BIF701 : Genomic Analysis
Practical file - Semester 3

Prathamesh Shelke
M. Sc. H. Bioinformatics
A270150423001
Department of Bioinformatics
Amity School of Biological Sciences
Amity University Punjab

Dr. Jaishree Meena
Department of Bioinformatics
Amity School of Biological Sciences
Amity University Punjab

Table of Content

Experiment 1: Prokaryotic Genome Annotation using PROKKA

Aim: To use PROKKA for Bacterial genome annotation of *Mycobacterium ulcerans* Agy99

Introduction:

Genome Annotation is the process of describing the structure and function of the components of a genome, by analysing and interpreting them in order to extract their biological significance and understand the biological processes in which they participate. The process of analysing, locating, labelling and assigning functions to genomic features on a genome is termed as Genome Annotation. These genomic features include Protein coding regions, RNA (mRNA, tRNA, ncRNA, etc), etc. This genome annotation is performed on the contigs (A contig is a set of overlapping DNA segments that together represent a consensus region of DNA) and scaffolds (scaffolds refer to three-dimensional structures that provide support and organisation for cells, tissues, and organs). Different tools available for such studies viz. PGAP (Prokaryotic Genome Annotation Pipeline), RAST (Rapid Annotation using Subsystem Technology), PROKKA, etc.

Examples of Prokka Options:

--prefix	Specifies a prefix for output filenames.
--outdir	Sets the output directory where all the annotated files will be stored.
--locustag	Assigns a locus tag prefix for annotated genes.
--cpus	Defines the number of CPU cores to use during the annotation process.
--kingdoms	Specifies the kingdom of the organism being annotated.
--addgenes	Forces PROKKA to add gene features to the output.
--proteins	Allows you to provide a FASTA file of known proteins for use in the annotation.

Materials and Methods:

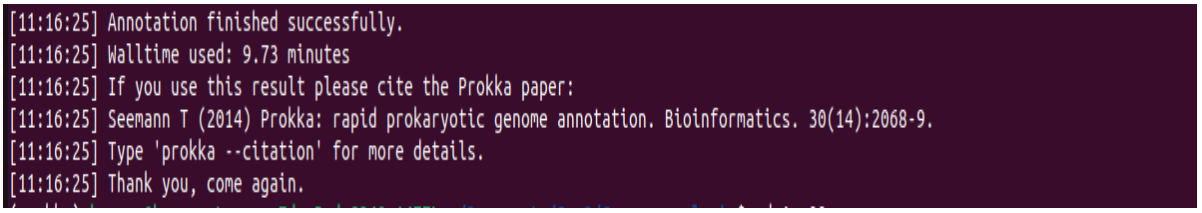
Tools used for the study were Prokka and Linux

The following command was used

```
prokka --cpus 4 --prefix Agy99 --kingdom Bacteria --locustag Agy99 ~/Downloads/Genomic_Analysis/Genome \Annotation/Agy99.fasta
```

```
(base) harman@harman-Lenovo-IdeaPad-S340-14IIL:~/Documents/Sem3/Genome analysis$ conda activate prokka
(prokka) harman@harman-Lenovo-IdeaPad-S340-14IIL:~/Documents/Sem3/Genome analysis$ prokka --cpus 4 --prefix Agy99 --kingdom Bacteria --locustag Agy99 Agy99.fasta
[11:06:41] This is prokka 1.14.6
[11:06:41] Written by Torsten Seemann <torsten.seemann@gmail.com>
[11:06:41] Homepage is https://github.com/tseemann/prokka
[11:06:41] Local time is Thu Oct 24 11:06:41 2024
[11:06:41] You are harman
[11:06:41] Operating system is linux
[11:06:41] You have BioPerl 1.7.8
Argument "1.7.8" isn't numeric in numeric lt (<) at /home/harman/anaconda3/envs/prokka/bin/prokka line 259.
[11:06:41] System has 8 cores.
```

…



A directory named Agy99 is created that contains the files generated by prokka.
Here's a tabulated summary of the different files generated by Prokka:

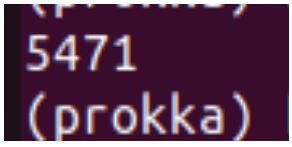
File Name	Description	Format/Use
<basename>.txt	Summary of annotation statistics (genes, CDS, rRNA, tRNA, etc.).	Plain text (summary overview).
<basename>.fna	Nucleotide FASTA file of annotated sequences (contigs/scaffolds).	FASTA (DNA sequences).
<basename>.faa	Protein FASTA file of predicted amino acid sequences from CDS.	FASTA (protein sequences).
<basename>.ffn	Nucleotide FASTA file of all annotated features (CDS, rRNA, tRNA, etc.).	FASTA (DNA sequences for all features).
<basename>.gff	GFF3 file containing all annotated genomic features.	GFF3 (genome annotation and feature visualization).
<basename>.gbk	GenBank file with annotations and sequences.	GenBank (standard genome sharing format).
<basename>.tsv	Tab-delimited file listing annotated features (locus tags, product descriptions, coordinates, etc.).	TSV (structured feature table).
<basename>.tbl	Feature table in NCBI submission format.	TBL (for GenBank submission).
<basename>.sqn	Sequin file for direct submission to NCBI.	NCBI Sequin format.
<basename>.hmm	Hidden Markov Model file for specific annotations (if HMMER is used).	HMM (used for annotation refinement).

`<basename>.err`	Log file of errors or warnings encountered during annotation.	Plain text (debugging errors).
`<basename>.log`	Detailed log of Prokka's processing steps.	Plain text (process log).
`<basename>.raw`	Raw output from Prodigal (gene prediction tool).	Prodigal-specific format (optional).
`<basename>.tbl out`	Tabular output of HMMER searches (if used).	Tab-delimited (optional, HMMER results).

Each file provided complementary information about the structural and functional components of the genome.

