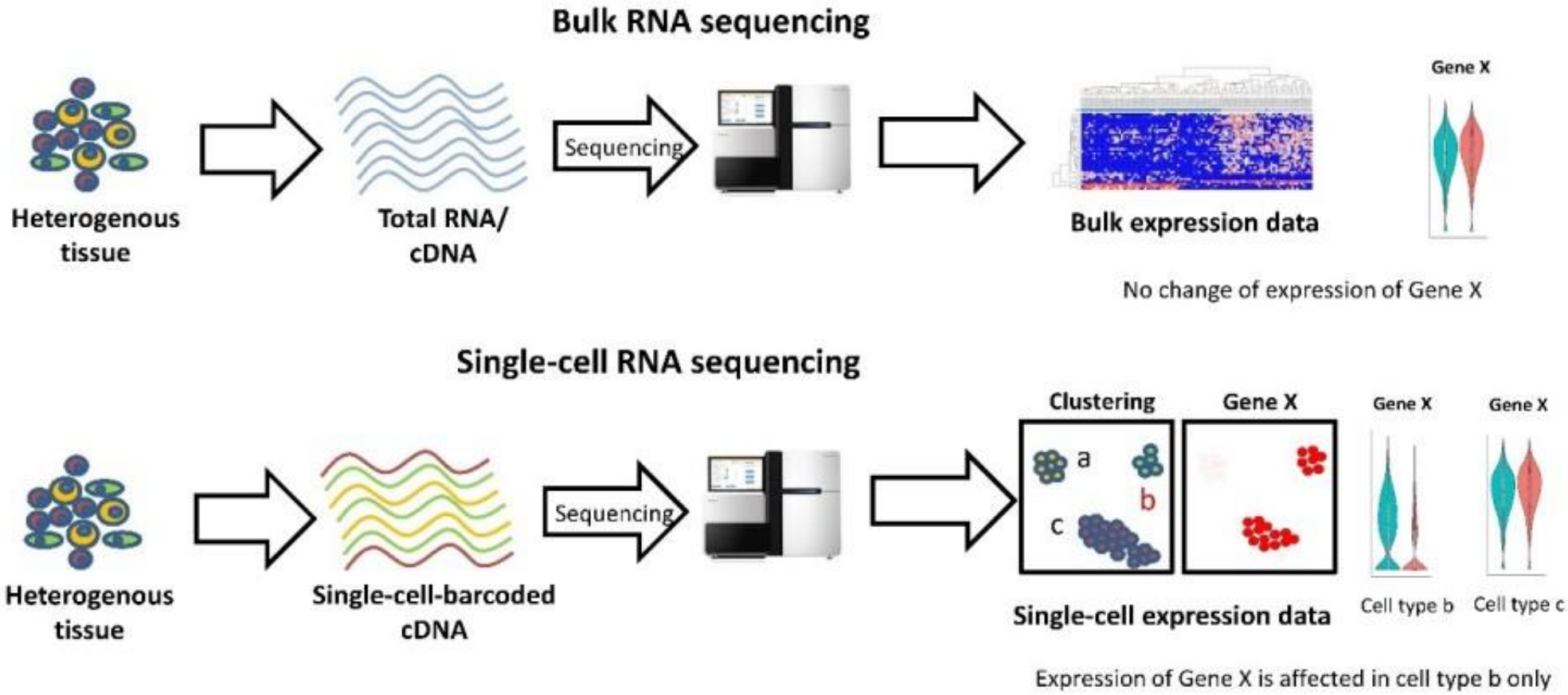
숭실대학교
Soongsil University

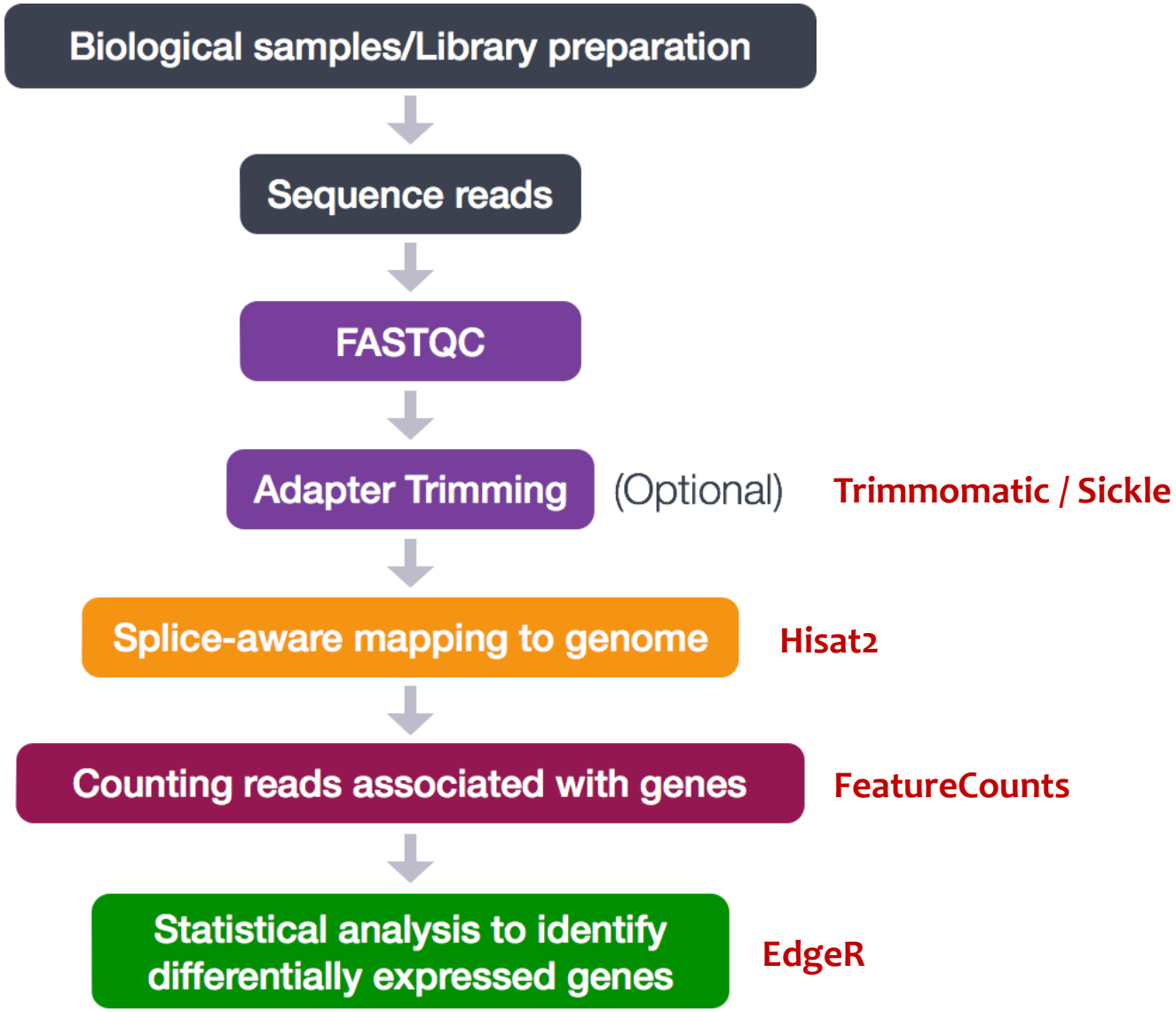
Bulk RNA-seq analysis pipeline 소개

숭실대학교 생명정보학과 석사과정 박정운

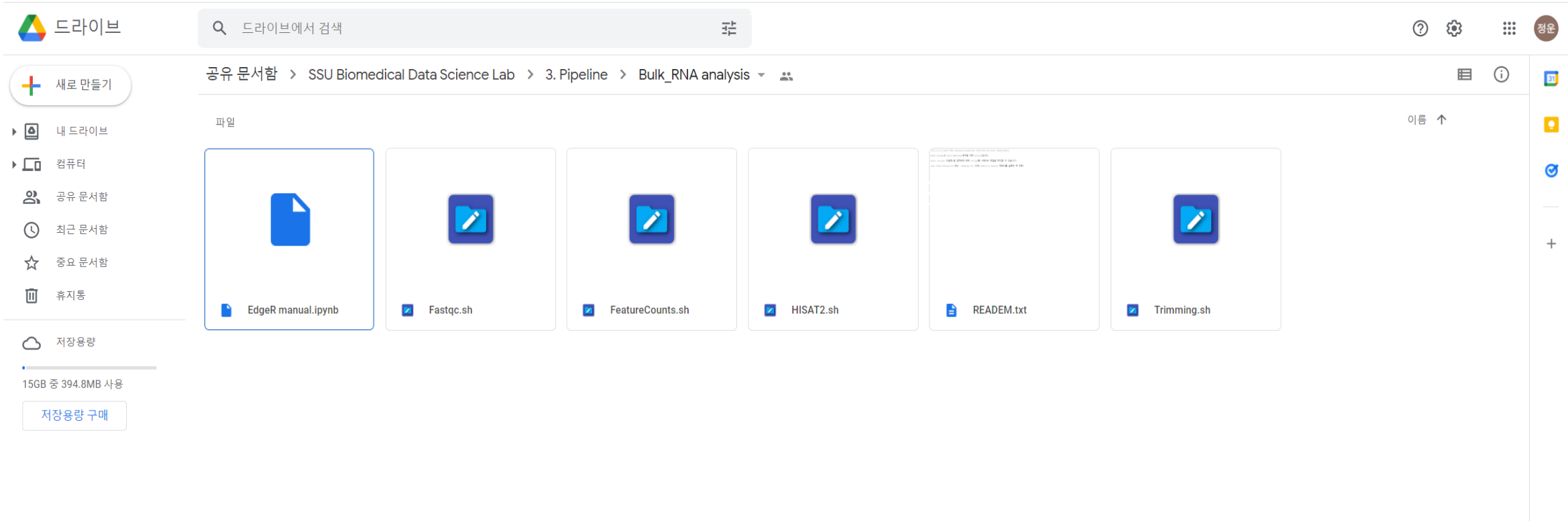
RNA-sequencing



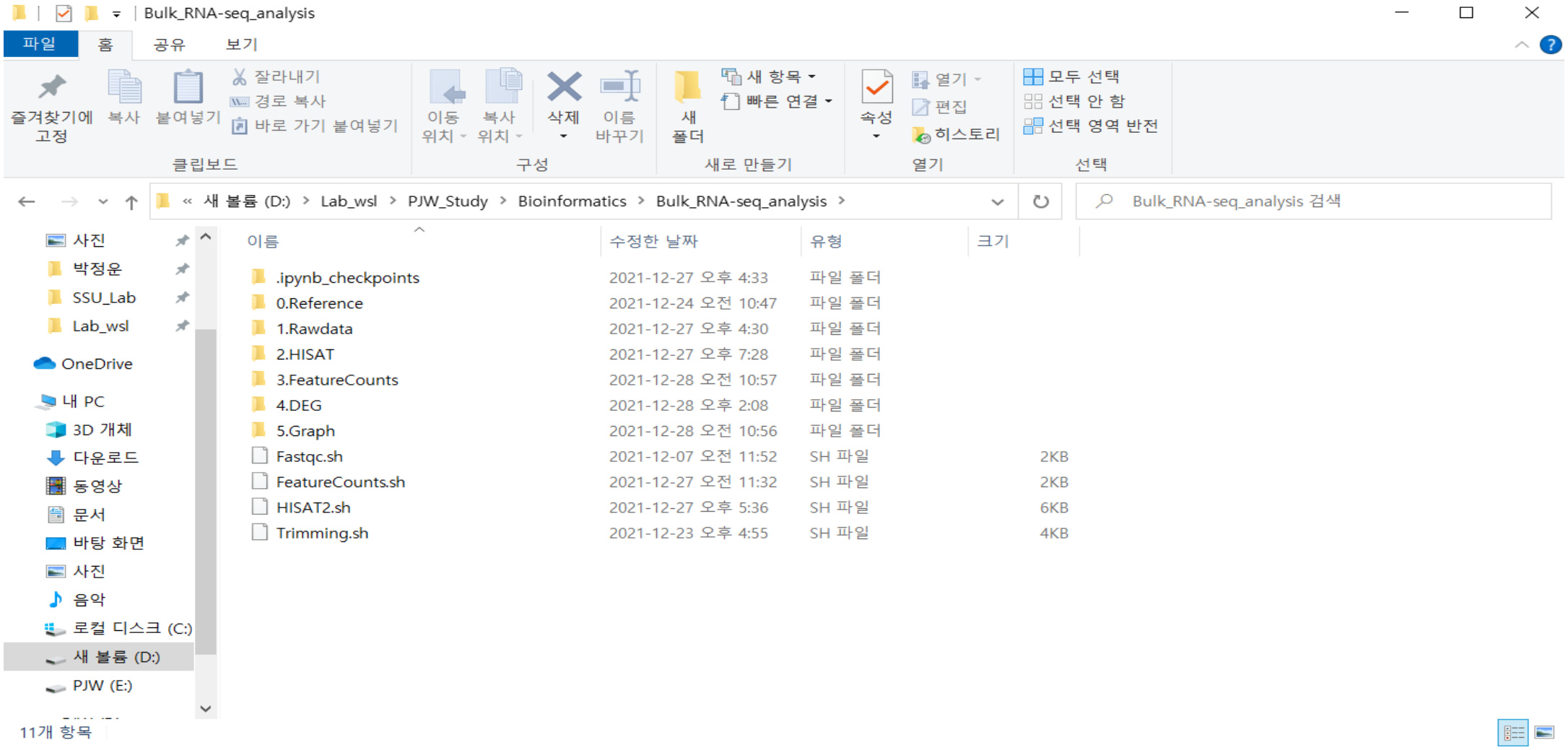
Bulk RNA-seq analysis pipeline



- https://drive.google.com/drive/folders/1ck9UFfl0dDGcwerMvyi5dgln_GgrJmkS
