Assignmment-2

Q.1: How can I calculate the length of each gene from the ASM584v2 GTF file?

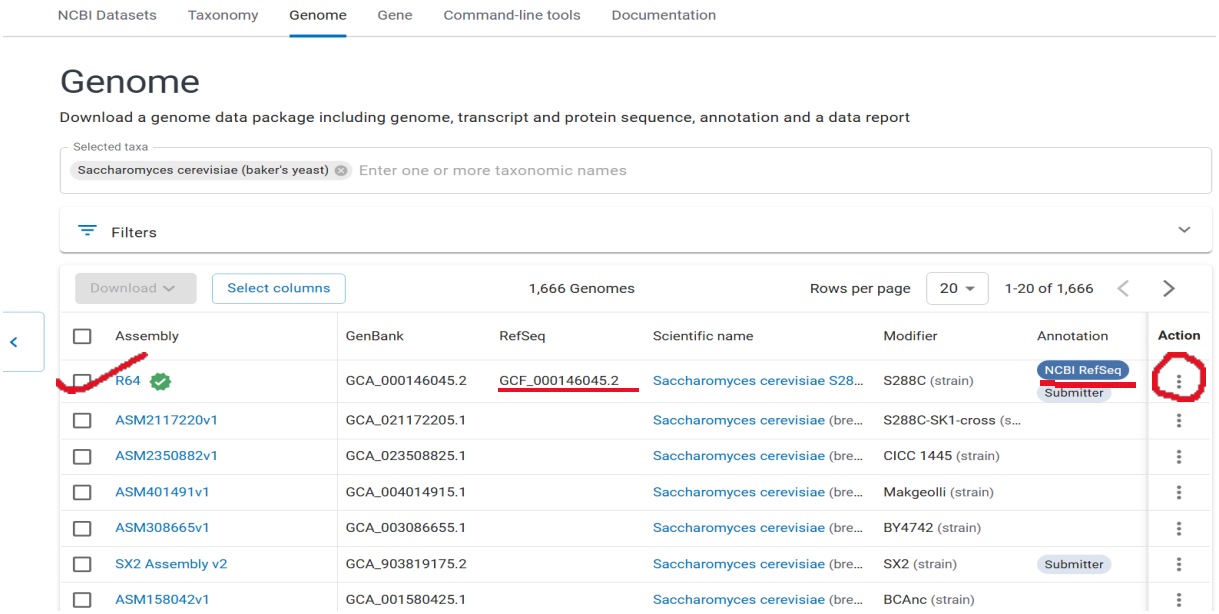
- Visit NCBI homepage: National Center for Biotechnology Information
- Choose ‘Genome’ from the dropdown menu.
- Type the scientific name of the organism you want to search for (e.g.: *Escherichia coli*). Click on ‘search’ button.
- Now, you will see a list of assemblies. Select the top one, i.e., ASM584v2 to download the GTF file.
- Run the below following -

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS_Lab/ngs_files$ awk '$3 == "gene" {len = $5 - $4 + 1; print $1, len}' GCF_000005845.2_ASM584v2_genomic.gtf > gene_lengths.txt
```

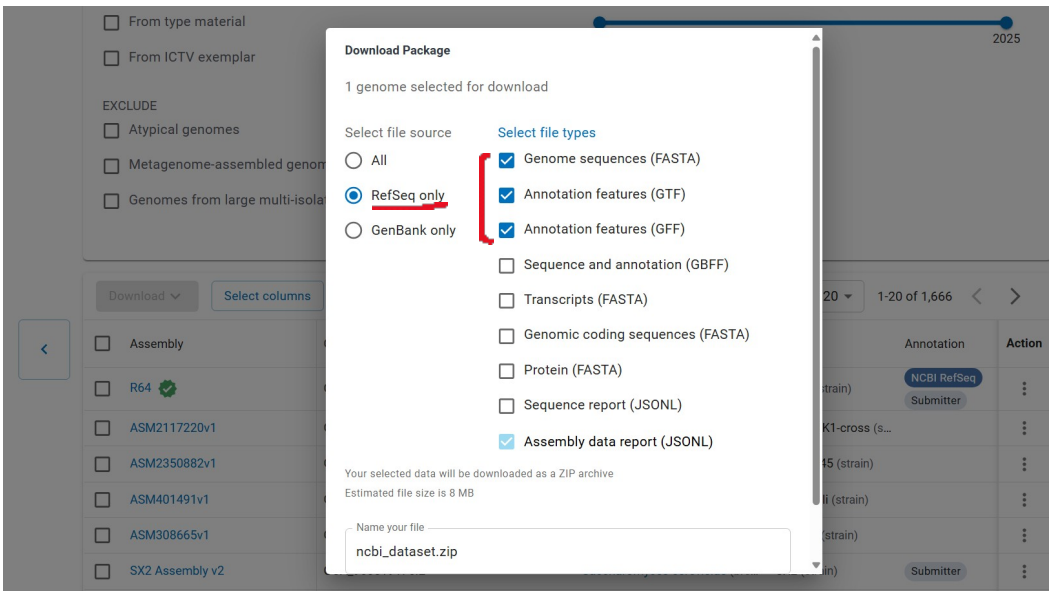
```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS_Lab/ngs_files$ cat gene_lengths.txt
NC_000913.3 66
NC_000913.3 2463
NC_000913.3 933
NC_000913.3 1287
NC_000913.3 297
NC_000913.3 777
NC_000913.3 1431
NC_000913.3 954
NC_000913.3 588
NC_000913.3 567
NC_000913.3 714
NC_000913.3 486
NC_000913.3 405
NC_000913.3 1917
NC_000913.3 1131
NC_000913.3 1113
NC_000913.3 210
NC_000913.3 153
NC_000913.3 59
NC_000913.3 1167
NC_000913.3 906
NC_000913.3 504
NC_000913.3 276
NC_000913.3 264
NC_000913.3 219
NC_000913.3 942
NC_000913.3 2817
NC_000913.3 495
NC_000913.3 450
NC_000913.3 951
NC_000913.3 915
```

Q.2: Create a directory and download the reference genome of *Saccharomyces cerevisiae* along with annotation files in GTF and GFF formats.