Result and Analysis: Exploring some of the output files

Command: `grep '>' Agy99.faa|wc -l`



Command: `head Agy99.faa`

Command: `grep 'hypothetical' Agy99.faa'`

Command: `grep 'hypothetical' Agy99.faa|wc -l`

```
(prokka)
2516
```

Command: `grep '>' Agy99.ffn|wc -l`

```
(prokka)
5523
```

Command: `grep 'gene' Agy99.ffn|wc -l`

```
(prokka)
7
```

Command: `grep 'gene' Agy99.ffn`

```
>Agy99_00246 Formylglycine-generating enzyme
>Agy99_00685 Cytochrome c biogenesis protein Ccs1
>Agy99_00686 Cytochrome c biogenesis protein CcsA
>Agy99_00887 Formylglycine-generating enzyme
>Agy99_01991 Putative gluconeogenesis factor
>Agy99_02794 Small ribosomal subunit biogenesis GTPase RsgA
>Agy99_03250 putative hydrogen peroxide-inducible genes activator
```

Command: `head Agy99.tsv`

```
(prokka) harman@harman-Lenovo-IdeaPad-S340-14IIL:~/Documents/Sem3/Genome analysis/Agy99$ head Agy99.tsv
locus_tag       ftype   length_bp       gene    EC_number       COG     product
Agy99_00001     CDS     1500    dnaA            COG0593 Chromosomal replication initiator protein DnaA
Agy99_00002     CDS     1209    dnaN            COG0592 Beta sliding clamp
Agy99_00003     CDS     1080    recF            COG1195 DNA replication and repair protein RecF
Agy99_00004     CDS     564                             hypothetical protein
Agy99_00005     CDS     2079    gyrB1   5.6.2.2         DNA gyrase subunit B
Agy99_00006     CDS     2520    gyrA1   5.6.2.2         DNA gyrase subunit A
Agy99_00007     tRNA    75                              tRNA-Ile(gat)
Agy99_00008     tRNA    74                              tRNA-Ala(tgc)
Agy99_00009     CDS     840     tam_1   2.1.1.144               Trans-aconitate 2-methyltransferase
```

Conclusion:

	locus_tag	ftyp	length_b	gene	EC_numb	COG	product
2	Agy99_00001	CDS	1500	dnaA		COG0593	Chromosomal replication initiator protein DnaA
3	Agy99_00002	CDS	1209	dnaN		COG0592	Beta sliding clamp
4	Agy99_00003	CDS	1080	recF		COG1195	DNA replication and repair protein RecF
5	Agy99_00004	CDS	564				hypothetical protein
6	Agy99_00005	CDS	2079	gyrB1	5.6.2.2		DNA gyrase subunit B
7	Agy99_00006	CDS	2520	gyrA1	5.6.2.2		DNA gyrase subunit A
8	Agy99_00007	tRNA	75				tRNA-Ile(gat)
9	Agy99_00008	tRNA	74				tRNA-Ala(tgc)
10	Agy99_00009	CDS	840	tam_1	2.1.1.144		Trans-aconitate 2-methyltransferase
11	Agy99_00010	CDS	762	tipA			HTH-type transcriptional activator TipA
12	Agy99_00011	CDS	405	cwsA			Cell wall synthesis protein CwsA
13	Agy99_00012	CDS	549	ppiA	5.2.1.8	COG0652	Peptidyl-prolyl cis-trans isomerase A
14	Agy99_00013	CDS	426				hypothetical protein
15	Agy99_00014	CDS	282	crgA			Cell division protein CrgA
16	Agy99_00015	CDS	744				hypothetical protein
17	Agy99_00016	CDS	687	trpG	4.1.3.27	COG0512	Anthranilate synthase component 2
18	Agy99_00017	CDS	1881	pknB	2.7.11.1	COG0515	Serine/threonine-protein kinase PknB
19	Agy99_00018	CDS	1359	pknA	2.7.11.1	COG0515	Serine/threonine-protein kinase PknA
20	Agy99_00019	CDS	1476	pbpA		COG0768	Penicillin-binding protein A
21	Agy99_00020	CDS	1410	rodA		COG0772	putative FtsW-like protein
22	Agy99_00021	CDS	1560	pstP	3.1.3.16	COG0631	PP2C-family Ser/Thr phosphatase
23	Agy99_00022	CDS	468	fhaB		COG1716	FHA domain-containing protein FhaB
24	Agy99_00023	CDS	1599	fhaA		COG1716	FHA domain-containing protein FhaA
25	Agy99_00024	CDS	1335				IS256 family transposase IS2606
26	Agy99_00025	tRNA	84				tRNA-Leu(cag)
27	Agy99_00026	CDS	744				hypothetical protein
28	Agy99_00027	CDS	675		1.-.-.-		putative oxidoreductase
29	Agy99_00028	CDS	753				hypothetical protein
30	Agy99_00029	CDS	525	osmX		COG1732	Osmoprotectant-binding protein OsmX
31	Agy99_00030	CDS	459				hypothetical protein
32	Agy99_00031	CDS	315	opuCB		COG1174	Glycine betaine/carnitine/choline transport system permease protein OpuCB
33	Agy99_00032	CDS	234				hypothetical protein
34	Agy99_00033	CDS	1200	opuCA		COG1125	Glycine betaine/carnitine/choline transport ATP-binding protein OpuCA
35	Agy99_00034	CDS	669	opuBB_1		COG1174	Choline transport system permease protein OpuBB
36	Agy99_00035	CDS	408				hypothetical protein
37	Agy99_00036	CDS	948				hypothetical protein
38	Agy99_00037	CDS	1314	nrgA			Ammonium transporter
39	Agy99_00038	CDS	300				hypothetical protein
40	Agy99_00039	CDS	771				hypothetical protein
41	Agy99_00040	CDS	417	whiB5			Transcriptional regulator WhiB5
42	Agy99_00041	CDS	792			COG1396	putative HTH-type transcriptional regulator
43	Agy99_00042	CDS	828				hypothetical protein
44	Agy99_00043	CDS	381				putative protein
45	Agy99_00044	CDS	1266				hypothetical protein
46	Agy99_00045	CDS	318				hypothetical protein
47	Agy99_00046	CDS	405				hypothetical protein

