\*Successfully submitted files are under “History”. The files will turn green when ready to use. The uploading of the files can take a few hours. If the files do not turn green after 8 hours, cancel the upload and restart again.\*\*\*\*\***

**Step 3: Uploading Fastq files.**

In a separate web browser tab go to <https://www.encodeproject.org/>

a. In ‘Search Encode portal’ type in ***your experiment data ID (ENCSR061RDC and ENCSR669GBC)***

b. Click on result and scroll down to the ***Files*** section (This is around halfway down the page.)

c. Under ‘**Files**’ section click in the tab ‘**File details’**

d. Right click on the download symbol *(arrow pointing to a bar)* next to the Accession number for *Isogenic replicate 1* SE101nt *(Single-End 50 nucleotide)* ***read 1*** (should be the top file in the “**Raw sequencing data**” section) and **copy link address.**

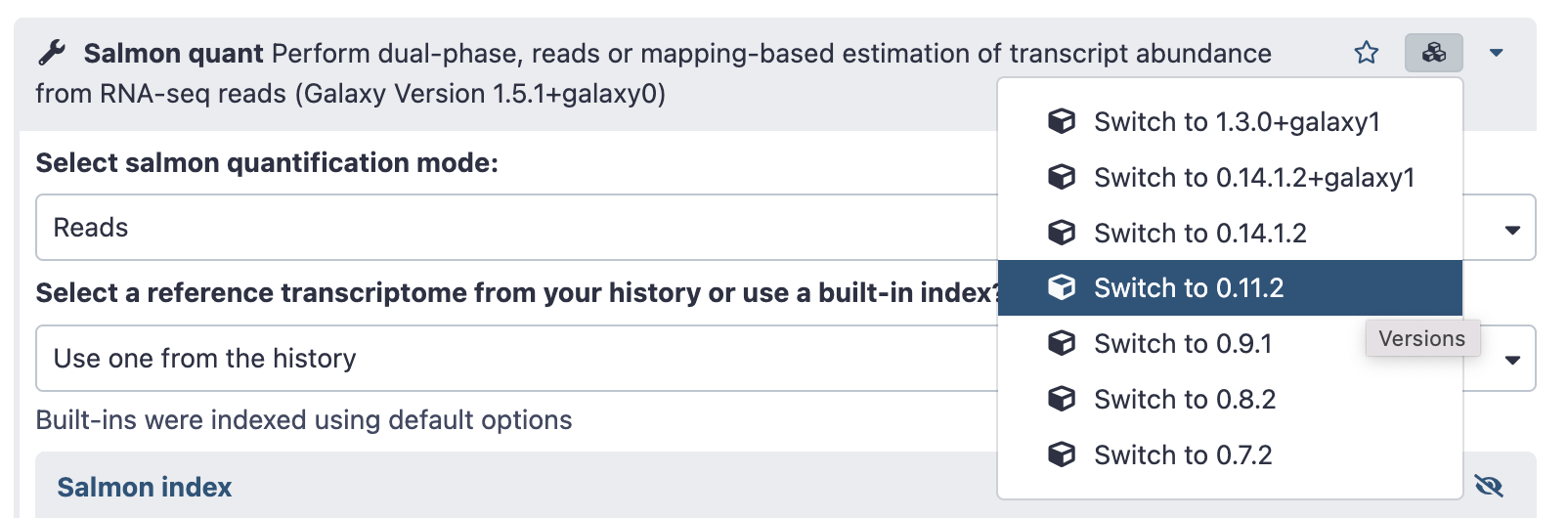
d. Repeat these steps for the second sample.

**\*\*\*\*Successfully submitted files are under “History”. Keep track of all the files you upload to usegalaxy.org. The files will turn green when ready to use. Uploading all these files can take a few hours. If the files do not turn green after 8 hours, cancel the upload and restart again.\*\*\*\***

**Step 4: Map reads (in fastq files) to transcriptome and quantify gene expression.**

The next step is mapping the processed reads to the genome. The major challenge when mapping RNA-seq reads is that they come from RNA which includes cross splice junction boundaries. Splice junctions are not present in a genome's sequence, and hence typical Next Generation Sequencer (NGS) mappers such as Bowtie and BWA are not ideal without modifying the genome sequence. Instead, it is better to use a mapper such as [Salmon](http://tophat.cbcb.umd.edu/) that is designed to map RNA-seq reads.