- RNA-seq preprocessing 과정은 bash script로 저장함. (Fastqc.sh, Trimming.sh, HISAT2.sh, FeatureCounts.sh)
- EdgeR을 이용한 DEG 분석은 jupyter notebook script에 저장함. (EdgeR manual.ipynb)



- 앞에 언급한 사이트로부터 bash script 파일들을 다운 받는다.
- 해당 bash script는 chmod +x (해당 script 파일명) 명령어를 사용하여 executable 권한을 부여한다!
- 다운을 받은 이후에 Bulk_RNA-seq_analysis 폴더를 만들고, 해당 폴더 안에 아래 그림과 같이 배치해 놓는다.



- 그리고 나서, anaconda를 설치 및 bulk-RNA sequencing 용 가상환경을 만든다.

(anaconda를 설치할 때 home directory에서 설치할 것!)

- 가상환경을 만든 후에는 분석 관련 package를 설치한다. (conda를 이용해 설치한다.)

```
(base) wjddns037@DESKTOP-39ALU06:~$ conda create -n RNA_seq
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /home/wjddns037/anaconda3/envs/RNA_seq



Proceed ([y]/n)? y
```

Workflow step	packages
Quality Check	Fastqc 0.11.9
Trimming	Trimmomatic 0.39, Sickle-trim 1.33
Alignment	Hisat2 2.2.1, Samtools 1.14
Quantification	Subread 2.0.1
DEG analysis	EdgeR (R version ≥ 4.0)

1. MultiQC 또는 fastqc 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v1.6
osx-64 v1.6
noarch v1.11

To install this package with conda run one of the following:

```
conda install -c bioconda multiqc
conda install -c bioconda/label/cf201901 multiqc
```

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v0.11.8
osx-64 v0.11.8
noarch v0.11.9

