Turn on your Laptop

Connect to Internet (Check your connection)

Launch WSL (Windows) or Terminal (macOS/Linux)

Tutorial Outline

1. Downloading data from NCBI
   a. RefSeq vs GenBank
   b. FNA vs GTF vs GFF vs GBFF

2. Brief introduction to k-mers, GC, Clump and Origin

3. Practice Tasks
   a. ORI signal checker: K-mer enrichment and plotting
   b. Clump finder: L,k,t clumps
   c. GC skew calculator
   d. ORI (Origin of Replication) finder

National Center for Biotechnology Information (NCBI)

1. **Public bioinformatics resource** maintained by the NIH (US).
2. Provides access to a w**ide range of biological databases, analysis tools, and reference datasets**.
3. Covers **genomics, transcriptomics, proteomics, and biomedical literature** datasets.

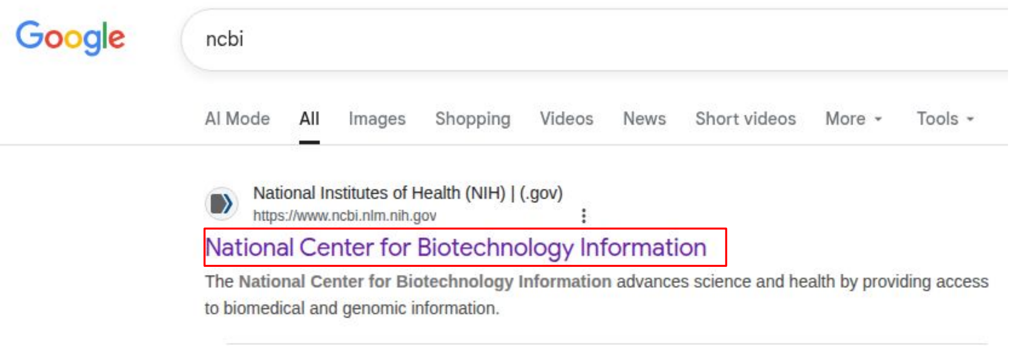
Key NCBI Databases:

1. **GenBank**: comprehensive, public repository of sequence data
2. **RefSeq**: non-redundant, curated, and standardized reference sequence data.

Common NCBI File Formats:

1. FASTA Nucleotide (**FNA**): Genomic Sequence
2. Gene Transfer Format (**GTF**): gene and transcript annotations
3. General Feature Format (**GFF**): genomic features such as genes, CDS, exons, etc.
4. GenBank Flat File (**GBFF**): human-readable annotated genome file

Downloading Data

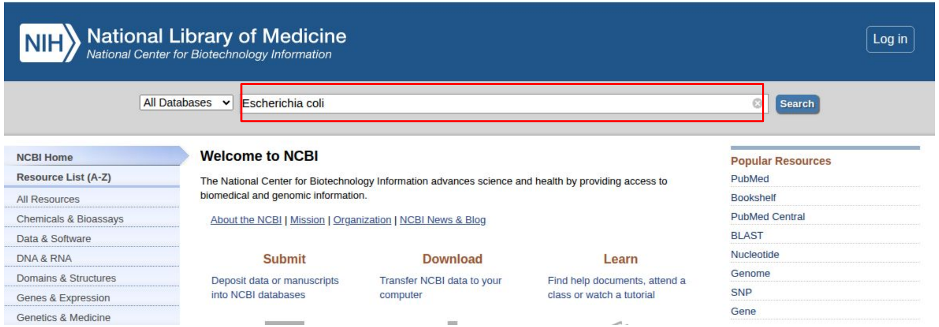


ncbi

National Institutes of Health (NIH) | (.gov)
https://www.ncbi.nlm.nih.gov

National Center for Biotechnology Information

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.



