Part 1 - Obtaining TDF files for IGV

Author: Daniel Ramírez, 2022

--- STAGING WORKING AREA IN SCRATCH ---

Go to ***your*** scratch user directory

/scratch/Users/dara6367/SR2022/day5

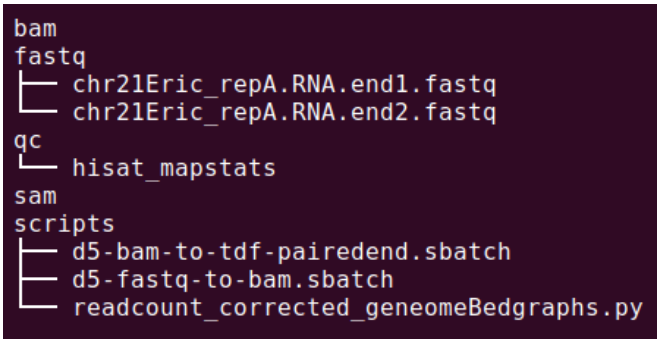
Next, to catch up to speed to where you should have ended your work for Day 4, copy these two RNA-seq paired-end FASTQ files to a new Day 5 staging area in your scratch user directory:

/scratch/Shares/public/sread2022/data_files/day5/fastq/
chr21Eric_repA.RNA.end1.fastq
chr21Eric_repA.RNA.end2.fastq

And copy the following 3 scripts as well:

/scratch/Shares/public/sread2022/scripts/day5/
d5-bam-to-tdf-pairedend.sbatch
d5-fastq-to-bam.sbatch
readcount_corrected_geneomeBedgraphs.py

Copy all necessary files to your own scratch user directory, in my case it will be to this directory and the files and subdirectories such look like this below.
Do not forget to create the necessary new subdirectories. Including somewhere to store eofiles!



The script **d5-fastq-to-bam.sbatch** can quickly be used to process input FASTQ files onto output BAM files, just as you did for Day 4. If you already have correctly generated BAM files from Day 4, then you can skip using script **d5-fastq-to-bam.sbatch**. Otherwise, if you are in need of obtaining a BAM file for today, then go ahead and edit **d5-fastq-to-bam.sbatch** appropriately to run in your terminal. This script should run fairly quickly, as the input files are

only a subset of a real RNA-seq dataset, but only with reads of one of the relatively small chromosome 21.

--- EDITING AND RUNNING SCRIPTS ---

Edit **d5-fastq-to-bam.sbatch** so that it has the correct paths. Specifically on lines 2, 4, 10, 11, 22, 23, 24, and 25.

```
 1 #!/bin/bash
 2 #SBATCH --job-name=<NAME OF JOB>                          # Job name
 3 #SBATCH --mail-type=ALL                                  # Mail events (NONE, BEGIN, END, FAIL, ALL)
 4 #SBATCH --mail-user=<YOUR EMAIL ADDRESS>                 # Where to send mail
 5 #SBATCH --nodes=1                                        # Numbers of nodes
 6 #SBATCH --ntasks=4                                       # Number of CPU (tasks)
 7 #SBATCH --time=00:10:00                                  # Time limit hrs:min:sec
 8 #SBATCH --partition=short                                # Partition/queue requested on server
 9 #SBATCH --mem=10gb                                       # Memory limit
10 #SBATCH --output=/YOUR/EOFILES/PATH/%x_%j.out
11 #SBATCH --error=/YOUR/EOFILES/PATH/%x_%j.err
12
13 ################ RUNNING SCRIPT IN TERMINAL #############################################
14
15 # Just copy and paste this command in the same directory where sbatch script is located.
16 # sbatch --export=FILENAME=chr21Eric_repA.RNA d5_fastq_to_bam.sbatch
17
18 ################ SET REQUIRED VARIABLES #################################################
19 ## the fastq files will be used as input to fastqc.
20 ## output will be a fastqc file used to assess quality
21
22 INDIR=/PATH/TO/FASTQ/FILES/IN/YOUR/SCRATCH/DIRECTORY
23 BAM=/PATH/TO/BAM/OUTPUT/DIRECTORY
24 SAM=/PATH/TO/SAM/OUTPUT/DIRECTORY
25 QC=/PATH/TO/HISAT2/QC/
26 INDICES=/scratch/Shares/public/genomes/hisatfiles/hg38/HISAT2/genome
27
```

To run the script, you can pass the value of the variable FILENAME as you type your sbatch command, as follows:

```
sbatch --export=FILENAME=chr21Eric_repA.RNA d5-fastq-to-bam.sbatch
```

