University of
TEHRAN

Introduction to Bioinformatics

RNA-Seq analysis

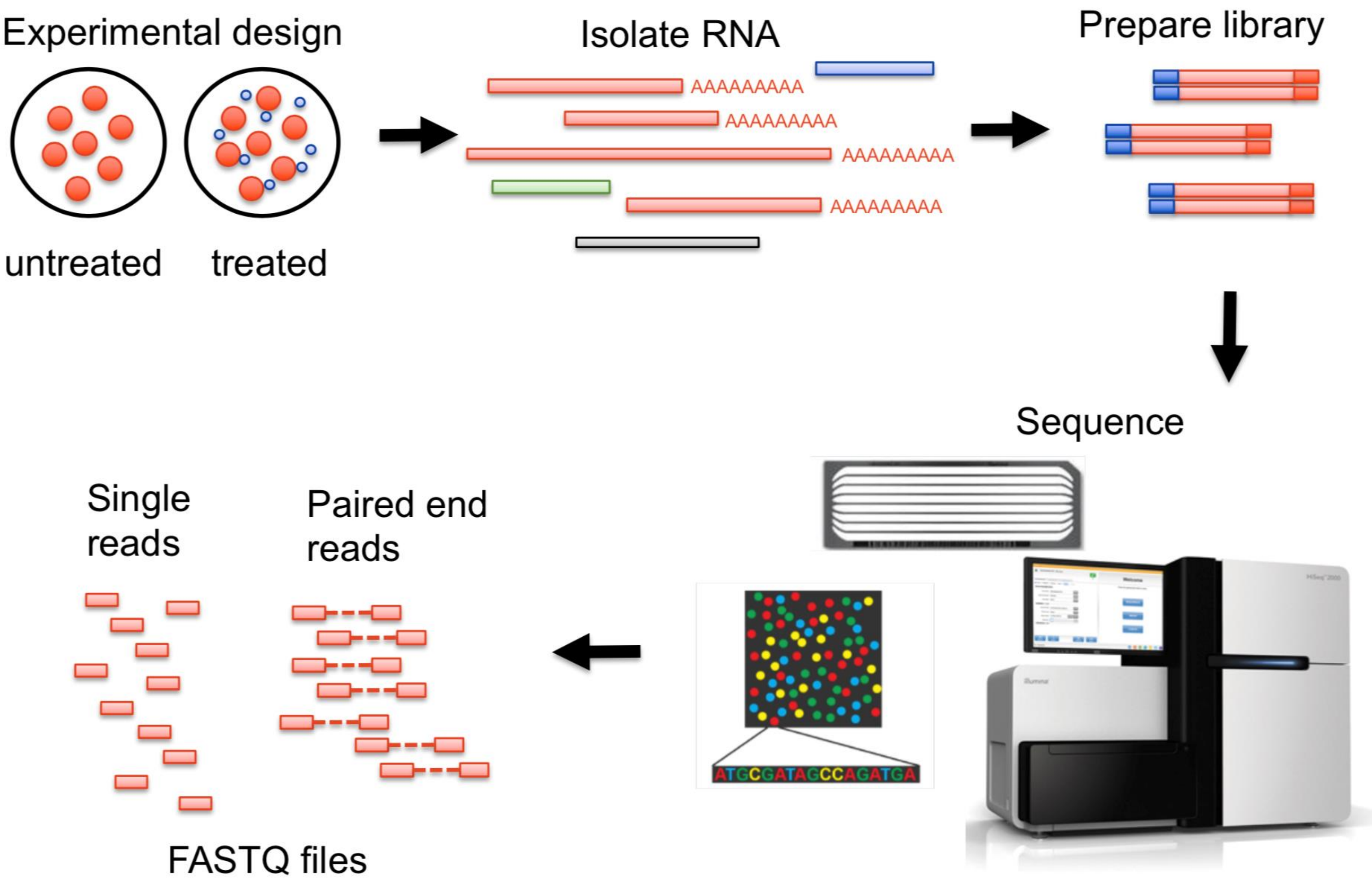
Department of Bioinformatics, IBB, University of Tehran