Escherichia coli

Welcome to NCBI

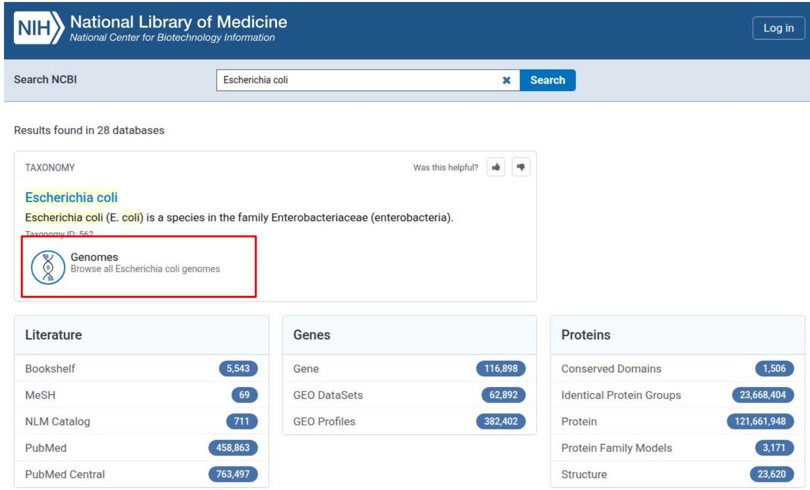
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial


NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
Search
Was this helpful?

Escherichia coli

Results found in 28 databases

TAXONOMY

Escherichia coli

Escherichia coli (E. coli) is a species in the family Enterobacteriaceae (enterobacteria).

Taxonomy ID: 562

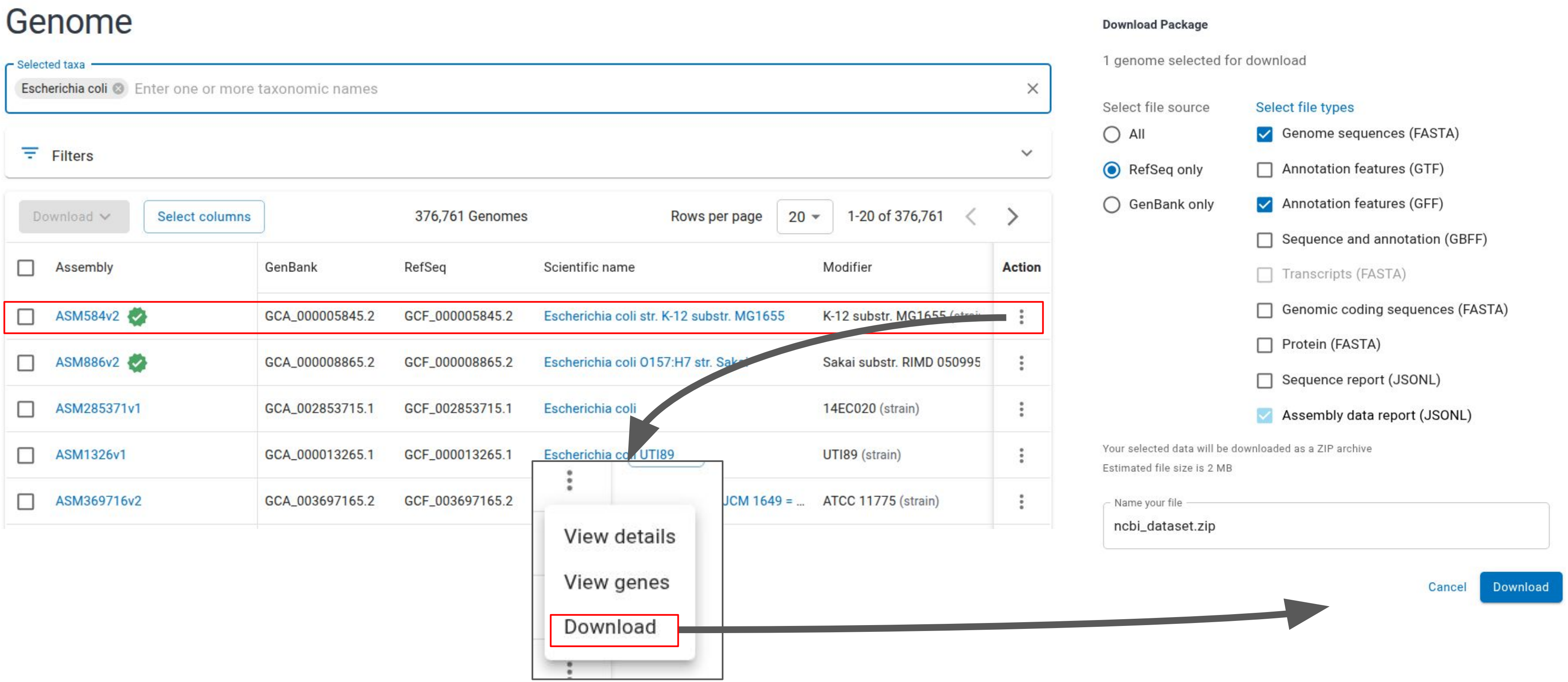
Genomes
Browse all Escherichia coli genomes