If your script ran successfully, your directory structure should have been populated as follows:

```
bam
├── chr21Eric_repA.RNA.sorted.bam
└── chr21Eric_repA.RNA.sorted.bam.bai
fastq
├── chr21Eric_repA.RNA.end1.fastq
└── chr21Eric_repA.RNA.end2.fastq
qc
└── hisat_mapstats
    └── chr21Eric_repA.RNA.hisat2_mapstats.txt
sam
└── chr21Eric_repA.RNA.sam
scripts
├── d5-bam-to-tdf-pairedend.sbatch
├── d5-fastq-to-bam.sbatch
└── readcount_corrected_geneomeBedgraphs.py
```

Now we have BAM files, just as we learned during Day 4.

For today, we want to convert this big BAM file onto a smaller TDF file, using the script **d5-bam-to-tdf-pairedend.sbatch** (which you have to edit) which itself automatically uses the python script **readcount_corrected_geneomeBedgraphs.py** (which you do *not* have to edit).

The script **d5-bam-to-tdf-pairedend.sbatch** is fairly long, with almost 180 lines. Each section is separated by comments briefly explaining what each section does. You only need to edit certain lines (marked < >), the ones that define the SBATCH parameters, and the script variables. You do not have to change anything else in the script (below line 36) to make it work.

```
#!/bin/bash
#SBATCH --job-name=<NAME OF JOB>
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<YOUR EMAIL ADDRESS>
#SBATCH --nodes=1
#SBATCH --ntasks=8
#SBATCH --mem=10gb
#SBATCH --time=00:10:00
#SBATCH --output=/YOUR/EOFILES/PATH/%x_%j.out
#SBATCH --error=/YOUR/EOFILES/PATH/%x_%j.err

################################################################################################
############## LOAD REQUIRED MODULES ###########################################################

module load samtools/1.8
module load bedtools/2.25.0
module load python/2.7.14/
module load igvtools/2.3.75

################################################################################################
############### SET REQUIRED VARIABLES #########################################################

BAMroot=<NAME OF DATASET> # E.g. chr21Eric_repZ.RNA
inDir=/scratch/Users/<YOUR USERNAME>/SR2022/day5/bam
outDir=/scratch/Users/<YOUR USERNAME>/SR2022/day5
pythonScript=/scratch/Users/<YOUR USERNAME>/SR2022/day5/scripts/readcount_corrected_geneomeBedgraphs.py
genomeFasta=/scratch/Shares/public/genomes/hisatfiles/hg38/hg38.fa
genomeSizes=/scratch/Shares/public/genomes/hisatfiles/hg38/hg38.chrom.sizes
BAM=${inDir}/${BAMroot}.sorted.bam

BEDGRAPHdir=${outDir}/bedgraphForTdf
TDFdir=${outDir}/tdf
STATSdir=${outDir}/stats
mkdir -p ${BEDGRAPHdir} ${TDFdir} ${STATSdir}

################################################################################################
```

If your script ran successfully, your directory structure should have been populated as shown below. Observe the single TDF file in its own directory, and all the intermediary files needed to make a correct and depth-normalized BEDGRAPH file.

```
bam
├── chr21Eric_repA.RNA.sorted.bam
└── chr21Eric_repA.RNA.sorted.bam.bai
bedgraphForTdf
├── chr21Eric_repA.RNA.bed
├── chr21Eric_repA.RNA.BedGraph
├── chr21Eric_repA.RNA.mp.BedGraph
├── chr21Eric_repA.RNA.neg.Bedgraph
├── chr21Eric_repA.RNA.neg.Bedgraphcol
├── chr21Eric_repA.RNA.pairfirst.bam
├── chr21Eric_repA.RNA.pairfirst.neg.bed
├── chr21Eric_repA.RNA.pairfirst.neg.BedGraph.sort
├── chr21Eric_repA.RNA.pairfirst.pos.bed
├── chr21Eric_repA.RNA.pairfirst.pos.BedGraph.sort
├── chr21Eric_repA.RNA.pairsecond.bam
├── chr21Eric_repA.RNA.pairsecond.neg.bed
├── chr21Eric_repA.RNA.pairsecond.neg.BedGraph.sort
├── chr21Eric_repA.RNA.pairsecond.pos.bed
├── chr21Eric_repA.RNA.pairsecond.pos.BedGraph.sort
├── chr21Eric_repA.RNA.pos.Bedgraph
└── chr21Eric_repA.RNA.pos.Bedgraphcol
fastq
├── chr21Eric_repA.RNA.end1.fastq
└── chr21Eric_repA.RNA.end2.fastq
qc
└── hisat_mapstats
    └── chr21Eric_repA.RNA.hisat2_mapstats.txt
sam
└── chr21Eric_repA.RNA.sam
scripts
├── d5-bam-to-tdf-pairedend.sbatch
├── d5-fastq-to-bam.sbatch
├── igv.log
└── readcount_corrected_geneomeBedgraphs.py
stats
├── chr21Eric_repA.RNA.bam.flagstat
└── chr21Eric_repA.RNA.bam.flagstat.err
tdf
└── chr21Eric_repA.RNA.tdf
```