Winter 2024

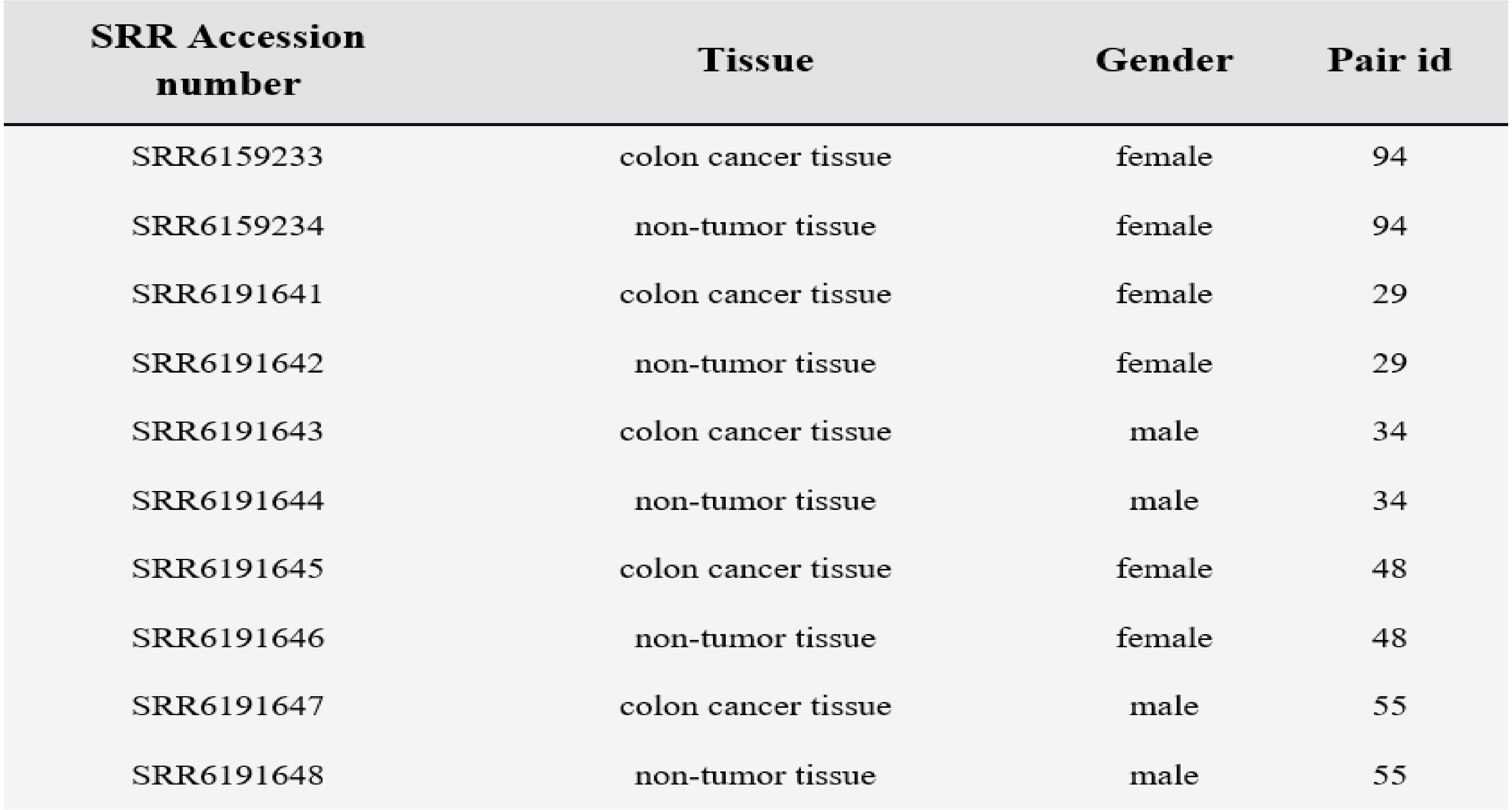
Presenter: Fereshteh Noroozi

RNA-Seq analysis

- Sample Preparation
- Library Preparation
- Sequencing
- Data Preprocessing
- Quantification of Gene Expression
- Differential Expression Analysis
- Functional Analysis
- Visualization