The PROKKA pipeline successfully annotated the genome of *Mycobacterium ulcerans* Agy99. The results included protein-coding regions, RNAs (mRNAs, tRNAs, etc.), and other genomic features. PROKKA proved to be an efficient tool for prokaryotic genome annotation, offering user-friendly options to customize outputs. By generating a comprehensive set of files (e.g., .fa, .faa, .gbk), it allowed further exploration of the genome's functional and structural components. This annotation serves as a foundation for studying *M. ulcerans* at a deeper level, enabling insights into its biology, pathogenicity, and potential drug targets.

For instance, examining the .faa file could reveal critical enzymes or proteins unique to this bacterium and so on.

Experiment 2: Eukaryotic Genome Annotation using PROKKA

Aim: To perform genome annotation for the Plasmodium falciparum 3D7 Apicoplast genome using Prokka.

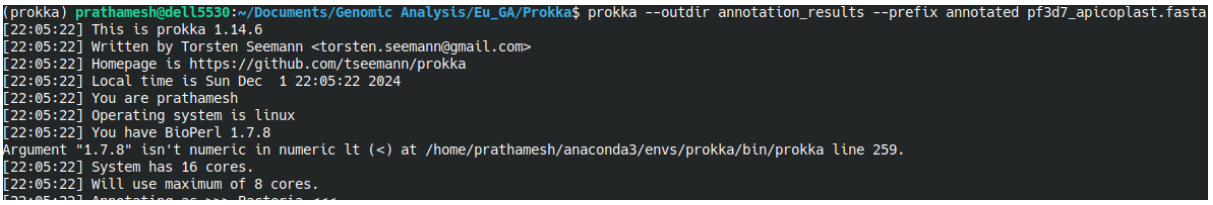
Introduction: Genome annotation is a critical step in understanding the functional elements encoded in a genome. It involves identifying genes, regulatory elements, and other features within the DNA sequence and assigning functions to them based on available biological data. Eukaryotic genome annotation is particularly challenging due to the complexity of eukaryotic genomes, which often include large introns, repetitive sequences, and non-coding RNA genes.

Materials and Methods:

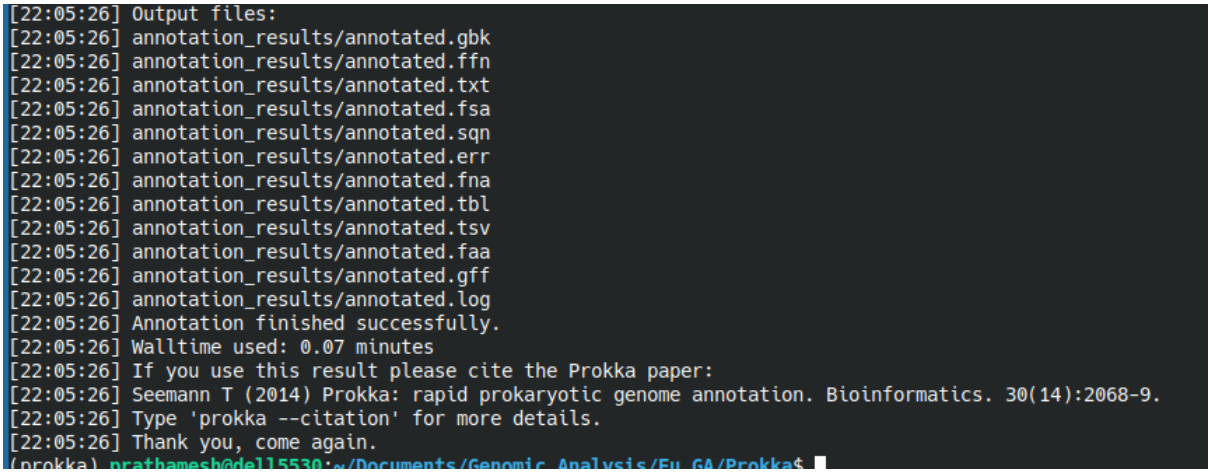
> Genome sequence (*Plasmodium falciparum* 3D7 Apicoplast FASTA format).
> https://www.ncbi.nlm.nih.gov/nuccore/CP131995.1

Tools used for the study were Prokka and Linux
The following command was used:

```
prokka --outdir annotation_results --prefix annotated
pf3d7_apicoplast.fasta
```



…



A directory named Annotated is created that contains the files generated by prokka.
Here's a tabulated summary of the different files generated by Prokka:

File Name	Description	Format/Use
<basename>.txt	Summary of annotation statistics (genes, CDS, rRNA, tRNA, etc.).	Plain text (summary overview).
<basename>.fna	Nucleotide FASTA file of annotated sequences (contigs/scaffolds).	FASTA (DNA sequences).
<basename>.faa	Protein FASTA file of predicted amino acid sequences from CDS.	FASTA (protein sequences).
<basename>.ffn	Nucleotide FASTA file of all annotated features (CDS, rRNA, tRNA, etc.).	FASTA (DNA sequences for all features).
<basename>.gff	GFF3 file containing all annotated genomic features.	GFF3 (genome annotation and feature visualization).
<basename>.gbk	GenBank file with annotations and sequences.	GenBank (standard genome sharing format).
<basename>.tsv	Tab-delimited file listing annotated features (locus tags, product descriptions, coordinates, etc.).	TSV (structured feature table).
<basename>.tbl	Feature table in NCBI submission format.	TBL (for GenBank submission).
<basename>.sqn	Sequin file for direct submission to NCBI.	NCBI Sequin format.
<basename>.hmm	Hidden Markov Model file for specific annotations (if HMMER is used).	HMM (used for annotation refinement).
<basename>.err	Log file of errors or warnings encountered during annotation.	Plain text (debugging errors).
<basename>.log	Detailed log of Prokka's processing steps.	Plain text (process log).
<basename>.raw	Raw output from Prodigal (gene prediction tool).	Prodigal-specific format (optional).