To install this package with conda run one of the following:

```
conda install -c bioconda fastqc
conda install -c bioconda/label/broken fastqc
conda install -c bioconda/label/cf201901 fastqc
```

2. Sickle 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v1.33
osx-64 v1.33

To install this package with conda run one of the following:

```
conda install -c bioconda sickle-trim
conda install -c bioconda/label/cf201901 sickle-trim
```

3. Hisat2 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v2.2.1
osx-64 v2.2.1

To install this package with conda run one of the following:

```
conda install -c bioconda hisat2
conda install -c bioconda/label/cf201901 hisat2
```

4. Samtools 및 IGV 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v1.14
osx-64 v1.14

To install this package with conda run one of the following:

```
conda install -c bioconda samtools
conda install -c bioconda/label/cf201901 samtools
```

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v2.4.9
osx-64 v2.4.9
noarch v2.11.3

To install this package with conda run one of the following:

```
conda install -c bioconda igv
conda install -c bioconda/label/cf201901 igv
```

5. Subread 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v2.0.1
osx-64 v2.0.1

To install this package with conda run one of the following:

```
conda install -c bioconda subread
conda install -c bioconda/label/cf201901 subread
```

6. R 4.1.1 설치

Installers

Info: This package contains files in non-standard labels.

conda install

linux-ppc64le v4.1.1
osx-arm64 v4.1.1
linux-64 v4.1.1
win-32 v3.4.1
linux-aarch64 v4.1.1
osx-64 v4.1.1
win-64 v4.1.1

To install this package with conda run one of the following:

conda install -c conda-forge r-base=4.1.1

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v3.36.0
osx-64 v3.36.0

To install this package with conda run one of the following:

```
conda install -c bioconda bioconductor-edger
conda install -c bioconda/label/gcc7 bioconductor-edger
conda install -c bioconda/label/cf201901 bioconductor-edger
```

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64 v1.34.0
osx-64 v1.34.0

To install this package with conda run one of the following:

```
conda install -c bioconda bioconductor-deseq2
conda install -c bioconda/label/gcc7 bioconductor-deseq2
conda install -c bioconda/label/broken bioconductor-deseq2
conda install -c bioconda/label/cf201901 bioconductor-deseq2
```

Fastqc.sh

```
                    Description

This script is used for quality control check.
Before using this script, you need to install Fastqc package (conda install -c bioconda fastqc)

Usage: Bash script for executing quality check. --> ./Fastqc.sh
       Input directory where all the neccesary files are saved. --> ./1.Rawdata (SRR391535.fastq.gz or SRR391535_1.fastq.gz / SRR391535_2.fastq.gz)
       Output directory where all the results goes. --> ./1.Rawdata/QC_result

Executable code --> ./Fastqc.sh ./1.Rawdata ./1.Rawdata/QC_result
```

→ Bash Fastqc.sh 명령어 입력할 시에 Description이 나온다.

```
(RNA_analysis) wjddns037@DESKTOP-39ALU06:~/Lab_wsl/PJW_Study/Bioinformatics/Bulk_RNA-seq_analysis$ ./Fastqc.sh ./1.Rawdata
```

- **Executable code를 입력해 실행시키면, 자동적으로 raw data들을 quality check가 진행된다.**
- **해당 script를 실행하기 전에, 1.Rawdata 폴더 안에 QC_result 폴더를 생성했는지 확인해야 된다.**

Trimming.sh

```
                  Description

This script is used for trimming the nucleo acid sequence.
Before using this script, you need to install Trimming tools (conda install -c bioconda sickle-trim or conda install -c bioconda trimmomatic)

Usage: Bash script for executing trimming the nucleo acid sequence. --> ./Trimming.sh
       Input directory where all the neccesary files are saved. --> ./1.Rawdata (SRR391535.fastq.gz or SRR391535_1.fastq.gz / SRR391535_2.fastq.gz)

       (Note) Output directory is same as input directory!

Executable code --> ./Trimming.sh ./1.Rawdata/
```