Literature	
Bookshelf	5,543
MeSH	69
NLM Catalog	711
PubMed	458,863
PubMed Central	763,497

Genes	
Gene	116,898
GEO DataSets	62,892
GEO Profiles	382,402

Proteins	
Conserved Domains	1,506
Identical Protein Groups	23,668,404
Protein	121,661,948
Protein Family Models	3,171
Structure	23,620

Downloading Data *Continued*



Genome

Selected taxa: Escherichia coli ⊗ Enter one or more taxonomic names ✕

Filters

376,761 Genomes — Rows per page 20 — 1-20 of 376,761

□ Assembly	GenBank	RefSeq	Scientific name	Modifier	Action
□ ASM584v2 ✓	GCA_000005845.2	GCF_000005845.2	Escherichia coli str. K-12 substr. MG1655	K-12 substr. MG1655 (strai	⋮
□ ASM886v2 ✓	GCA_000008865.2	GCF_000008865.2	Escherichia coli O157:H7 str. Sak	Sakai substr. RIMD 050995	⋮
□ ASM285371v1	GCA_002853715.1	GCF_002853715.1	Escherichia coli	14EC020 (strain)	⋮
□ ASM1326v1	GCA_000013265.1	GCF_000013265.1	Escherichia coli UTI89	UTI89 (strain)	⋮
□ ASM369716v2	GCA_003697165.2	GCF_003697165.2	UCM 1649 = ...	ATCC 11775 (strain)	⋮

Download Package

1 genome selected for download

Select file source

- ○ All
- ● RefSeq only
- ○ GenBank only

Select file types

- ☑ Genome sequences (FASTA)
- □ Annotation features (GTF)
- ☑ Annotation features (GFF)
- □ Sequence and annotation (GBFF)
- □ Transcripts (FASTA)
- □ Genomic coding sequences (FASTA)
- □ Protein (FASTA)
- □ Sequence report (JSONL)
- ☑ Assembly data report (JSONL)