1. In usegalaxy.com go to ‘GENOMICS ANALYSIS’ then click on **“RNA-seq”** then **“Salmon”. Make sure to switch the version to 0.11.2 as in the image below:**

****

b. There a several options for Salmon, first we have to choose the correct transcriptome and lastly the correct Fastq file. We will map to transcriptome the fastq files uploaded in step 3.

i. **Select a reference transcriptome from your history or use a built-in index?**

- Use one from the History

**ii. Select the reference transcriptome**

- Select file: “gencode.v32.transcripts.fa.gz”

**iii. Is this library mate-paired?**

- Select “Single-end”

**iv. FASTQ/FASTA file**

- Select the file which correspond to “*Isogenic replicate 1: read1”*

**v. File containing a mapping of transcripts to genes**

- Select your GTF file (gencode.v32.basic.annotation.gtf.gz)

c. Scroll to the bottom of the page and click on **“Execute”.** A new file will be generated under **“History”.** The file will be ready to use when **green. This step of mapping the reads can take several hours.**

d. Repeat Step 4 b and c for your other fastq

**Step 5. Change the format of the files output from Step 4**

The output files for Salmon are in the .tabular format which will not be recognized by Excel or Google Sheets. The output file format will be changed to txt.

a. For the file generated from step 4 “Salmon on data ##, data ## and data ## (Quantification)” click on the “**pencil icon” (edit attributes).**

b. Select ‘**Datatypes**’ tab and set the “**New Type**” to “**txt**” and press Change datatype.

c. **Wait until this new task turn green.**

d. You can now download this file. Click on the “Salmon on data ##, data ## and data ## (Quantification)”, then click on the **“floppy disk icon”**. The file will download to your computer.

e. Import the file into either Excel or a Google Sheet.

**Step 6. Interpret data from quantification file using Excel or Google Sheets.**

The quantification file generated by Salmon has information about how much a particular gene is expressed in your sample. We have to use Excel or Google sheets to manipulate the data and extract information to answer the questions below.

a. Using Excel or Google sheet import your quantification file using tabs as the delimiter.  
b. file format:  
 -The 1st column = **name** contains all of the metadata about a gene. ***The information in this column is important for differentiating if a row correspond to a protein coding gene, a pseudo-gene, a lncRNA, etc.***

-The 2nd and 3rd columns = **length** and **effective length** contain the gene length (not important)  
 -The 4th column = **TPM** is transcript per million a fancy version of RPKM. This is important for question 3.  
 -The last column = **NumReads** ***contains the number of reads that mapped to that gene.***

**Step 7. Count how many reads you have for each Fastq file.**

In lecture we went over the format of a Fastq file, and the definition of one read. For this section, you will count how many lines you have in each Fastq file, and answer questions below.

a. You have to “uncompressed” you Fastq files, in other words change its format, so the software can count the lines**. You have to do the following steps for both Fastq files from step 3.**

i. On the files you uploaded in step 3, click on “pencil button (edit attributes)”.

ii. Click on **“Convert”** tab

iii. Under **“Datatypes”** choose “fastq”

iv. Click on **“Save”**

v. A new file will be generated under history *“Convert compressed file to uncompressed. on data ##”*

b. Counting the number of lines for the uncompressed **files** generated in step 7a.

i. under **“general text tool”,** click on **“Text manipulation”**

ii. click on “**Line/Word/Character count of a dataset”**

iii**. Text file**

- Choose one of the uncompressed files

iv. **Desired values**

- deselect word count and character count.

- Make sure Line count is checked .

v. Click **“execute”**

vi. Repeat for the other uncompressed file.

(1) Answer the questions bellow.

(2) Combine the screenshot and answers in **one PDF** document and submit on canvas

**Questions:**

1. What is your experimental sample (biological specimen)? What kind of cells are you analyzing? Include all the details offered in the website such as age, gender and type of tissue.

2. How many lines were in your fastq files from ENCODE? How many reads is that?

3. How many total reads were mapped by Salmon for each replicate? What percentage of reads were mapped for each biological replicate?

4. What are the top 20 ***genes*** (excluding mitochondrial) expressed in **each replicate** (TPM)? Look up 3 of the genes for each replicate on genecards.org. Explain the function of each gene in your own words, including information that is relevant to its function. What would happen if these genes acquired mutations over the lifespan of a person? What would happen if these genes were deleted?