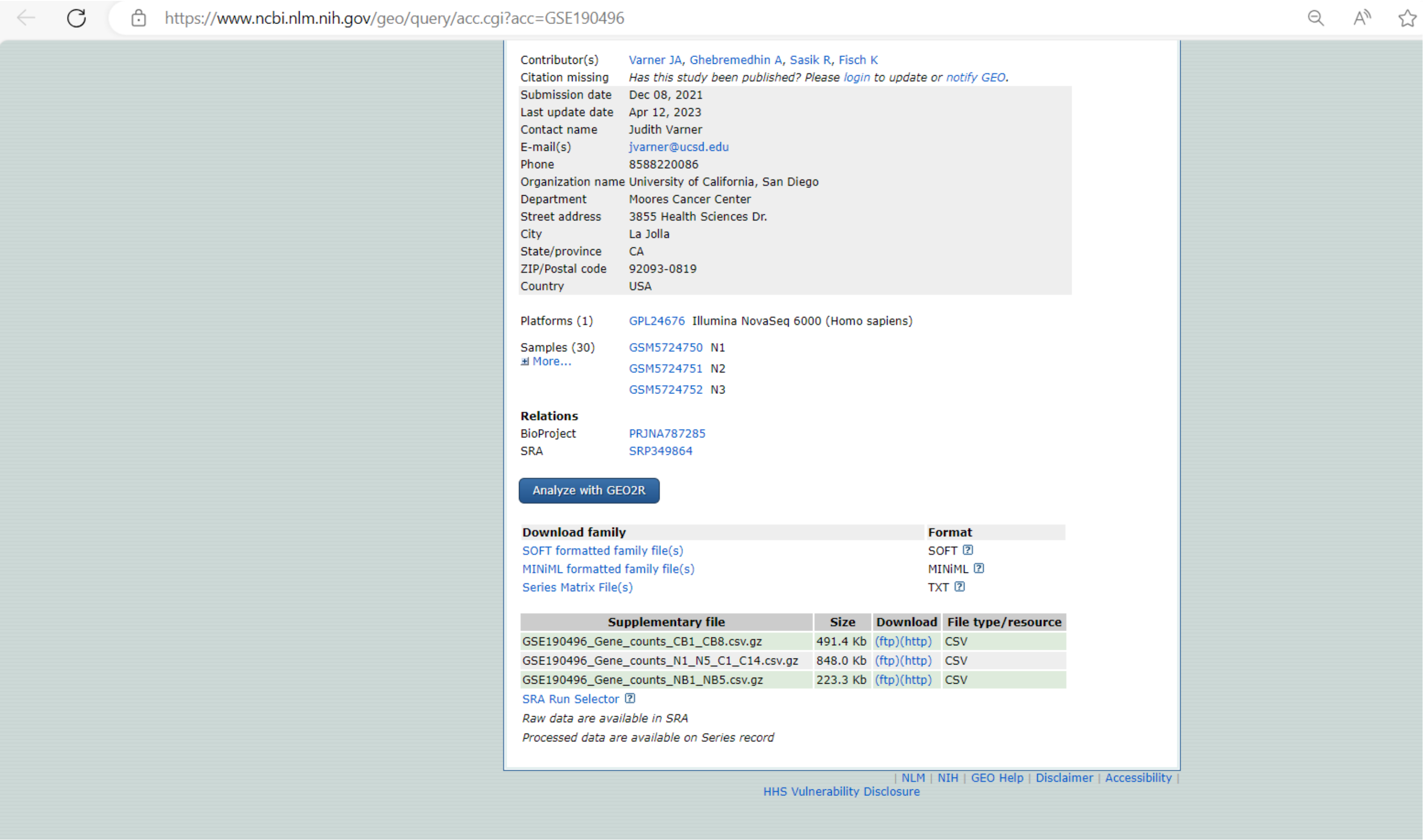
Accession number GSE104836



SRR Accession number	Tissue	Gender	Pair id
SRR6159233	colon cancer tissue	female	94
SRR6159234	non-tumor tissue	female	94
SRR6191641	colon cancer tissue	female	29
SRR6191642	non-tumor tissue	female	29
SRR6191643	colon cancer tissue	male	34
SRR6191644	non-tumor tissue	male	34
SRR6191645	colon cancer tissue	female	48
SRR6191646	non-tumor tissue	female	48
SRR6191647	colon cancer tissue	male	55
SRR6191648	non-tumor tissue	male	55

Accession number GSE104836

- SRA Run Selector



https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190496

Contributor(s)	Varner JA, Ghebremedhin A, Sasik R, Fisch K
Citation missing	*Has this study been published? Please login to update or notify GEO.*
Submission date	Dec 08, 2021
Last update date	Apr 12, 2023
Contact name	Judith Varner
E-mail(s)	jvarner@ucsd.edu
Phone	8588220086
Organization name	University of California, San Diego
Department	Moores Cancer Center
Street address	3855 Health Sciences Dr.
City	La Jolla
State/province	CA
ZIP/Postal code	92093-0819
Country	USA
Platforms (1)	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)
Samples (30) ⊞ More...	GSM5724750 N1 GSM5724751 N2 GSM5724752 N3

Relations

BioProject	PRJNA787285
SRA	SRP349864

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE190496_Gene_counts_CB1_CB8.csv.gz	491.4 Kb	(ftp)(http)	CSV
GSE190496_Gene_counts_N1_N5_C1_C14.csv.gz	848.0 Kb	(ftp)(http)	CSV
GSE190496_Gene_counts_NB1_NB5.csv.gz	223.3 Kb	(ftp)(http)	CSV

