RNA Seq Pipeline Docs | Vijay

meningioma AND [RNA-Seq]

Data Sample for Preliminary Set Up:

GSE136661: GSM4054990 - WHO III Meningioma; cDNA
SRR: SRX6777597

Obtaining Data Sample: http://www.ebi.ac.uk/ena/data/view/SRX6777597

Anaconda Environment: pipeline

Source: https://docs.anaconda.com/free/anaconda/packages/using-r-language/

Creating an environment with R

1. Download and install Anaconda.
2. Create a new conda environment with all the r-essentials conda packages built from CRAN:
   ```
   conda create -n pipeline r-essentials r-base
   ```
3. Activate the environment:
   ```
   conda activate pipeline
   ```
4. List the packages in the environment:
   ```
   conda list
   ```

The list shows that the package r-base is installed and r is listed in the build string of the other R packages in the environment.

Anaconda Navigator, the Anaconda graphical package manager and application launcher, creates R environments by default.

Reproducing:

https://stackoverflow.com/a/64094923

FastQC

Install fastqc Using apt-get on Ubuntu

Update apt database with `apt-get` using the following command.

```
sudo apt-get update
```

After updating apt database, We can install `fastqc` using `apt-get` by running the following command:

```
sudo apt-get -y install fastqc
```

Anaconda Installation of fastqc

Source: https://anaconda.org/bioconda/fastqc

```
conda install -c bioconda fastqc
```

Trimmomatic

Conda Installation

Source: https://anaconda.org/bioconda/trimmomatic

```
conda install -c bioconda trimmomatic
```

Hisat2

Conda Installation

Source: https://anaconda.org/bioconda/hisat2

```
conda install -c bioconda hisat2
```

Stringtie

Conda Installation

Source: https://anaconda.org/bioconda/stringtie

```
conda install -c bioconda stringtie
```

edgeR

Conda Installation

Source:https://anaconda.org/bioconda/bioconductor-edger

```
conda install -c bioconda bioconductor-edger
```

Deseq2

Conda Installation (**)

Source: https://anaconda.org/bioconda/bioconductor-deseq2

```
conda install -c bioconda bioconductor-deseq2
```

Samtools

Conda Installation

Source: https://anaconda.org/bioconda/samtools

```
conda install -c bioconda samtools
```

Grch38 Reference Genome

1. Download and unzip reference:
    - `wget http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz`
    - `gunzip hg38.fa.gz`
2. Getting the annotated file: https://www.biostars.org/p/174331/

Executing Trimmomatic

Command Template:

```
trimmomatic PE -threads 1 ./data/SRR1672666_1.fastq.gz
./data/SRR1672666_2.fastq.gz -baseout ./trim/SRR1672666.fastq.gz
ILLUMINACLIP:./genome/genome.fa:4:30:10 MINLEN:30
```

Faced problem: Out of Memory Error on aforementioned sample

```
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
        at java.base/java.util.Arrays.copyOfRangeByte(Arrays.java:3863)
        at java.base/java.util.Arrays.copyOfRange(Arrays.java:3854)
        at java.base/java.lang.String.<init>(String.java:4784)
        at java.base/java.lang.String.<init>(String.java:1507)
        at java.base/java.lang.StringBuilder.toString(StringBuilder.java:475)
        at org.usadellab.trimmomatic.fasta.FastaParser.parseOne(FastaParser.java:48)
        at org.usadellab.trimmomatic.fasta.FastaParser.next(FastaParser.java:71)
        at
org.usadellab.trimmomatic.trim.IlluminaClippingTrimmer.loadSequences(IlluminaClippingTrimmer.java:121)
        at
org.usadellab.trimmomatic.trim.IlluminaClippingTrimmer.makeIlluminaClippingTrimmer(IlluminaClippingTrimmer.java
:71)
        at org.usadellab.trimmomatic.trim.TrimmerFactory.makeTrimmer(TrimmerFactory.java:32)
        at org.usadellab.trimmomatic.Trimmomatic.createTrimmers(Trimmomatic.java:59)
        at org.usadellab.trimmomatic.TrimmomaticPE.run(TrimmomaticPE.java:552)
        at org.usadellab.trimmomatic.Trimmomatic.main(Trimmomatic.java:80)
```

Solution: Within the environment execute:

```
export _JAVA_OPTIONS=”-Xmx16g”
```