→ Bash Trimming.sh 명령어 입력할 시에 Description이 나온다.

```
(RNA_analysis) wjddns037@DESKTOP-39ALU06:~/Lab_wsl/PJW_Study/Bioinformatics/Bulk_RNA-seq_analysis$ ./Trimming.sh ./1.Rawdata
Enter your enviroment name. > RNA_analysis
Is paired-end? (Yes / No) > Yes
What do you use tool? (Sickle / Trimmomatic) > Sickle
```

- **Executable code를 입력해 실행시키면, 위와 같은 질문이 보이게 된다.**
- **Environment name은 anaconda 가상환경 이름을 입력하면 되고, 나머지는 질문에 맞게 입력하면 된다.**

  (ex) bulk-RNA sequencing 용 가상환경 이름이 RNA-seq이면, RNA-seq이라고 입력하면 된다.

Hisat2.sh

```
                        Description

This script is used for genome alignment.
Before using this script, you need to install Hisat2 and Samtools package.

Usage: Bash script for executing genome alignment --> ./HISAT2.sh
       Input directory where all the neccesary files are saved. --> ./1.Rawdata (SRR391535.fastq.gz or SRR391535_1.fastq.gz / SRR391535_2.fastq.gz)
       Output directory where all the results goes. --> ./2.HISAT
       Reference directory where all the neccesary files are saved. --> ./0.Reference (Homo_sapiens.GRCh38.dna.primary_assembly.fa, Homo_sapiens.GRCh38.87.gtf)

Executable code --> ./HISAT2.sh ./1.Rawdata ./2.HISAT
```

→ Bash HISAT2.sh 명령어 입력할 시에 Description이 나온다.

```
(RNA_analysis) wjddns037@DESKTOP-39ALUO6:~/Lab_wsl/PJW_Study/Bioinformatics/Bulk_RNA-seq_analysis$ ./HISAT2.sh ./1.Rawdata ./2.HISAT
Do you want to build index? (Yes / No) No
Is the data trimmed? (Yes / No) > Yes
Is paired-end? (Yes / No) Yes
Without indexing build, align genome right away!
What did you use trimming tool? (Sickle / Trimmomatic) > Sickle
```

- **Executable code를 입력해 실행시키면, 위와 같은 질문이 보이게 된다.**
- **Alignment를 하기 전에 reference genome에 대한 index를 생성해야 되며, 첫 번째 질문에서 Yes를 입력하면 index를 생성한 후에 alignment를 진행하게 된다. (만약 index가 미리 생성되어 있을 경우에는, No을 입력하면 index 생성을 생략하고 alignment를 진행하게 된다.)**
- **나머지는 질문에 맞게 입력하면 된다.**

FeatureCounts.sh

```
                    Description

This script is used for gene counting by using featurecounts.
Before using this script, you need to install subread package. (conda install -c bioconda subread)

Usage: Bash script for executing gene counting. --> ./FeatureCounts.sh
       Input directory where all the neccesary files are saved. --> ./2.HISAT (SRR391535.bam)
       Output directory where all the results goes. --> ./3.FeatureCounts
       Reference directory where all the neccesary files are saved. --> ./0.Reference (Homo_sapiens.GRCh38.dna.primary_assembly.fa, Homo_sapiens.GRCh38.87.gtf)

Precautions: Separately store paired-end bam files and single-end bam files in the input directory.

Executable code --> ./FeatureCounts.sh ./2.HISAT ./3.FeatureCounts
```