`<basename>.tbl`	Tabular output of HMMER searches (if used).	Tab-delimited (optional, HMMER results).
`out`		

Each file provided complementary information about the structural and functional components of the genome.

Result and Analysis: Exploring some of the output files

Command: `grep ">" annotated.faa|wc -l`

```
(prokka) pratha
0
```

Command: `head annotated.faa`

```
(prokka) prathamesh@dell5530:~/Documents/Genomic Analysis/Eu_GA/Prokka/annotation_results$ head annotated.faa
>CDEGBLJO_00001 hypothetical protein
MIKFLKPKIKILKKLNIPFLLYLSSKYNYKYLNYKISYKSYFDLKLKFIRYICYNYCITY
KKYLYYLNKIDNKNINILYFKLLKILELRLDIFLVNIGFFKTILQSRYYIKYKNIYINNI
INKYYNINLKNNDILFFNNKIKYIILKNLIYKYNIYIYISNLYKYNFIKIYSYNKYFIIC
IYNFKIKILNINNILNNILYIYNDIYYI
>CDEGBLJO_00011 hypothetical protein
MNIIILNNNTLNNIIFKYKYNFFIKLYFNNYIKICKLIIYIIKYLYIYNIYMYKHTKNKS
KVYFSNKKIRVQKGLGKARLKNFKSPVCKQGACNFGPFYKENKIISKINYRLIFVYLLIN
KRSNIIIIKLENIINLLNIFYKNKNYCIFKLLYLKGIINNKYILINLNNKLFNKNIFINI
IMYNYLIFLI
```

Command: `grep '>' annotated.ffn|wc -l`

```
(prokka) prat
21
```

Command: `grep '>' annotated.ffn|wc -l`

```
(prokka)
67
```

Command: `head annotated.tsv`

```
(prokka) prathamesh@dell5530:~/Documents/Genomic Analysis/Eu_GA/Prokka/annotation_results$ head annotated.tsv
locus_tag       ftype   length_bp       gene    EC_number       COG     product
CDEGBLJO_00001  CDS     627                             hypothetical protein
CDEGBLJO_00002  tRNA    74                              tRNA-His(gtg)
CDEGBLJO_00003  tRNA    72                              tRNA-Cys(gca)
CDEGBLJO_00004  tRNA    87                              tRNA-Met(cat)
CDEGBLJO_00005  tRNA    84                              tRNA-Tyr(gta)
CDEGBLJO_00006  tRNA    92                              tRNA-Ser(gct)
CDEGBLJO_00007  tRNA    75                              tRNA-Asp(gtc)
CDEGBLJO_00008  tRNA    73                              tRNA-Lys(ttt)
CDEGBLJO_00009  tRNA    71                              tRNA-Glu(ttc)
```

Conclusion:

	A	B	C	D	E	F	G
1	locus_tag	ftype	length_bp	gene	EC_number	COG	product
2	CDEGBLJO_00001	CDS	627				hypothetical protein
3	CDEGBLJO_00002	tRNA	74				tRNA-His(gtg)
4	CDEGBLJO_00003	tRNA	72				tRNA-Cys(gca)
5	CDEGBLJO_00004	tRNA	87				tRNA-Met(cat)
6	CDEGBLJO_00005	tRNA	84				tRNA-Tyr(gta)
7	CDEGBLJO_00006	tRNA	92				tRNA-Ser(gct)
8	CDEGBLJO_00007	tRNA	75				tRNA-Asp(gtc)
9	CDEGBLJO_00008	tRNA	73				tRNA-Lys(ttt)
10	CDEGBLJO_00009	tRNA	71				tRNA-Glu(ttc)
11	CDEGBLJO_00010	tRNA	74				tRNA-Pro(tgg)
12	CDEGBLJO_00011	CDS	573				hypothetical protein
13	CDEGBLJO_00012	CDS	228				hypothetical protein
14	CDEGBLJO_00013	CDS	738	rplB		COG0090	50S ribosomal protein L2
15	CDEGBLJO_00014	CDS	273				hypothetical protein
16	CDEGBLJO_00015	CDS	468				hypothetical protein
17	CDEGBLJO_00016	CDS	303	rplP		COG0197	50S ribosomal protein L16
18	CDEGBLJO_00017	CDS	225				hypothetical protein
19	CDEGBLJO_00018	CDS	201				hypothetical protein
20	CDEGBLJO_00019	CDS	387				hypothetical protein
21	CDEGBLJO_00020	CDS	507				hypothetical protein
22	CDEGBLJO_00021	CDS	720				hypothetical protein
23	CDEGBLJO_00022	CDS	276				hypothetical protein
24	CDEGBLJO_00023	CDS	399				hypothetical protein
25	CDEGBLJO_00024	CDS	369	rpsL			30S ribosomal protein S12
26	CDEGBLJO_00025	CDS	429				hypothetical protein
27	CDEGBLJO_00026	CDS	1233	tuf		COG0050	Elongation factor Tu
28	CDEGBLJO_00027	tRNA	72				tRNA-Phe(gaa)
29	CDEGBLJO_00028	tRNA	72				tRNA-Gln(ttg)
30	CDEGBLJO_00029	tRNA	71				tRNA-Gly(acc)
31	CDEGBLJO_00030	tRNA	73				tRNA-Trp(cca)
32	CDEGBLJO_00031	CDS	390				hypothetical protein
33	CDEGBLJO_00032	CDS	2304				hypothetical protein
34	CDEGBLJO_00033	tRNA	72				tRNA-Gly(tcc)
35	CDEGBLJO_00034	CDS	240				hypothetical protein
36	CDEGBLJO_00035	tRNA	89				tRNA-Ser(tga)
37	CDEGBLJO_00036	CDS	381				hypothetical protein
38	CDEGBLJO_00037	CDS	330				hypothetical protein
39	CDEGBLJO_00038	CDS	1308	rpoC2	2.7.7.6		DNA-directed RNA polymerase subunit beta'
40	CDEGBLJO_00039	CDS	1578	rpoC_1	2.7.7.6		DNA-directed RNA polymerase subunit beta'
41	CDEGBLJO_00040	CDS	1728	rpoC_2	2.7.7.6	COG0086	DNA-directed RNA polymerase subunit beta'
42	CDEGBLJO_00041	CDS	1593	rpoB	2.7.7.6	COG0085	DNA-directed RNA polymerase subunit beta