- Create a directory using ‘mkdir’ command.
- Visit NCBI homepage: National Center for Biotechnology Information
- Choose ‘Genome’ from the dropdown menu.
- Type the scientific name of the organism you want to search for (e.g.: *Saccharomyces cerevisiae*). Click on ‘search’ button.
- Now, you will see a list of assemblies. Select the top one (hint: green tick) as shown below.

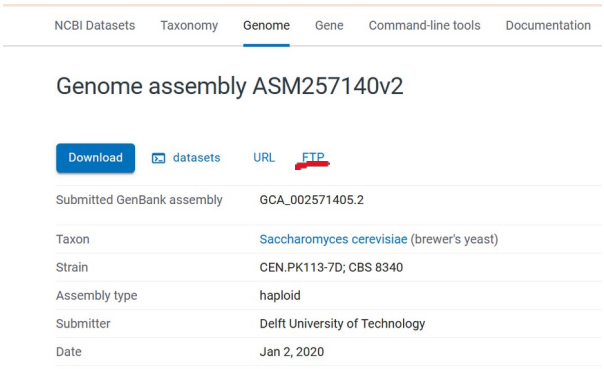


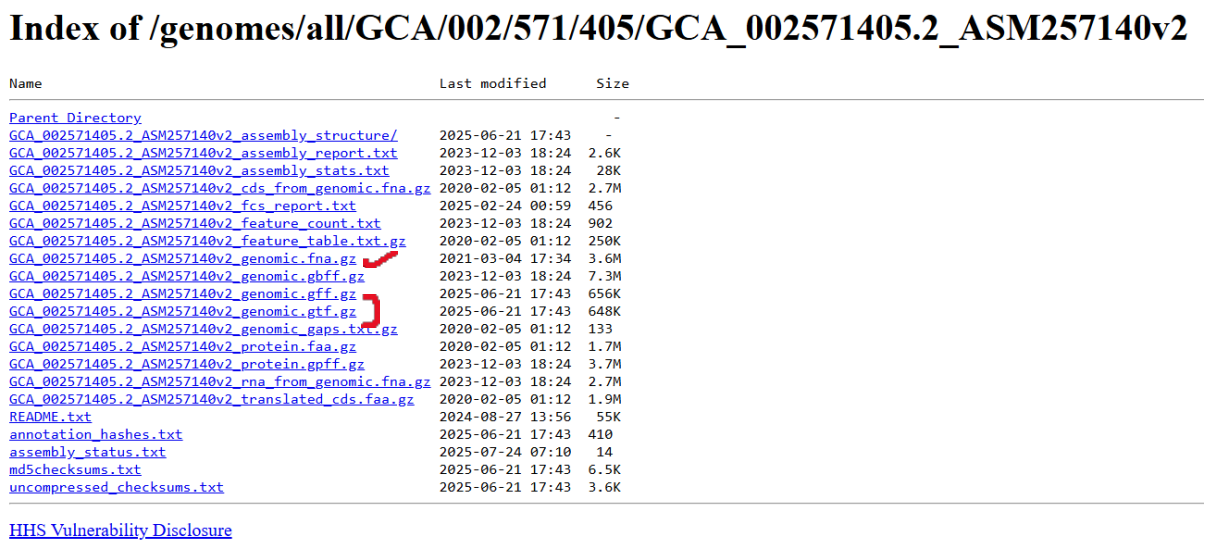
- Go to ‘Action’ and then click on ‘Download’. Select ‘RefSeq only’ and choose those that you want to download.