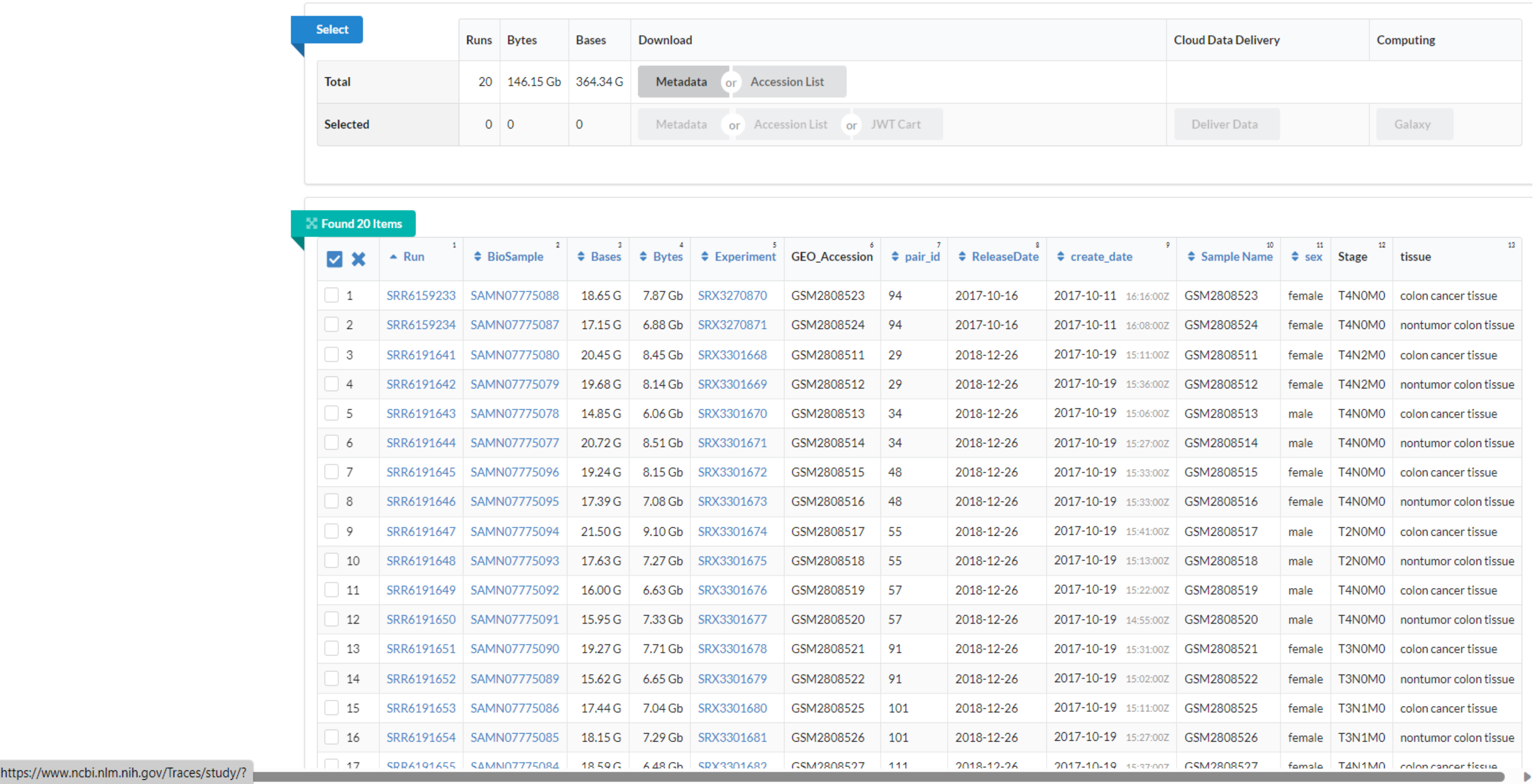
SRA Run Selector ?

Raw data are available in SRA

Processed data are available on Series record

| NLM | NIH | GEO Help | Disclaimer | Accessibility |
HHS Vulnerability Disclosure