The genome annotation of the Plasmodium falciparum 3D7 apicoplast using Prokka successfully identified and categorized various genomic features, including coding sequences (CDS), rRNA, and tRNA regions. The generated files, such as GFF3, GenBank, and FASTA formats, provide comprehensive information for further analysis and functional characterization. This study demonstrates the utility of Prokka in automating and streamlining eukaryotic genome annotation, contributing valuable insights into the structural and functional components of the apicoplast genome. These results serve as a foundation for future research on the biological roles of annotated genes and their potential as drug targets or therapeutic pathways.

Experiment 3: Understanding data type of transcriptomics, proteomics and metabolomics

Aim: To explore the data types generated in transcriptomics, proteomics, and metabolomics, understand their characteristics, and relate them to biological insights.

Introduction:

Omics sciences—**transcriptomics, proteomics, and metabolomics**—are key fields in systems biology that study molecules at different biological levels. Each omics field generates unique data types due to differences in the molecules analyzed and the techniques used.

Explanation:

A. Data types of Transcriptomics

Transcriptomic data encompass a variety of RNA types and analysis methods, each providing unique insights into gene expression and regulation. One key data type is messenger RNA (mRNA), which reflects the coding transcripts translated into proteins. Studying mRNA levels reveals gene expression patterns, identifies differentially expressed genes, and sheds light on pathways active under specific conditions. Long non-coding RNAs (lncRNAs) represent another data type, comprising transcripts longer than 200 nucleotides that do not encode proteins but play regulatory roles in processes like chromatin remodeling and transcriptional regulation. These are often tissue-specific and are increasingly associated with diseases such as cancer. MicroRNAs (miRNAs), small non-coding RNAs (~22 nucleotides), regulate gene expression post-transcriptionally by targeting mRNAs for degradation or translational repression. They are key players in biological processes and are widely studied as potential biomarkers for diseases. Another emerging class of RNAs is circular RNAs (circRNAs), which are covalently closed loops formed by back-splicing. These molecules are stable, tissue-specific, and implicated in regulating miRNA activity, making them valuable for studying diseases and cellular regulation. Ribosomal RNA (rRNA) and small nuclear and nucleolar RNAs (snRNAs and snoRNAs) are structural and functional components of the cell's RNA machinery. While rRNA forms the backbone of ribosomes and is often depleted in transcriptomic studies, snRNAs and snoRNAs are essential for RNA splicing and modification, contributing to the cell's processing and translational efficiency. Beyond these, alternative splicing data focus on the diversity of transcript isoforms resulting from the inclusion or exclusion of specific exons or introns, highlighting tissue-specific or disease-associated variations in gene expression.RNA editing data investigate post-transcriptional modifications, such as A-to-I or C-to-U editing, which alter RNA sequences and potentially their functions. Single-cell RNA sequencing (scRNA-Seq) takes transcriptomics further by capturing gene expression at the resolution of individual cells, offering insights into cellular heterogeneity and identifying unique cell subpopulations. Similarly, spatial transcriptomics adds another layer of complexity by integrating gene expression data with spatial coordinates, enabling the study of tissue architecture and microenvironments, especially in contexts like tumor biology.

Collectively, these transcriptomic data types, analyzed through high-throughput sequencing technologies and computational tools, provide a comprehensive view of the transcriptional landscape, advancing our understanding of molecular biology, development, and disease processes.

Data Type	Key Methods	Tools for Analysis
mRNA	RNA-Seq, Microarrays	DESeq2, edgeR, limma
lncRNA	RNA-Seq	GENCODE, NONCODE, StringTie
miRNA	Small RNA-Seq	miRDeep, miRBase
circRNA	RNA-Seq	CIRCexplorer, CIRI
snRNA/snoRNA	RNA-Seq	snoStrip, RiboGalaxy
Alternative Splicing	RNA-Seq, Iso-Seq	Cufflinks, SUPPA2
scRNA-Seq	10x Genomics, Smart-seq	Seurat, Scanpy, Monocle
Spatial Transcriptomics	10x Visium	SpaceRanger, Giotto