- Unzip the file using ‘gunzip’ command and navigate into the sub-folders in it to see and analyse the .fna / .gtf / .gff files.

- Another way to download is through ‘FTP’ links or even using CLI.





Index of /genomes/all/GCA/002/571/405/GCA_002571405.2_ASM257140v2

Name	Last modified	Size
Parent Directory		-
GCA_002571405.2_ASM257140v2_assembly_structure/	2025-06-21 17:43	-
GCA_002571405.2_ASM257140v2_assembly_report.txt	2023-12-03 18:24	2.6K
GCA_002571405.2_ASM257140v2_assembly_stats.txt	2023-12-03 18:24	28K
GCA_002571405.2_ASM257140v2_cds_from_genomic.fna.gz	2020-02-05 01:12	2.7M
GCA_002571405.2_ASM257140v2_fcs_report.txt	2025-02-24 00:59	456
GCA_002571405.2_ASM257140v2_feature_count.txt	2023-12-03 18:24	902
GCA_002571405.2_ASM257140v2_feature_table.txt.gz	2020-02-05 01:12	250K
GCA_002571405.2_ASM257140v2_genomic.fna.gz	2021-03-04 17:34	3.6M
GCA_002571405.2_ASM257140v2_genomic.gbff.gz	2023-12-03 18:24	7.3M
GCA_002571405.2_ASM257140v2_genomic.gff.gz	2025-06-21 17:43	656K
GCA_002571405.2_ASM257140v2_genomic.gtf.gz	2025-06-21 17:43	648K
GCA_002571405.2_ASM257140v2_genomic_gaps.txt.gz	2020-02-05 01:12	133
GCA_002571405.2_ASM257140v2_protein.faa.gz	2020-02-05 01:12	1.7M
GCA_002571405.2_ASM257140v2_protein.gpff.gz	2023-12-03 18:24	3.7M
GCA_002571405.2_ASM257140v2_rna_from_genomic.fna.gz	2023-12-03 18:24	2.7M
GCA_002571405.2_ASM257140v2_translated_cds.faa.gz	2020-02-05 01:12	1.9M
README.txt	2024-08-27 13:56	55K
annotation_hashes.txt	2025-06-21 17:43	410
assembly_status.txt	2025-07-24 07:10	14
md5checksums.txt	2025-06-21 17:43	6.5K
uncompressed_checksums.txt	2025-06-21 17:43	3.6K

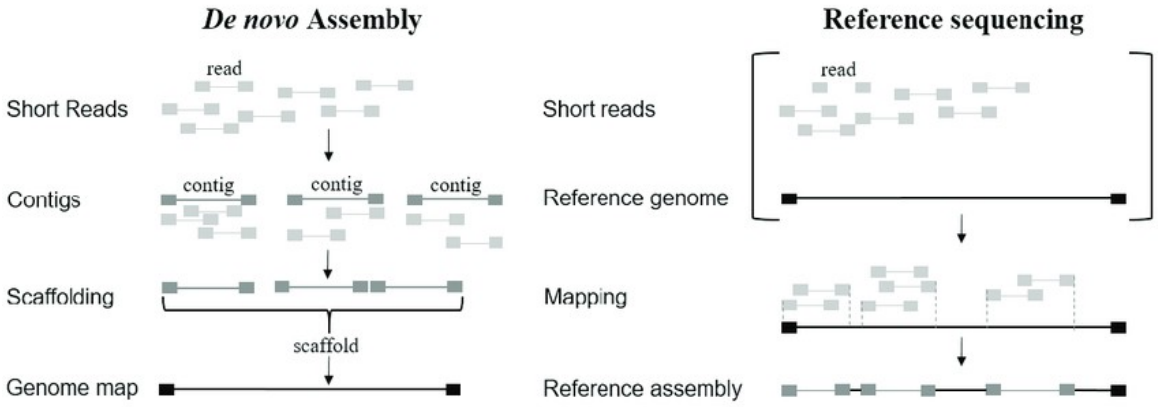
HHS Vulnerability Disclosure

Q.3: What is the difference between assembly and alignment?

	Assembly	**Alignment**
Definition Input Reference? Types	Computational construction of a longer sequence from smaller sequence reads. It can be de-novo (without a reference) or reference-guided.	Similarity-based arrangement of DNA, RNA or protein sequences. It matches sequencing reads to an existing reference genome to determine their origin. Types:- Global or local alignment
Purpose	To build the entire genome or transcript when the reference genome is missing or incomplete.	To find where each read maps on a known reference genome or transcriptome is.
Output	Contigs or scaffolds (assembled sequences) – FASTA format.	SAM or BAM files (coordinates of reads on the reference)
Tools Used	SPAdes, Velvet, MEGAHIT, Trinity (RNA)	BWA, Bowtie2, STAR, HISAT2, Minimap2

Use Cases	To find novel species, metagenomic studies, to discover structural variants, etc.	Variant calling, gene expression analysis, RNA-seq quantification, SNP detection, to annotate genomic features, etc.