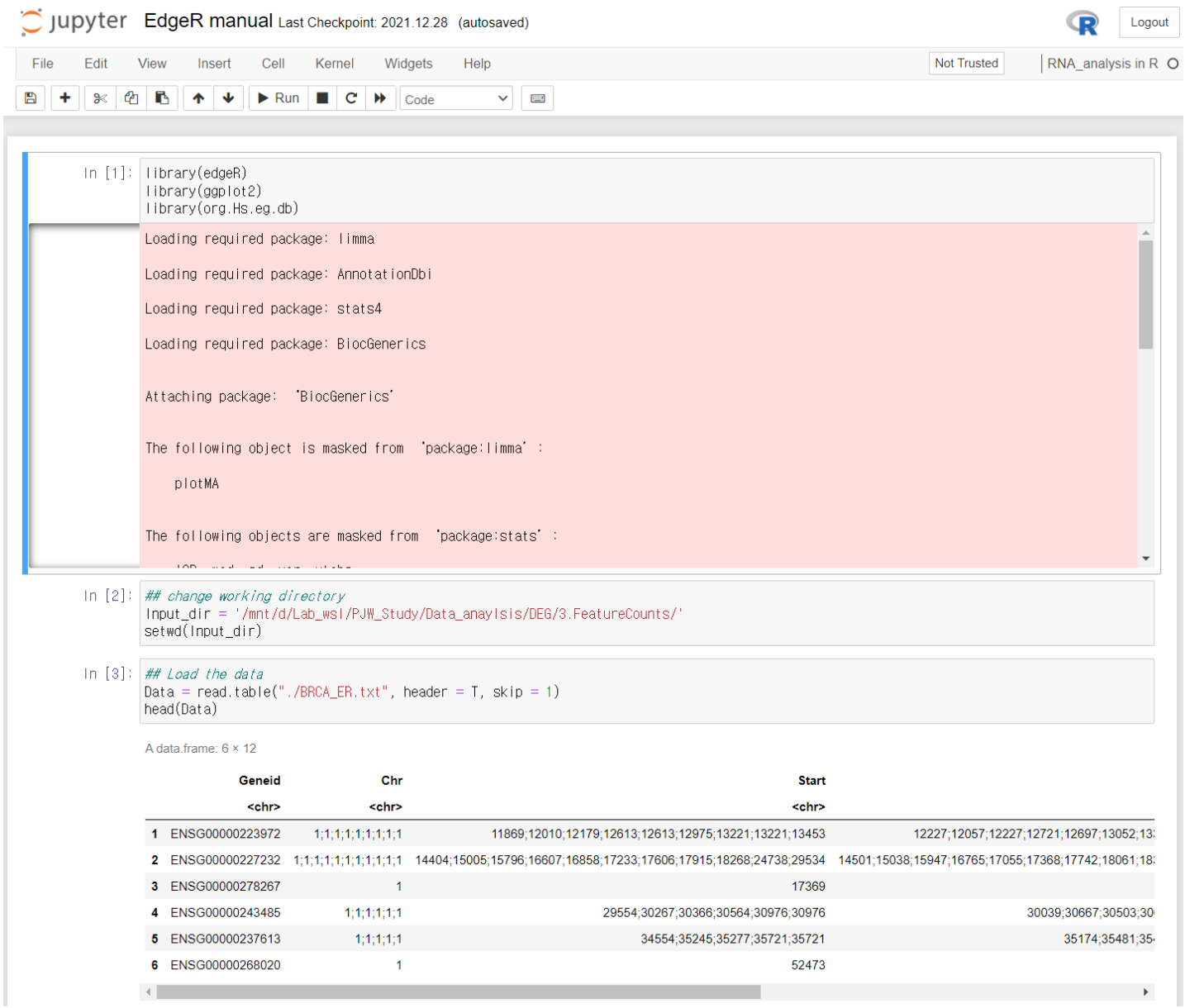
→ Bash FeatureCounts.sh 명령어 입력할 시에 Description이 나온다.

```
(RNA_analysis) wjddns037@DESKTOP-39ALUO6:~/Lab_wsl/PJW_Study/Bioinformatics/Bulk_RNA-seq_analysis$ ./FeatureCounts.sh ./2.HISAT ./3.FeatureCounts
Is paired-end? (Yes / No) > Yes
```

- **Executable code를 입력해 실행시키면, 위와 같은 질문이 등장하게 되며 이에 맞게 입력하면 된다.**
- **해당 scrip를 실행하기 전에 주의사항은 Paired-end와 Single-end을 따로 저장 해야 된다.**

(ex) Paired-end --> ./3.FeatureCounts_PE, Single-end --> ./3.FeatureCounts_SE

EdgeR manual.ipynb



- 해당 파일에는 FeatureCount에서 얻은 countmatrix 파일을 읽어서 Differentially gene expression 분석 과정이 적혀 있다.
- 본인이 가지고 있는 데이터에 맞춰 수정을 해서 사용할 것.
- 또한 분석 하기 전에 countmatrix 파일을 한번 확인해본 뒤, 진행하는 것을 추천!