Accession number GSE104836

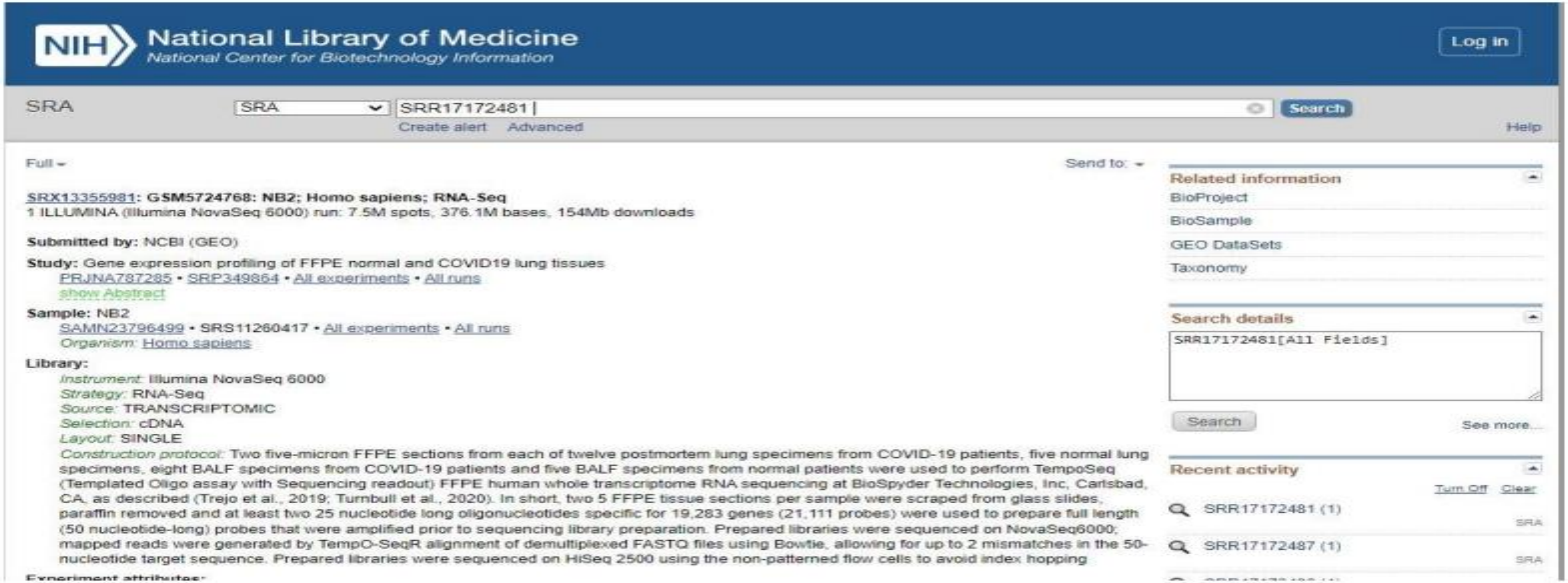


	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	20	146.15 Gb	364.34 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

☑ ✖	Run	BioSample	Bases	Bytes	Experiment	GEO_Accession	pair_id	ReleaseDate	create_date	Sample Name	sex	Stage	tissue
□ 1	SRR6159233	SAMN07775088	18.65 G	7.87 Gb	SRX3270870	GSM2808523	94	2017-10-16	2017-10-11 16:16:00Z	GSM2808523	female	T4N0M0	colon cancer tissue
□ 2	SRR6159234	SAMN07775087	17.15 G	6.88 Gb	SRX3270871	GSM2808524	94	2017-10-16	2017-10-11 16:08:00Z	GSM2808524	female	T4N0M0	nontumor colon tissue
□ 3	SRR6191641	SAMN07775080	20.45 G	8.45 Gb	SRX3301668	GSM2808511	29	2018-12-26	2017-10-19 15:11:00Z	GSM2808511	female	T4N2M0	colon cancer tissue
□ 4	SRR6191642	SAMN07775079	19.68 G	8.14 Gb	SRX3301669	GSM2808512	29	2018-12-26	2017-10-19 15:36:00Z	GSM2808512	female	T4N2M0	nontumor colon tissue
□ 5	SRR6191643	SAMN07775078	14.85 G	6.06 Gb	SRX3301670	GSM2808513	34	2018-12-26	2017-10-19 15:06:00Z	GSM2808513	male	T4N0M0	colon cancer tissue
□ 6	SRR6191644	SAMN07775077	20.72 G	8.51 Gb	SRX3301671	GSM2808514	34	2018-12-26	2017-10-19 15:27:00Z	GSM2808514	male	T4N0M0	nontumor colon tissue
□ 7	SRR6191645	SAMN07775096	19.24 G	8.15 Gb	SRX3301672	GSM2808515	48	2018-12-26	2017-10-19 15:33:00Z	GSM2808515	female	T4N0M0	colon cancer tissue
□ 8	SRR6191646	SAMN07775095	17.39 G	7.08 Gb	SRX3301673	GSM2808516	48	2018-12-26	2017-10-19 15:33:00Z	GSM2808516	female	T4N0M0	nontumor colon tissue
□ 9	SRR6191647	SAMN07775094	21.50 G	9.10 Gb	SRX3301674	GSM2808517	55	2018-12-26	2017-10-19 15:41:00Z	GSM2808517	male	T2N0M0	colon cancer tissue
□ 10	SRR6191648	SAMN07775093	17.63 G	7.27 Gb	SRX3301675	GSM2808518	55	2018-12-26	2017-10-19 15:13:00Z	GSM2808518	male	T2N0M0	nontumor colon tissue
□ 11	SRR6191649	SAMN07775092	16.00 G	6.63 Gb	SRX3301676	GSM2808519	57	2018-12-26	2017-10-19 15:22:00Z	GSM2808519	male	T4N0M0	colon cancer tissue
□ 12	SRR6191650	SAMN07775091	15.95 G	7.33 Gb	SRX3301677	GSM2808520	57	2018-12-26	2017-10-19 14:55:00Z	GSM2808520	male	T4N0M0	nontumor colon tissue
□ 13	SRR6191651	SAMN07775090	19.27 G	7.71 Gb	SRX3301678	GSM2808521	91	2018-12-26	2017-10-19 15:31:00Z	GSM2808521	female	T3N0M0	colon cancer tissue
□ 14	SRR6191652	SAMN07775089	15.62 G	6.65 Gb	SRX3301679	GSM2808522	91	2018-12-26	2017-10-19 15:02:00Z	GSM2808522	female	T3N0M0	nontumor colon tissue
□ 15	SRR6191653	SAMN07775086	17.44 G	7.04 Gb	SRX3301680	GSM2808525	101	2018-12-26	2017-10-19 15:11:00Z	GSM2808525	female	T3N1M0	colon cancer tissue
□ 16	SRR6191654	SAMN07775085	18.15 G	7.29 Gb	SRX3301681	GSM2808526	101	2018-12-26	2017-10-19 15:27:00Z	GSM2808526	female	T3N1M0	nontumor colon tissue
□ 17	SRR6191655	SAMN07775084	18.59 G	6.48 Gb	SRX3301682	GSM2808527	111	2018-12-26	2017-10-19 15:37:00Z	GSM2808527	female	T4N1M0	colon cancer tissue

https://www.ncbi.nlm.nih.gov/Traces/study/?