Difficulty **CPU usage** **Storage**	Much more complex & resource-intensive, high CPU usage (especially for de-novo) and requires more storage (especially with intermediate contigs).	Relatively fast and lightweight, moderate to low CPU usage, easy to interpret results.
Disadvantages	Prone to fragmentation or misassembly.	Sensitive to reference genome errors.

Q.4: What is the difference between de-novo assembly and reference-based assembly?



	De-Novo Assembly	**Reference-Based Assembly**
Primary Goal	Reconstruct a genome or transcriptome without any reference.	Map reads to a known reference genome to infer variation or expression.
When to Use?	• New or poorly studied organisms. • Metagenomics • Novel virus or bacterial strains.	• Model organisms (human, mouse, yeast) • Clinical genomics • Variant calling, RNA-seq quantification
Genome Bias	No bias; entire sequence assembled from scratch.	Biased towards the reference; novel sequences may be missed.
Data Requirement	Needs high coverage (30x – 100x +) & good quality data to generate reliable contigs.	Can work with lower coverage (~10x – 30x) since structure is already known.
Novel sequence detection	Can detect new genes, rearrangements, or repeat expansions.	Rare or novel variants may be missed or misaligned.

Repetitive Regions	Difficult to resolve repeats accurately; may result in gaps or fragmented assembly.	Easier to handle with reference, as repeats are known and aligned accordingly.
Output Examples	• FASTA files with contigs/scaffolds • Entire genome sequence • Graphs (e.g., De Bruijn)	• BAM/SAM alignment files • VCF for SNP/indel calls • Coverage maps or expression matrices
Computational load	High RAM & CPU needed; more time-consuming (large genomes).	Faster & less memory-intensive due to guided alignment.
Error Sensitivity	Sensitive to sequencing errors; can affect assembly quality or contiguity.	Errors are easier to spot & correct against reference genome.
Tools (Genomic)	SPAdes, SOAPdenovo, ABySS, Canu (for long reads), Flye (for nanopore), etc.	BWA, Bowtie2, Minimap2, GATK, Samtools for post-processing, etc.
Tools (Transcriptomic)	Trinity, Oases (RNA-seq de-novo), RNA-Bloom (long-read RNA), etc.	HISAT2, STAR, Kallisto, Salmon (for quantification), etc.
Use in evolutionary studies	To discover divergent or unique genes in different species or populations.	To compare known variations & SNPs across individuals of the same species.
Scaffolding & Gap Filling	Often requires paired-end or mate-pair reads or long reads to scaffold properly.	Not needed as the reference already provides structure.

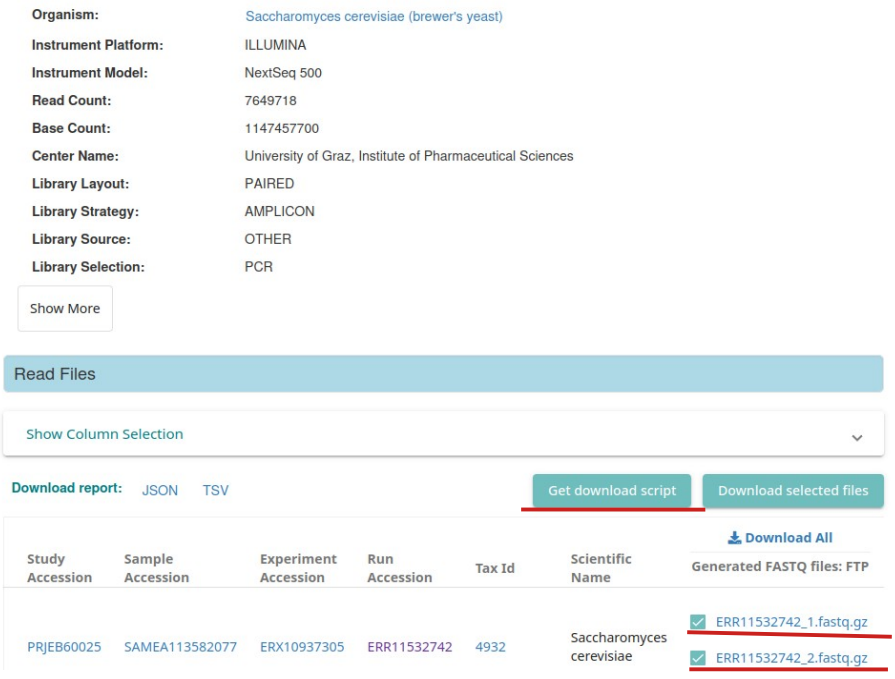
Q.5: What is the difference between FASTA and FASTQ file format?