Here 16g instructs Java to offer a default heap size of 16GB limit on the DRAM

Faced Problem: Prints a lot of AGCT stuff on the console
Solution: Add the “`-quiet`” option to shut it up

IMPORTANT!: The file given to trimmomatic as fasta is the TrueSeq3 fasta for adapter cropping. It has nothing to do with genome.fa

Executing Hisat2

Source: https://www.reneshbedre.com/blog/hisat2-sequence-aligner.html
Building the Genome Index from the fasta:

```
hisat2-build -p 6 -f <filename>.fa <basename-for-output>
```

Problem: Trimmomatic gave 4 files, how do we run hisat on this??
Solution: Two possible methods:

1. [CURRENT IMPLEMENTATION] Just use hisat2 on the paired outputs, or
2. [CURRENTLY UNCONSIDERED] Use samtools to merge the result of the paired output with the unpaired ones. [Idea: https://www.biostars.org/p/360675/ and https://www.biostars.org/p/16533/]
    a. First generate the separate .sam files. Source:https://www.biostars.org/p/360675/
    b. Convert sam to bam https://www.biostars.org/p/107794/
    c. Now merge the bam files. Source: https://www.biostars.org/p/9864/
    d. Also see: https://www.biostars.org/p/362149/

Executing StringTie

Source: http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

Executing DESeq2 with RScript

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("DESeq2")
Error: With R version 3.5 or greater, install Bioconductor packages using
BiocManager; see https://bioconductor.org/install
Reference: https://www.bioconductor.org/install/
Installing BiocManager
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install()

Installing DeSeq2
BiocManager::install("DeSeq2")

Reference: https://lashlock.github.io/compbio/R_presentation.html

Make your own metadata if not available

dds <- DESeqDataSetFromMatrix(countData, colData = metadata, design = ~
Condition + Group)
```

Metadata for DESeq2

- To get metadata for all samples in an RNA-Seq dataset such as GSE136661 from the Gene Expression Omnibus (GEO), you can follow these steps:

- Go to the GEO website: https://www.ncbi.nlm.nih.gov/geo/
- In the search bar, enter the dataset identifier "GSE<xxxx>" and press Enter to search for the dataset.
- Click on the dataset link to access its homepage.
- Look for the "Series Matrix File(s)" section, which typically contains information about the metadata and expression data files. If available, you can download the series matrix file, which is usually in a tab-delimited format.
- Open the series matrix file in a text editor or spreadsheet software (e.g., Excel, Google Sheets).
- The file should contain metadata information for all samples in the dataset, including sample names, titles, characteristics, and other relevant information

With e-search utilities provided by SRA Toolkit, one can use the command:

```
esearch -db sra -query SRR10042623 | efetch -format runinfo
```

But this misses every data classification of importance.

However,

```
esearch -db bioproject -query "GSE136661" | elink -target biosample | efetch | head -20
```

Seems to give exactly what is needed.

Bulk Download of SRR fastq files:

Reference: https://www.ncbi.nlm.nih.gov/books/NBK179288/#_chapter6_Getting_Started_
Installing EDirect(NCBI E-Utilities)

```
>> export PATH=${HOME}/edirect:${PATH}
To set path for current terminal session
```

References: https://bioinformaticsworkbook.org/dataAcquisition/fileTransfer/sra.html#gsc.tab=0
Loading SRR.numbers for given PRJNA BioProject.
ISSUE: PRJNA662780, corresponding to GSE157783, has 11 samples but got 12 SRR files. Check!

Reference: https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit
Installing SRA-toolkit and adding path to binaries
Go inside `sratoolkit.3.0.0-ubuntu64/bin`
`And run command ```export PATH=$PATH:$PWD````

Refer: https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit

Reference: https://erilu.github.io/python-fastq-downloader/
Prefetch and fastq-dump

ISSUE: Conda installation and other installations of SRA Toolkit yield deprecated/obsolete versions of the toolkit. The green text has been used finally to achieve the purpose.

SAMN to SRR mapping

This seemed to be the major problem with metadata matching. The below command line argument suggested by ChatGPT maps the SAMN to the SRR.

```
esearch -db biosample -query "SAMNXXXXXXXX" | elink -target sra |
efetch -format docsum | xtract -pattern Runs -element Run@acc
```