EdgeR manual.ipynb

In [12]:
```
# Perform quasi-likehoold F-test.
fit = glmQLFit(y, Design)
qlt = glmQLFTest(fit, coef = 2)
topTags(qlt)
```

$table

A data.frame: 10 × 6

	Symbol	logFC	logCPM	F	PValue	FDR
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
48811	NA	10.002453	6.621865	12403.444	4.851566e-25	9.060784e-21
28220	NA	3.355587	7.649512	10874.484	1.424209e-24	1.184355e-20
21228	NA	3.622280	9.219858	10241.759	2.326261e-24	1.184355e-20
3971	NA	6.251815	6.452778	10133.991	2.536635e-24	1.184355e-20
52914	NA	3.154560	8.271813	8932.228	7.126634e-24	2.610991e-20
12346	NA	3.157542	7.432057	8628.958	9.454367e-24	2.610991e-20
21862	NA	3.555547	6.846893	8592.641	9.786324e-24	2.610991e-20
54521	WDR44	3.106577	7.684615	8153.931	1.502561e-23	3.507728e-20
39883	NA	3.283070	6.701488	7906.512	1.933357e-23	3.728342e-20
53347	UBASH3A	3.454711	7.337028	7875.596	1.996328e-23	3.728342e-20

$adjust.method
'BH'
$comparison
'groupCancer'
$test
'glm'

In [14]:
```
# Gene ontology and pathway analysis
go = goana(qlt, species = "Hs")
topGO(go, sort = "up")
```

A data.frame: 20 × 7

	Term	Ont	N	Up	Down	P.Up	P.Down
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0061351	neural precursor cell proliferation	BP	42	29	9	0.001117980	0.9954468
GO:1903047	mitotic cell cycle process	BP	190	105	62	0.001584623	0.9738216
GO:0009636	response to toxic substance	BP	75	46	19	0.002369279	0.9958661
GO:0001824	blastocyst development	BP	16	13	2	0.003052973	0.9959522
GO:0042743	hydrogen peroxide metabolic process	BP	21	16	2	0.003236271	0.9995667
GO:0062197	cellular response to chemical stress	BP	92	54	26	0.004099931	0.9889647
GO:0034451	centriolar satellite	CC	23	17	3	0.004118247	0.9986669
GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	MF	23	17	5	0.004118247	0.9766307
GO:0004601	peroxidase activity	MF	23	17	5	0.004118247	0.9766307
GO:2000177	regulation of neural precursor cell proliferation	BP	23	17	4	0.004118247	0.9934423
GO:1902882	regulation of response to oxidative stress	BP	23	17	3	0.004118247	0.9986669
GO:0070509	calcium ion import	BP	30	21	5	0.004224303	0.9979601
GO:0016072	rRNA metabolic process	BP	41	27	9	0.004704985	0.9938550
GO:0006364	rRNA processing	BP	41	27	9	0.004704985	0.9938550
GO:0031253	cell projection membrane	CC	87	51	32	0.005394931	0.7052970
GO:0046872	metal ion binding	MF	925	447	347	0.005427957	0.8642113
GO:0030501	positive regulation of bone mineralization	BP	15	12	1	0.005659181	0.9994112
GO:0030879	mammary gland development	BP	38	25	9	0.006587809	0.9853560
GO:0005874	microtubule	CC	84	49	25	0.007199904	0.9710792
GO:0042744	hydrogen peroxide catabolic process	BP	17	13	2	0.007685976	0.9973966