	FASTA	**FASTQ**
Full name	FAST-All (because it can store any sequence type)	FASTQ (FAST + Quality)
Origin	Developed for the sequence databases like GenBank.	Developed by the Sanger Institute for storing raw sequencing reads.
Lines per entry	2 lines (or more if sequence wraps): Header line (starts with >) and sequence line.	Always 4 lines: header line (starts with @), sequence line, + separator, ASCII-encoded quality.
Quality encoding	Not supported	Usually Phred+33 (Illumina) or Phred+64 (older platforms) $Q = -10 * \log_{10}P$; where, P = Probability of error

Header format	Flexible	Must match the identifier in line 1 and line 3.
Common tools	BLAST, Bowtie (for index), genome browsers, Prokka	FastQC, Trimmomatic, BWA, STAR, kallisto, HISAT2, Salmon
Storage purpose	Reference sequence, consensus genomes, contigs, scaffolds	Raw read data from sequencers (prior to alignment/assembly)
Example	>NC_090774.1 Yarrowia lipolytica chromosome 1E, complete sequence GCGTGCTCAGTCGAATCCT CCACTA	 @seqA ATGCAAGTC + IIIIIIII (I = high quality score)

Q.6: How can I download paired-end nucleotide sequencing data of *Saccharomyces cerevisiae* (e.g. – ERR11532742) from a whole genome sequencing (WGS) experiment using either the NCBI SRA or the ENA (European Nucleotide Archive)?