B. Data types of Proteomics

Proteomic data encompasses a wide range of information about proteins, their properties, and interactions within biological systems. Broadly, proteomic data can be classified into several categories. Protein identification data focuses on determining which proteins are present in a sample, typically through mass spectrometry (MS) and database comparisons, providing insights into the composition of cells or tissues. Protein quantification data measures the abundance of proteins, either using label-free methods like spectral intensities or label-based techniques such as SILAC (Stable Isotope Labeling by Amino Acids in Cell Culture) or TMT (Tandem Mass Tags). This data is crucial for comparative studies, such as examining changes in protein levels in response to different conditions. Another key type is post-translational modification (PTM) data, which identifies chemical modifications on proteins (e.g., phosphorylation, glycosylation, or ubiquitination) that play critical roles in regulating protein function and signaling pathways. Protein interaction data, on the other hand, examines interactions between proteins or between proteins and other molecules using techniques like co-immunoprecipitation (Co-IP), yeast two-hybrid screening, or crosslinking-MS, enabling the mapping of protein interaction networks. Proteome dynamics data captures temporal and spatial changes in protein expression, turnover, or degradation, often using time-course experiments or pulse-chase labeling approaches, providing insights into cellular

responses to stimuli. In addition to these, protein structural data reveals three-dimensional protein structures through methods like cryo-electron microscopy (Cryo-EM) or X-ray crystallography, which are essential for understanding protein folding, stability, and interactions. Proteogenomics data integrates proteomics with genomic and transcriptomic analyses to validate genetic variants, identify novel proteins, and explore proteoforms, bridging the gap between genotype and phenotype. In clinical contexts, clinical proteomics data is particularly valuable for identifying protein biomarkers for disease diagnosis, prognosis, and treatment, with targeted approaches such as selected reaction monitoring (SRM) often used for validation.

For microbiome studies, metaproteomics data focuses on the collective protein composition of microbial communities in environments like the gut or soil, shedding light on microbial diversity and functional activity. Lastly, protein localization data determines the subcellular distribution of proteins, using organelle enrichment techniques or imaging to study spatial organization and compartmentalized protein functions. Together, these diverse proteomic data types provide a comprehensive understanding of proteins in health, disease, and various biological processes, often requiring integrative approaches to uncover meaningful insights.

Data Type	Focus	Key Techniques	Applications
Protein Identification	Protein presence in sample	MS, database matching	Biomarker discovery, protein profiling
Protein Quantification	Protein abundance	LFQ, SILAC, TMT	Comparative studies, drug response
PTM Analysis	Post-translational modifications	Enrichment, MS	Signaling pathways, disease research
Protein Interaction	Protein-protein and protein-ligand interactions	Co-IP, Y2H, XL-MS	Interaction networks, complex assembly
Proteome Dynamics	Temporal/spatial changes	Time-course MS, dynamic SILAC	Systems biology, stress response

Protein Structural Data	3D structures	Cryo-EM, X-ray crystallography	Drug design, structural biology
Proteogenomics	Genome-protein integration	MS, genomics	Functional genomics, personalized medicine
Clinical Proteomics	Biomarkers and disease targets	SRM, MRM, targeted MS	Cancer, diagnostics
Metaproteomics	Microbial community proteomes	Shotgun proteomics	Microbiome research
Protein Localization	Subcellular localization	Organelle MS, imaging	Cell biology, disease targeting

C. Data types of Metabolomics

Metabolomics data can be categorized into different types based on the nature of the analysis, quantification methods, and the scope of the study. First, analytical data types include spectral and chromatographic data. Spectral data, such as mass spectra from mass spectrometry (MS) or chemical shifts from nuclear magnetic resonance (NMR), provide raw or processed information about the metabolites, including peak intensities, mass-to-charge ratios (m/z), and retention times. Chromatographic data, on the other hand, include retention times and peak areas from techniques like gas chromatography (GC) or liquid chromatography (LC), which are crucial for separating and identifying complex mixtures of metabolites. In terms of quantification, metabolomics data can be classified as absolute, relative, or semi-quantitative. Absolute quantitative data provide precise concentrations of metabolites using calibration curves, while relative quantitative data report metabolite levels as relative abundances (e.g., fold changes) without absolute units. Semi-quantitative data estimate metabolite levels, often in high-throughput untargeted studies. Based on the scope of the analysis, metabolomics data can be either untargeted or targeted. Untargeted metabolomics is an exploratory approach that aims to detect all measurable metabolites in a sample, often resulting in high-dimensional data with potential for discovering new biomarkers. In contrast, targeted metabolomics focuses on a predefined set of metabolites, offering higher accuracy and sensitivity, typically for hypothesis-driven studies. Metabolomics data can also be characterized by specific molecular features, such as m/z ratios from MS, retention times from chromatography, chemical shifts from NMR, and spectral intensities that indicate metabolite abundance. These features are used to identify and quantify metabolites. Additionally, processed data types include peak tables, pathway mappings, statistical results, and metabolite annotations. Peak tables list

detected features (e.g., m/z, retention times, and intensities), while pathway data integrate metabolites into biological pathways to provide functional insights. Statistical analyses generate multivariate data, such as PCA scores or VIP scores, to highlight significant patterns. Contextual data is essential for interpreting metabolomics results. This includes experimental metadata, such as sample preparation methods and instrument settings, as well as biological metadata, like phenotypic information and treatment conditions. These contextual details ensure reproducibility and enable integration with other omics datasets. By organizing metabolomics data into these categories, researchers can better design experiments, analyze data, and interpret biological systems.

Data Type	Description	Examples
Spectral Data	Raw data from MS, NMR, etc.	Mass spectra, chemical shifts
Quantitative Data	Metabolite levels in absolute or relative terms	µM concentrations, fold changes
Scope-Based Data	Broad or focused metabolite profiling	Untargeted (all metabolites) vs. targeted (specific metabolites)
Molecular Feature Data	Specific characteristics of metabolites	m/z ratios, retention times, chemical shifts
Processed Data	Analyzed and interpreted information	Peak tables, pathway maps, statistical results
Contextual Data	Experimental and biological metadata	Sample type, instrument settings, phenotypic data

Conclusion:

Understanding the data types generated by transcriptomics, proteomics, and metabolomics is essential for integrating these datasets in systems biology. While transcriptomics focuses on RNA and gene expression, proteomics provides insights into protein abundance and modifications, and metabolomics sheds light on metabolic activities. Together, these fields offer a comprehensive view of cellular processes across multiple biological layers.

Experiment 4: Data cleaning and developing volcano plots

Aim: To Clean the RNASeq Data and Develop Volcano Plots.