Data preparation



SRX13355981: GSM5724768: NB2; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina NovaSeq 6000) run: 7.5M spots, 376.1M bases, 154Mb downloads

Submitted by: NCBI (GEO)

Study: Gene expression profiling of FFPE normal and COVID19 lung tissues
PRJNA787285 • SRP349864 • All experiments • All runs
show Abstract

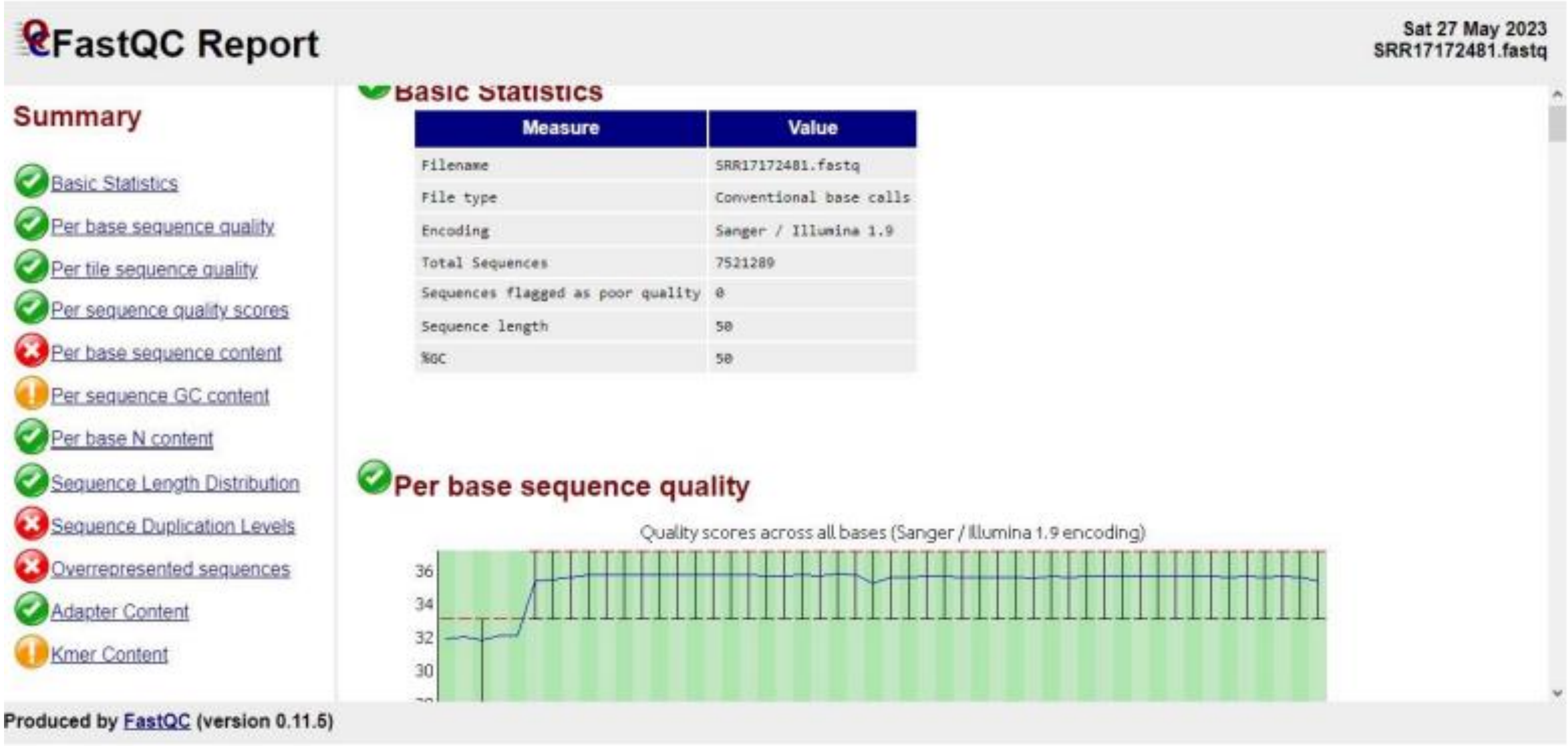
Sample: NB2
SAMN23796499 • SRS11260417 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina NovaSeq 6000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: Two five-micron FFPE sections from each of twelve postmortem lung specimens from COVID-19 patients, five normal lung specimens, eight BALF specimens from COVID-19 patients and five BALF specimens from normal patients were used to perform TempoSeq (Templated Oligo assay with Sequencing readout) FFPE human whole transcriptome RNA sequencing at BioSpyder Technologies, Inc, Carlsbad, CA, as described (Trejo et al., 2019; Turnbull et al., 2020). In short, two 5 FFPE tissue sections per sample were scraped from glass slides, paraffin removed and at least two 25 nucleotide long oligonucleotides specific for 19,283 genes (21,111 probes) were used to prepare full length (50 nucleotide-long) probes that were amplified prior to sequencing library preparation. Prepared libraries were sequenced on NovaSeq6000; mapped reads were generated by TempO-SeqR alignment of demultiplexed FASTQ files using Bowtie, allowing for up to 2 mismatches in the 50-nucleotide target sequence. Prepared libraries were sequenced on HiSeq 2500 using the non-patterned flow cells to avoid index hopping