- Visit ENA homepage - https://www.ebi.ac.uk/ena/browser/home
- Enter the accession ID (ERR11532742) or organism name in the search bar and press 'Enter'.
- Select both the fastq.gz files (forward & reverse-end) and click on 'Get download script'. This contains wget links, so if you run the script then it will automatically both the fastq.gz files. And then gunzip or uncompress these.

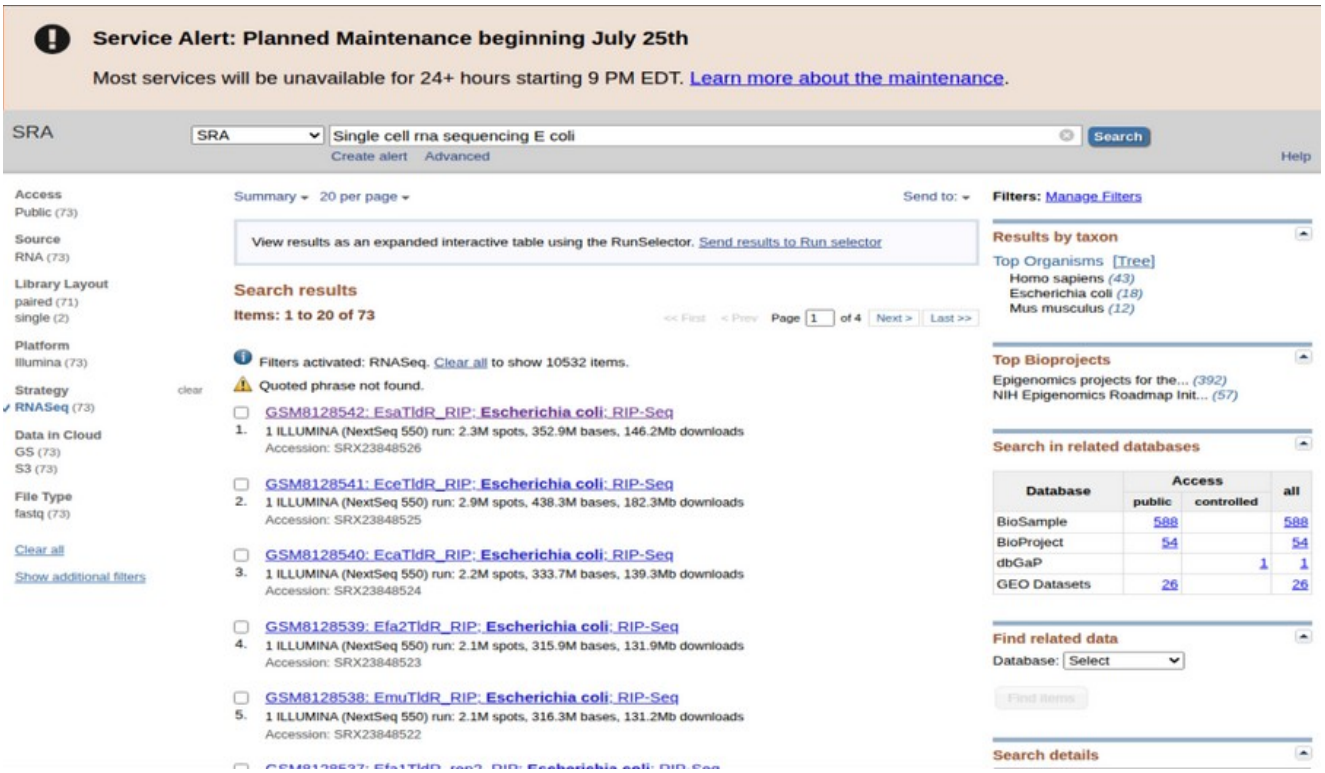


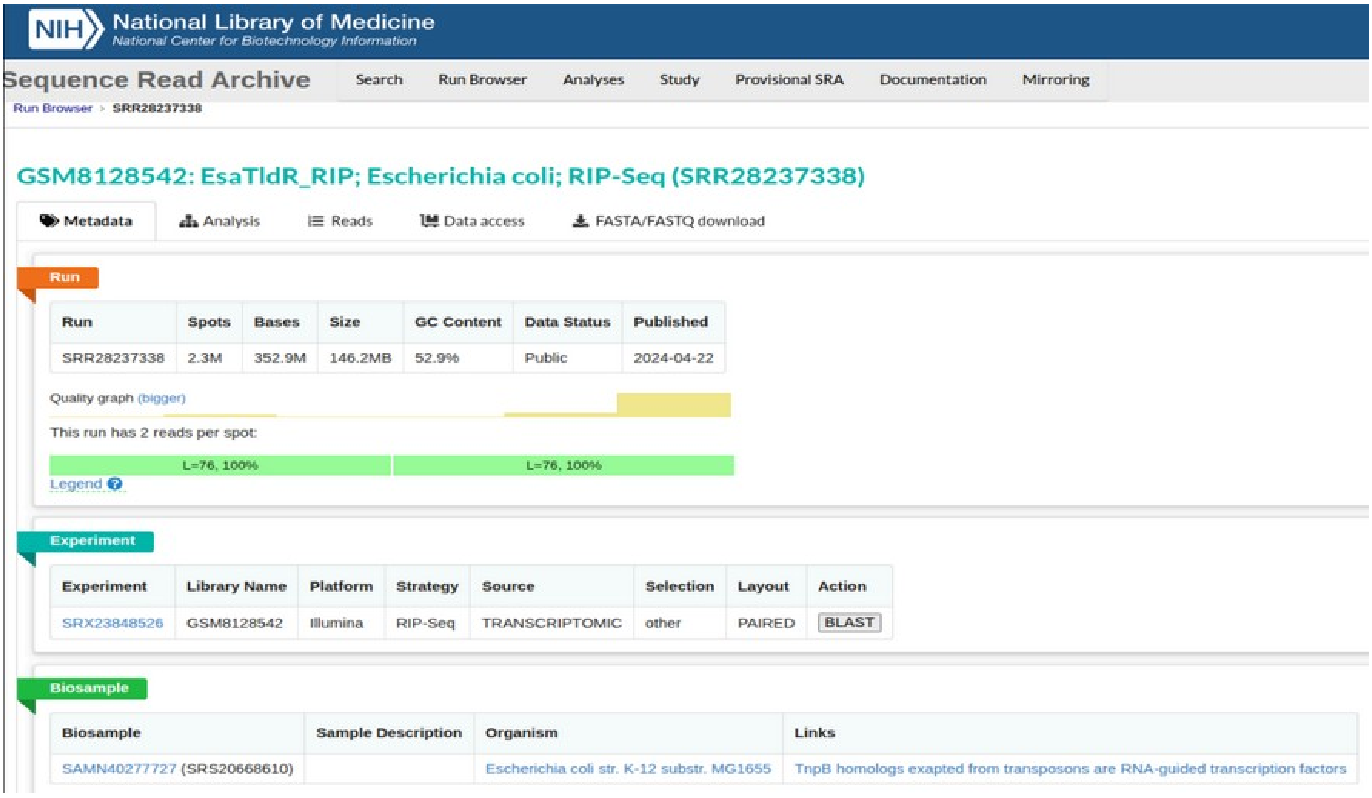

```
(base) ibab@IBAB-MScBDB2-Comp007:~/Downloads$ cat ena-file-download-selected-files-20250823-0627.sh
wget -nc ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR115/042/ERR11532742/ERR11532742_2.fastq.gz
wget -nc ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR115/042/ERR11532742/ERR11532742_1.fastq.gz
```

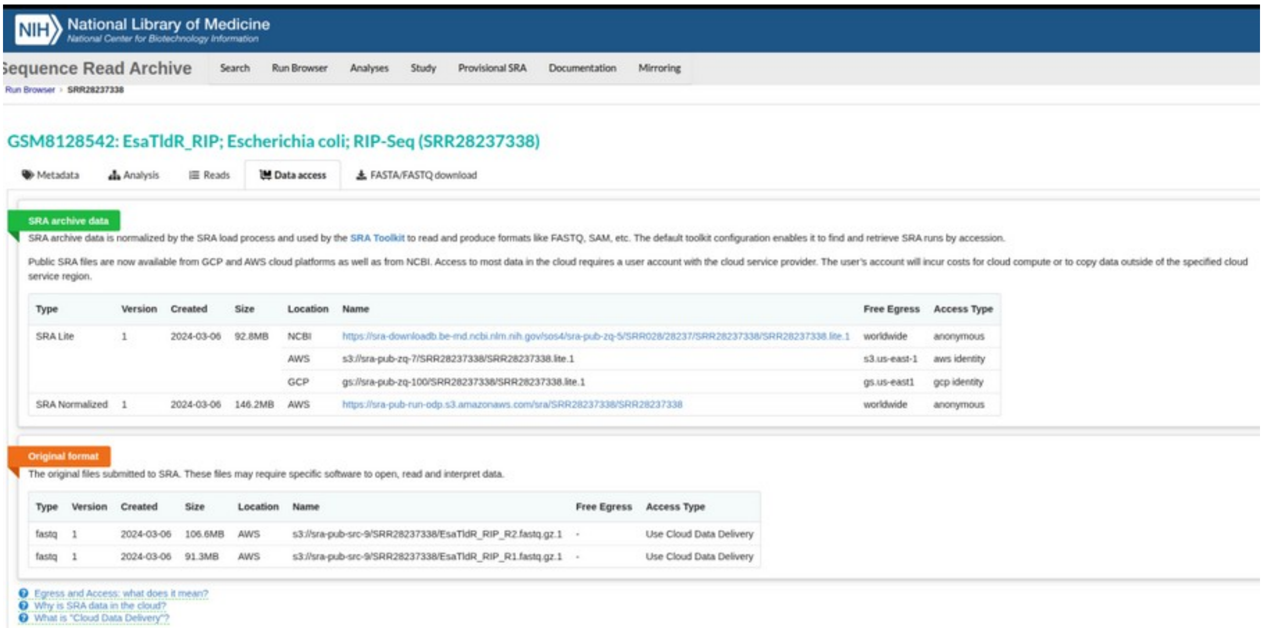
```
(base) ibab@IBAB-MScBDB2-Comp007:~/Downloads$ chmod +x ena-file-download-selected-files-20250823-0627.sh
(base) ibab@IBAB-MScBDB2-Comp007:~/Downloads$ ./ena-file-download-selected-files-20250823-0627.sh
--2025-08-23 11:58:25--  ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR115/042/ERR11532742/ERR11532742_2.fastq.gz
Resolving proxy.ibab.ac.in (proxy.ibab.ac.in)... 192.168.1.254