Introduction:

One of the primary applications of RNA-Seq is the analysis of differentially expressed genes (DEGs). By comparing transcript abundance between conditions (e.g., treated vs. control), researchers can identify genes that are upregulated or downregulated. These DEGs often highlight biological pathways and molecular mechanisms driving specific phenotypic outcomes or responses.

To visualize RNA-Seq results effectively, volcano plots are commonly used. A volcano plot combines statistical significance (e.g., p-value) and biological relevance (e.g., fold change) into a single, easily interpretable graph. The x-axis represents the log2 fold change (magnitude of expression change), while the y-axis represents the -log10 of the p-value (statistical significance). Genes with significant upregulation or downregulation appear as points on either side of the plot, while insignificant changes cluster near the center. Volcano plots are invaluable in highlighting genes with both large expression changes and strong statistical support, aiding in the prioritization of targets for further study. This approach is widely used in biomedical research to uncover biomarkers, study disease mechanisms, and explore responses to treatments. Together, RNA-Seq data and volcano plots provide a robust framework for transcriptomic analyses and hypothesis generation in molecular biology.

Materials and Methods:

The RNA-Seq Dataset was obtained from Gene Expression Omnibus (GEO) database for accesion number GSE123456 (https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE123456)

The paper was titled “Gene Level Expression Profiling in human tongue squamous carcinoma cell line (SAS): gene expression profile in SAS cells transfected with scrambled control or C6orf141 mimic”

1. Loading the Dataset

[2]:
```
# Loading the Dataset
import pandas as pd

# Load the dataset
file_path = "/home/harman/Documents/Sem3/Genome analysis/test_rna_seq_data.csv"
df = pd.read_csv(file_path)

# Preview the data
print(df.head())
```

```
                    ID  adj.P.Val   p-value      t     B  log2FoldChange  \
0     TC0600008212.hg.1          1  0.000009 -10.60 -4.09           -3.46
1  HTA2-neg-47419781_st          1  0.000077  -7.78 -4.12           -2.73
2  HTA2-neg-47420247_st          1  0.010962  -3.36 -4.31           -2.35
3     TC0200008433.hg.1          1  0.016518  -3.08 -4.33           -1.95
4     TC0500008271.hg.1          1  0.000739  -5.49 -4.18           -1.88

                              SPOT_ID
0                  NM_001145652 // RefSeq
1                      --normgene->intron
2                      --normgene->intron
3              T193910 // miTranscriptome
4  skoysloby.aAug10-unspliced // Ace View
```

2. Data Cleaning

[3]:
```
# Checking for Missing Values

print(df.isnull().sum())  # Check for missing values
df = df.dropna()          # Drop rows with missing values if any
```

```
ID                0
adj.P.Val         0
p-value           0
t                 0
B                 0
log2FoldChange    0
SPOT_ID           0
dtype: int64
```

[4]:
```
# Ensure necessary columns exist: Confirm that the dataset contains␣
↪log2FoldChange and p-value columns:

if 'log2FoldChange' not in df.columns or 'p-value' not in df.columns:
    raise ValueError("Required columns (log2FoldChange, p-value) are missing!")
```

[5]:
```
# Remove Duplicates

df = df.drop_duplicates()
```

[6]:
```
# Check for extreme values or outliers

print(df.describe())  # Summary statistics
```

```
       adj.P.Val        p-value              t              B  log2FoldChange
count   138745.0  138745.000000  138745.000000  138745.000000   138745.000000
mean         1.0       0.517037       0.026937      -4.604747        0.006581
std          0.0       0.285580       1.092457       0.078054        0.290315
min          1.0       0.000009     -10.600000      -4.670000       -3.460000
25%          1.0       0.272539      -0.646000      -4.660000       -0.167000
50%          1.0       0.523087       0.024000      -4.640000        0.006120
75%          1.0       0.765648       0.694000      -4.580000        0.179000
```

```
max          1.0        1.000000        9.290000       -4.090000        2.700000
```

3. Preparation of Data for Visualisation

[7]:
```
# Adding Derived Columns

import numpy as np
df['neg_log10_pval'] = -np.log10(df['p-value'])
```

[8]:
```
# Setting Dynamic Threshhold

fold_change_threshold = 0.6
p_value_threshold = 0.05
```