- For paired : SRA(Zip of fastq) files into two forward and reverse Fastq.gz files using the *fastq-dump* command from the SRA Toolkit. (Hint: use *fastq-dump [options] file.sra*)

Part a- Quality control and trimming

use fastqc and TrimmomaticPE



FastQC Report

Sat 27 May 2023
SRR17172481.fastq

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

Measure	Value
Filename	SRR17172481.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7521289
Sequences flagged as poor quality	0
Sequence length	50
%GC	50

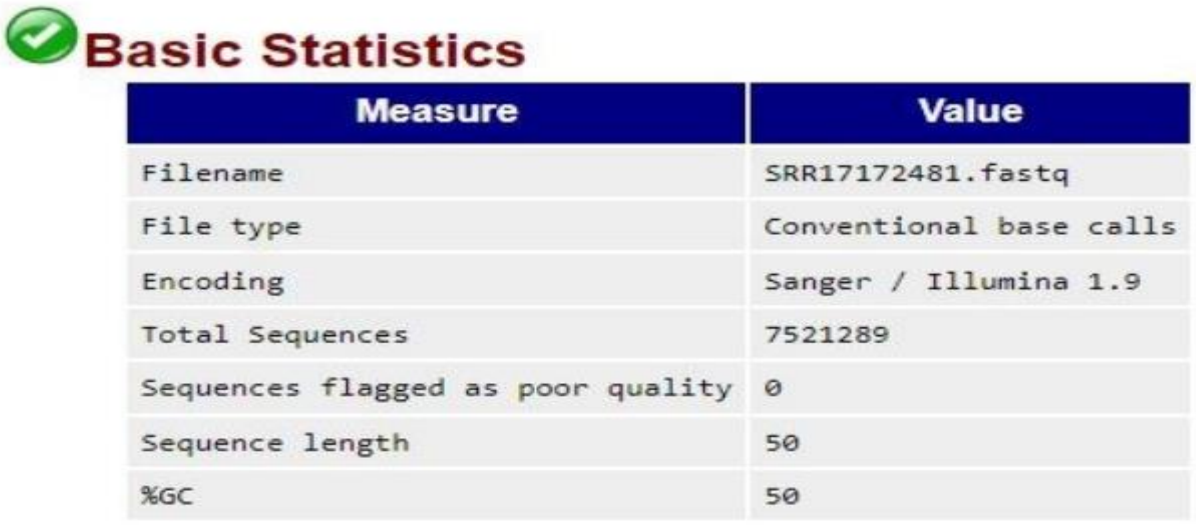
✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30

Produced by FastQC (version 0.11.5)

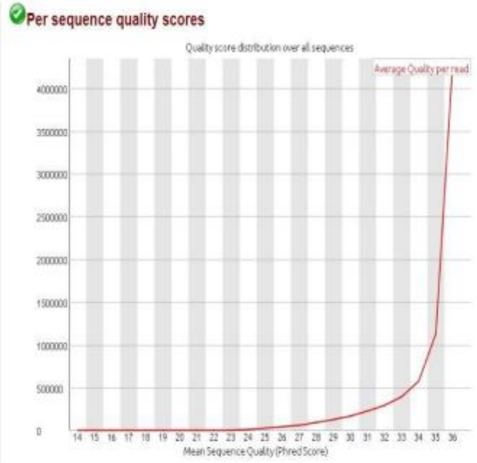
Part a- Quality control and trimming

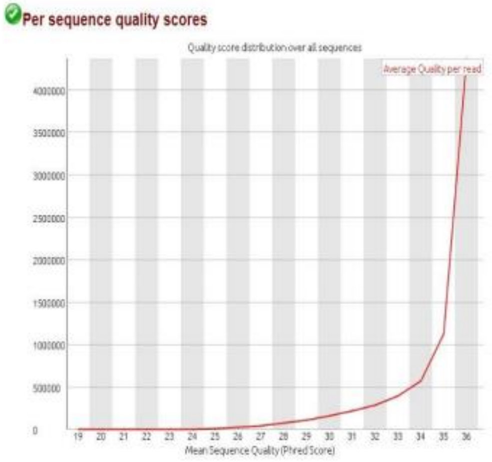
1-What is the average number of reads across samples before and after the read trimming?
2-Compare the read length averages in different samples before and after the read trimming?
3-Compare the read quality distributions over all sequences before and after the read trimming.
4-What does the Adaptor Content warning indicate?
5-Why do we first remove the Adapter sequences for the reads and then the low-quality bases?
6-What does the quality of bases mean, and how is it obtained?