Connecting to proxy.ibab.ac.in (proxy.ibab.ac.in)|192.168.1.254|:3128... connected.
Proxy request sent, awaiting response... 200 Gatewaying
Length: 229187791 (219M) [text/plain]
Saving to: 'ERR11532742_2.fastq.gz'

ERR11532742_2.fastq.gz              100%[=====================================================================>] 218.57M  2.44MB/s    in 99s

2025-08-23 12:00:16 (2.21 MB/s) - 'ERR11532742_2.fastq.gz' saved [229187791/229187791]

--2025-08-23 12:00:16--  ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR115/042/ERR11532742/ERR11532742_1.fastq.gz
Resolving proxy.ibab.ac.in (proxy.ibab.ac.in)... 192.168.1.254
Connecting to proxy.ibab.ac.in (proxy.ibab.ac.in)|192.168.1.254|:3128... connected.
Proxy request sent, awaiting response... 200 Gatewaying
Length: 232721564 (222M) [text/plain]
Saving to: 'ERR11532742_1.fastq.gz'

ERR11532742_1.fastq.gz              100%[=====================================================================>] 221.94M  3.03MB/s    in 76s

2025-08-23 12:01:34 (2.93 MB/s) - 'ERR11532742_1.fastq.gz' saved [232721564/232721564]
```

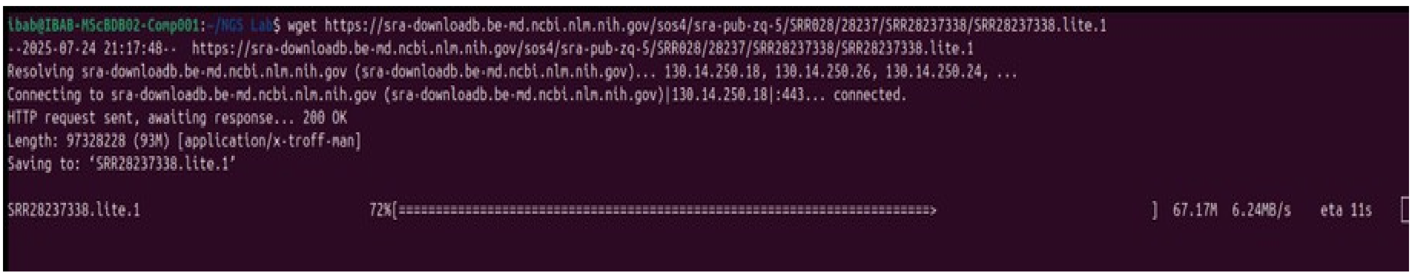
Q.7: Download the single-cell RNA-Seq data of E. Coli or Saccharomyces cerevisiae; the one having small size.