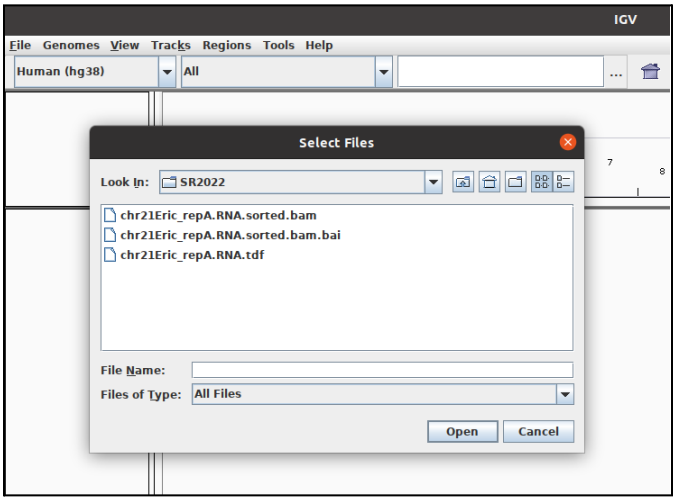
To see the TDF file, you will have to transfer the TDF file that you have just created that lives in the computer cluster, to your own local computer which has the IGV software installed. Once in your computer (see username is now “daniel” and not “dara6367”), it may be a good idea to make a special folder for the materials of this workshop. You can compare the BAM and TDF files, and see which file tells you what information.