Your selected data will be downloaded as a ZIP archive

Estimated file size is 2 MB

Name your file: ncbi_dataset.zip

Brief introduction to k-mers

1. Substring of length k extracted from a biological sequence data.

2. Have application in genome assembly, sequence comparison and and clustering, etc.

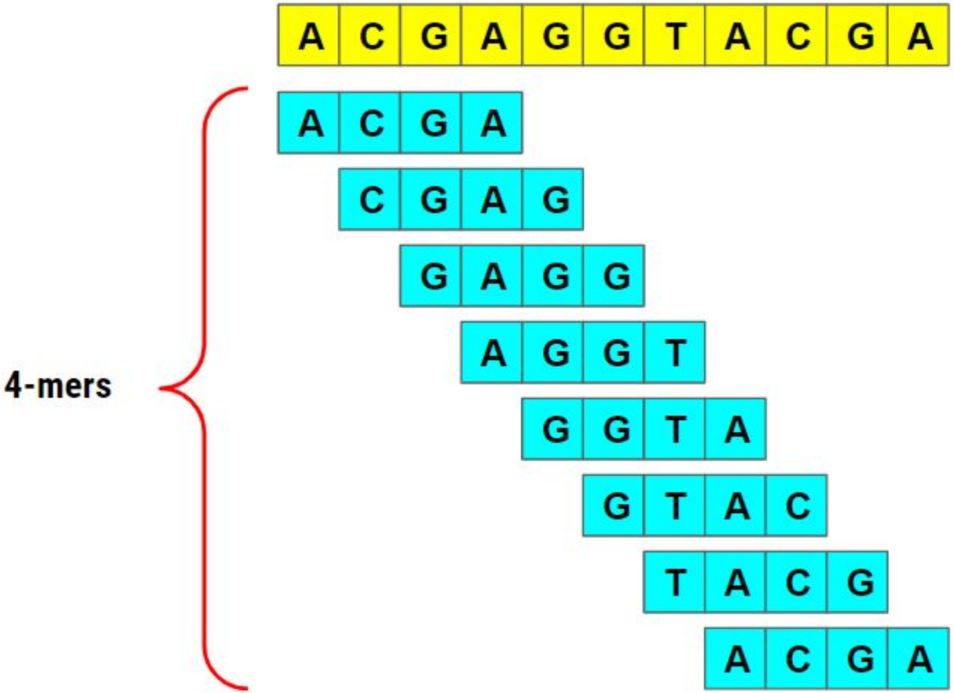


Image Source: https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29

Brief introduction to GC

1. Refers to the percentage of guanine (G) and cytosine (C) nucleotides in a DNA sequences.

2. GC content affects gene density, replication and transcription efficiency, etc.

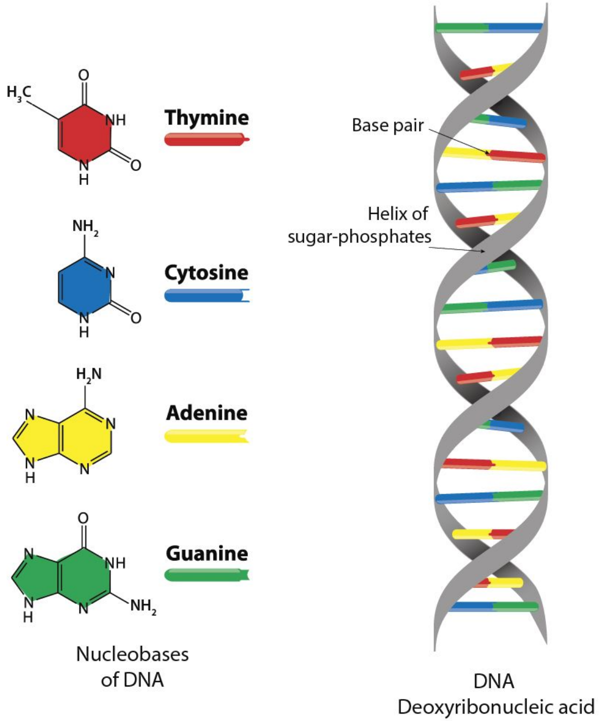


Image Source:
https://www.technologynetworks.com/genomics/articles/what-are-the-key-differences-between-dna-and-rna-296719

Brief introduction to (L, k, t)-Clumps

1. Identify locally frequent k-mers within a DNA sequence.

2. A k-mer is said to form an (L, k, t)-clump if it appears at least t times within any window of length L in the genome

3. Given:
   - k → length of the k-mer
   - L → length of the sliding window
   - t → minimum number of occurrences

Brief introduction to Origin (ORI)

1. Origin of Replication (ORI) is a specific genomic region where DNA replication begins.

2. Often AT-rich, making strand separation easier

3. Usually single ORI in bacteria and multiple ORIs in eukaryotes

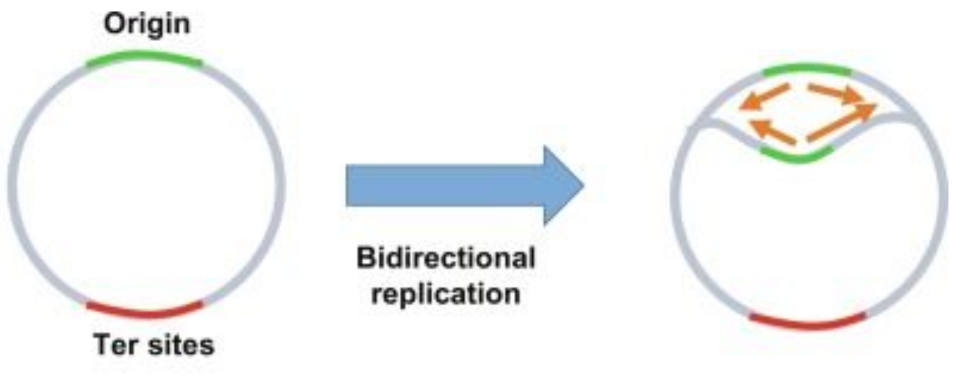


Image Source: https://doi.org/10.1016/B978-0-323-91788-9.00006-5

Vibe Coding Session

Practice Tasks

a. ORI signal checker: K-mer enrichment and plotting
b. Clump finder: L,k,t clumps
c. GC skew calculator
d. ORI (Origin of Replication) finder

Vibe Coders looking at their own code after exhausting their credits

Thank You