1. Go to the SRA website.
2. Type “Single cell RNA-seq E. Coli” in the search bar & press ‘Enter’.
3. Click on the first hit.
4. Go below the page and click on the SRA metadata link.
5. From the Data Access column, copy the NCBI link.
6. Open terminal and navigate to the required folder and download using the wget command.




NIH National Library of Medicine
National Center for Biotechnology Information
Sequence Read Archive
Search
Run Browser
Analyses
Study
Provisional SRA
Documentation
Mirroring
Run Browser › SRR28237338
GSM8128542: EsaTldR_RIP; Escherichia coli; RIP-Seq (SRR28237338)
Metadata
Analysis
Reads
Data access
FASTA/FASTQ download
Run
Run | Spots | Bases | Size | GC Content | Data Status | Published
SRR28237338 | 2.3M | 352.9M | 146.2MB | 52.9% | Public | 2024-04-22
Quality graph (bigger)
This run has 2 reads per spot:
L=76, 100%
L=76, 100%
Legend
Experiment
Experiment | Library Name | Platform | Strategy | Source | Selection | Layout | Action
SRX23848526 | GSM8128542 | Illumina | RIP-Seq | TRANSCRIPTOMIC | other | PAIRED | BLAST
Biosample
Biosample | Sample Description | Organism | Links
SAMN40277727 (SRS20668610) | | Escherichia coli str. K-12 substr. MG1655 | TnpB homologs exapted from transposons are RNA-guided transcription factors


NIH National Library of Medicine
National Center for Biotechnology Information
Sequence Read Archive
Search
Run Browser
Analyses
Study
Provisional SRA
Documentation
Mirroring
Run Browser › SRR28237338
GSM8128542: EsaTldR_RIP; Escherichia coli; RIP-Seq (SRR28237338)
Metadata
Analysis
Reads
Data access
FASTA/FASTQ download
SRA archive data
SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.
Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.
Type | Version | Created | Size | Location | Name | Free Egress | Access Type
SRA Lite | 1 | 2024-03-06 | 92.8MB | NCBI | https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos4/sra-pub-zq-5/SRR028/28237/SRR28237338/SRR28237338.lite.1 | worldwide | anonymous
AWS | s3://sra-pub-zq-7/SRR28237338/SRR28237338.lite.1 | s3.us-east-1 | aws identity
GCP | gs://sra-pub-zq-100/SRR28237338/SRR28237338.lite.1 | gs.us-east1 | gcp identity
SRA Normalized | 1 | 2024-03-06 | 146.2MB | AWS | https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR28237338/SRR28237338 | worldwide | anonymous
Original format
The original files submitted to SRA. These files may require specific software to open, read and interpret data.
Type | Version | Created | Size | Location | Name | Free Egress | Access Type
fastq | 1 | 2024-03-06 | 106.6MB | AWS | s3://sra-pub-src-9/SRR28237338/EsaTldR_RIP_R2.fastq.gz.1 | - | Use Cloud Data Delivery
fastq | 1 | 2024-03-06 | 91.3MB | AWS | s3://sra-pub-src-9/SRR28237338/EsaTldR_RIP_R1.fastq.gz.1 | - | Use Cloud Data Delivery
Egress and Access: what does it mean?
Why is SRA data in the cloud?
What is "Cloud Data Delivery"?


ibab@IBAB-MScBDB02-Comp001:~/NGS Lab$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos4/sra-pub-zq-5/SRR028/28237/SRR28237338/SRR28237338.lite.1
--2025-07-24 21:17:48-- https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos4/sra-pub-zq-5/SRR028/28237/SRR28237338/SRR28237338.lite.1
Resolving sra-downloadb.be-md.ncbi.nlm.nih.gov (sra-downloadb.be-md.ncbi.nlm.nih.gov)... 130.14.250.18, 130.14.250.26, 130.14.250.24, ...
Connecting to sra-downloadb.be-md.ncbi.nlm.nih.gov (sra-downloadb.be-md.ncbi.nlm.nih.gov)|130.14.250.18|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 97328228 (93M) [application/x-troff-man]
Saving to: 'SRR28237338.lite.1'
SRR28237338.lite.1 72%[===>] 67.17M 6.24MB/s eta 11s

Q.8: Download paired-end RNA-seq data of Escherichia coli from a public repository (NCBI SRA or ENA) with the following experimental design:

1. **Conditions: 3 samples under healthy (control) condition, 3 samples under diseased treatment condition at different time points.**
2. **Replicates: Each condition/time point has 3 biological replicates.**