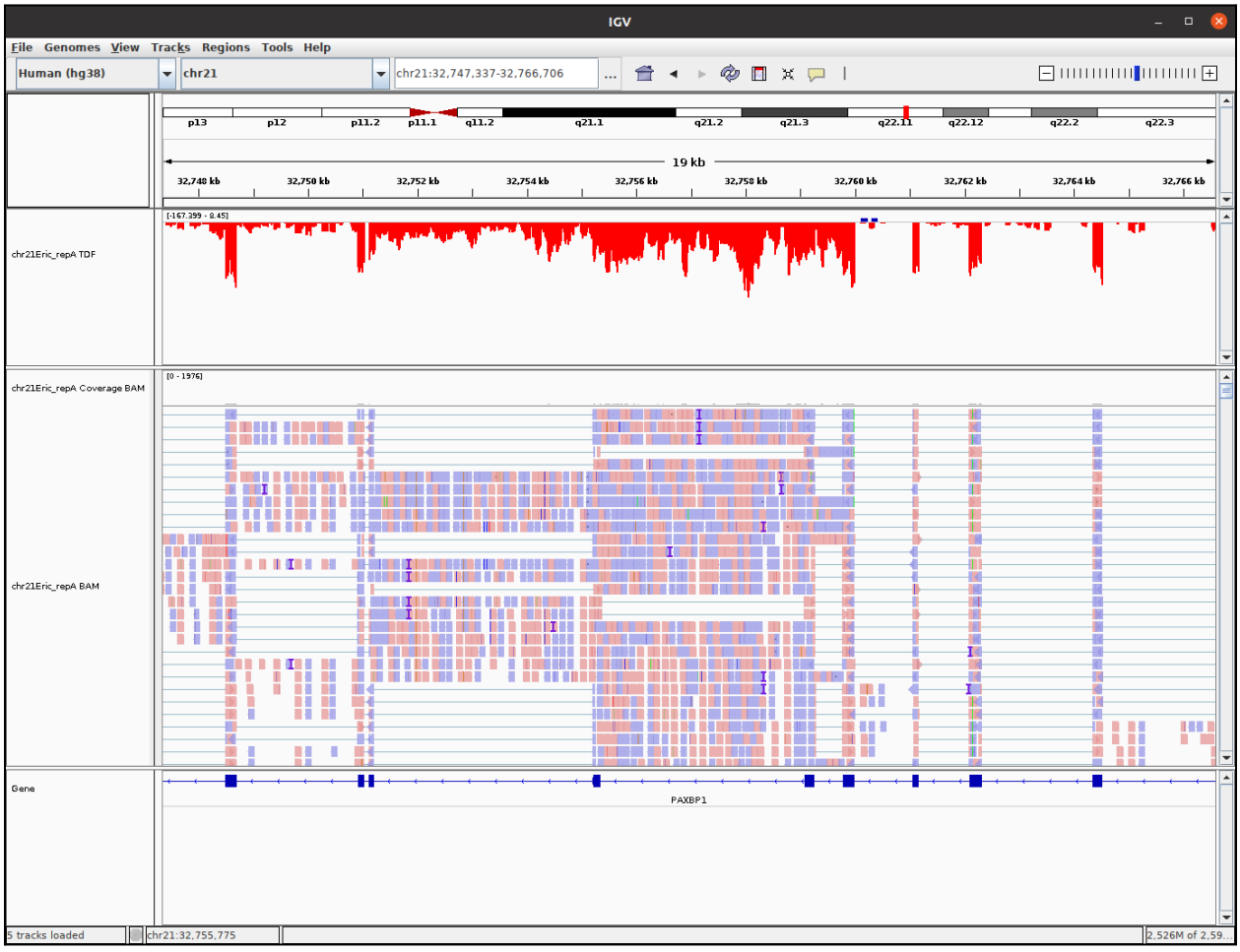
```
daniel@nebuchadnezzar:~$ pwd
/home/daniel
daniel@nebuchadnezzar:~$ mkdir SR2022
daniel@nebuchadnezzar:~$ cd SR2022/
daniel@nebuchadnezzar:~$ rsync -P daramirez@3.136.149.251:/scratch/Users/dara6367/SR2022/day5/bam/*bam* .
chr21Eric_repA.RNA.sorted.bam
     48,667,254 100%    8.35MB/s    0:00:05 (xfr#1, to-chk=1/2)
chr21Eric_repA.RNA.sorted.bam.bai
      1,506,208 100%    2.03MB/s    0:00:00 (xfr#2, to-chk=0/2)
daniel@nebuchadnezzar:~$ rsync -P dara6367@fiji.colorado.edu:/scratch/Users/dara6367/SR2022/day5/tdf/*tdf* .
chr21Eric_repA.RNA.tdf
      2,514,804 100%    3.51MB/s    0:00:00 (xfr#1, to-chk=0/1)
```

To view the BAM and TDF files, we can use IGV:

```
daniel@nebuchadnezzar:~$ cd ~/Downloads/IGV_Linux_2.8.2/
daniel@nebuchadnezzar:~$ sh igv.sh
echo Using bundled JDK.
WARNING: package com.sun.java.swing.plaf.windows not in java.desktop
WARNING: package sun.awt.windows not in java.desktop
openjdk version "11.0.5" 2019-10-15
OpenJDK Runtime Environment AdoptOpenJDK (build 11.0.5+10)
OpenJDK 64-Bit Server VM AdoptOpenJDK (build 11.0.5+10, mixed mode)
```

Open both the BAM and TDF file with IGV, and customize their tracks at your preference. As a reminder, the data is from chromosome 21.





Part 2 - Week 1 assessment

Pick any of the other available day5 fastq files (or do them all together with a loop!). Make a new sbatch script that takes as input these paired FASTQ files, and processes them, checks their quality by using FASTQC, trims them, maps them using HISAT2, converts them to BAM, and makes depth-normalized TDF files.

Condition	Biological Replicates	Read Pairs
Chr21 Eric	RepA	End1
		End2
	RepB	End1
		End2
	RepC	End1
		End2
Chr21 Ethan	RepA	End1
		End2
	RepB	End1
		End2
	RepC	End1
		End2

/scratch/Shares/public/sread2022/day5/fastq/

```
chr21Eric_repA.RNA.end1.fastq
chr21Eric_repA.RNA.end2.fastq
chr21Eric_repB.RNA.end1.fastq
chr21Eric_repB.RNA.end2.fastq
chr21Eric_repC.RNA.end1.fastq
chr21Eric_repC.RNA.end2.fastq
chr21Ethan_repA.RNA.end1.fastq
chr21Ethan_repA.RNA.end2.fastq
chr21Ethan_repB.RNA.end1.fastq
chr21Ethan_repB.RNA.end2.fastq
chr21Ethan_repC.RNA.end1.fastq
chr21Ethan_repC.RNA.end2.fastq
```