Basic Statistics

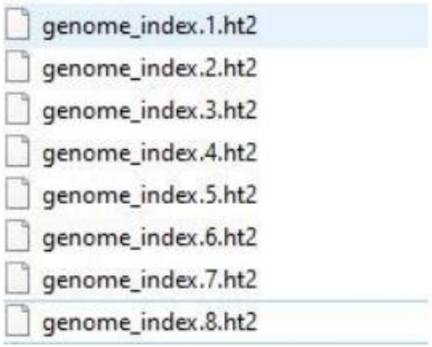
Measure	Value
Filename	SRR17172481.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7521289
Sequences flagged as poor quality	0
Sequence length	50
%GC	50





Part b- Read mapping

- Using the *HISAT2* software to map reads to the reference



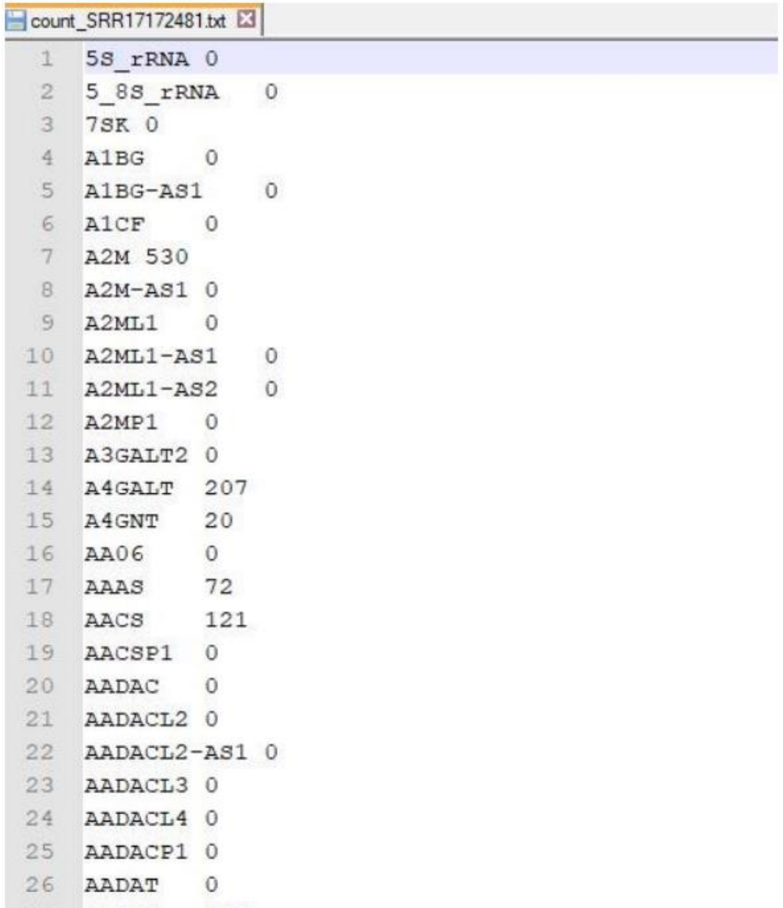
genome_index.1.ht2
genome_index.2.ht2
genome_index.3.ht2
genome_index.4.ht2
genome_index.5.ht2
genome_index.6.ht2
genome_index.7.ht2
genome_index.8.ht2

1. What is the difference between SAM and BAM files?
2. What is the purpose of indexing the genome?
3. Report mapping percentages of all samples in a table. Please explain why a low percentage of reads cannot be mapped.

Part c- Building gene expression matrix

- Run htseq-count then, to merge results files into a single matrix
- How many genes are not expressed in control and tumor samples? Explain the results.
- Compare the matrix obtained at this stage with the corresponding gene expression submatrix of the main study. Discuss the differences.
- What are other software available to do this step?

Name two other software and discuss their advantages and disadvantages.



```
5S_rRNA 0
5_8S_rRNA	0
7SK 0
A1BG	0
A1BG-AS1	0
A1CF	0
A2M 530
A2M-AS1 0
A2ML1	0
A2ML1-AS1	0
A2ML1-AS2	0
A2MP1	0
A3GALT2 0
A4GALT	207
A4GNT	20
AA06	0
AAAS	72
AACS	121
AACSP1	0
AADAC	0
AADACL2 0
AADACL2-AS1 0
AADACL3 0
AADACL4 0
AADACP1 0
AADAT	0
```

Part d- Differential gene expression analysis

- **Use the expression matrix of the main study (all samples) to answer the following question**
  Use edgeR
- How many genes are given to edgeR? How many of them are differentially expressed in tumor versus normal samples? How do you define statistical significance in this context?
- Determine the percentage of differentially expressed genes with |log2FoldChange| > 1.5.
- Explain the difference between P-value and FDR?

Part e- Gene Ontology enrichment analysis

- GOseq package in R
  To accomplish this step, select the genes with FDR < 0.1 and an absolute value of Log2FoldChange > 1.5.
- Display results related to Biological Process, Molecular Function, Cellular Component, and KEGG as separate plots using an R package of your choice.
- Do a brief study of each of the significant terms and discuss which terms you think may play an important role.
- Write a general biological conclusion about the final results of the project.