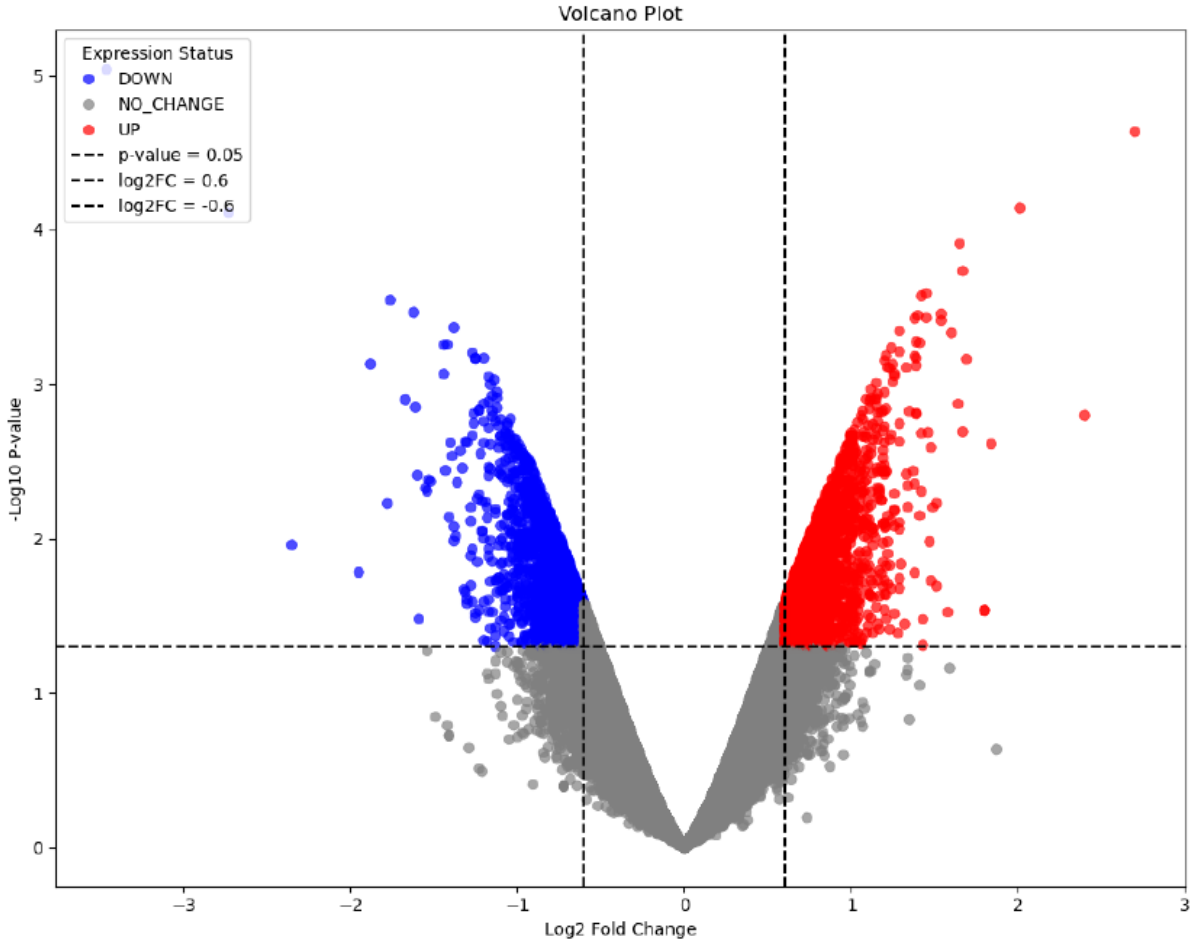
4. Generation of Volcano Plots

[9]:
```
# Check if Derived columns exist

print(df.columns)
print(df.head())
```

```
Index(['ID', 'adj.P.Val', 'p-value', 't', 'B', 'log2FoldChange', 'SPOT_ID',
       'neg_log10_pval'],
      dtype='object')
                     ID  adj.P.Val   p-value      t     B  log2FoldChange  \
0     TC0600008212.hg.1          1  0.000009 -10.60 -4.09           -3.46
1  HTA2-neg-47419781_st          1  0.000077  -7.78 -4.12           -2.73
2  HTA2-neg-47420247_st          1  0.010962  -3.36 -4.31           -2.35
3     TC0200008433.hg.1          1  0.016518  -3.08 -4.33           -1.95
4     TC0500008271.hg.1          1  0.000739  -5.49 -4.18           -1.88

                                SPOT_ID  neg_log10_pval
0                NM_001145652 // RefSeq        5.040005
1                    --normgene->intron        4.110810
2                    --normgene->intron        1.960092
3            T193910 // miTranscriptome        1.782051
4  skoysloby.aAug10-unspliced // Ace View        3.131097
```

[10]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Ensure categorization column exists
if 'expression' not in df.columns:
    df['expression'] = np.where(
        (df['log2FoldChange'] > fold_change_threshold) & (df['p-value'] <␣
 ↪p_value_threshold), 'UP',
        np.where(
            (df['log2FoldChange'] < -fold_change_threshold) & (df['p-value'] <␣
 ↪p_value_threshold), 'DOWN',
            'NO_CHANGE'
        )
    )

# Volcano Plot
plt.figure(figsize=(10, 8))
sns.scatterplot(
    data=df,
    x='log2FoldChange',
    y='neg_log10_pval',
    hue='expression',
    palette={"UP": "red", "DOWN": "blue", "NO_CHANGE": "grey"},
    alpha=0.7,
    edgecolor=None
)

# Add threshold lines
plt.axhline(y=-np.log10(p_value_threshold), color='black', linestyle='--',␣
 ↪label=f'p-value = {p_value_threshold}')
plt.axvline(x=fold_change_threshold, color='black', linestyle='--',␣
 ↪label=f'log2FC = {fold_change_threshold}')
plt.axvline(x=-fold_change_threshold, color='black', linestyle='--',␣
 ↪label=f'log2FC = {-fold_change_threshold}')

# Customize plot
plt.title("Volcano Plot")
plt.xlabel("Log2 Fold Change")
plt.ylabel("-Log10 P-value")
plt.legend(title="Expression Status", loc='upper left')
plt.tight_layout()
plt.show()
```

Result:



The RNA-Seq analysis identified genes with significant expression changes: Genes with **log2 fold change > 0.6** and **p-value < 0.05** were categorized as upregulated. Genes with **log2 fold change < -0.6** and **p-value < 0.05** were classified as downregulated. All other genes were considered to show no significant expression change. The volcano plot revealed a clear distinction between differentially expressed genes: **Upregulated genes** (red) and **downregulated genes** (blue) were visible on opposite sides of the x-axis. Genes near the center represented no significant change in expression. Threshold lines highlighted the significance cutoffs, making it easy to visualize the most critical genes for further study. The genes identified in the upregulated category may be associated with the molecular mechanisms induced by **C6orf141 mimic**. Downregulated genes could represent pathways suppressed during this condition. Future functional validation can confirm the role of these DEGs in tumor biology or response to treatments.

Conclusion:

The RNA-Seq analysis successfully identified differentially expressed genes in the SAS cell line transfected with C6orf141 mimic compared to a scrambled control. The use of volcano plots provided a clear and intuitive visualization of significant genes, aiding in the prioritization of potential targets for further functional and pathway analysis. This study underscores the utility of RNA-Seq and volcano plots in transcriptomic research, paving the way for deeper insights into molecular mechanisms driving phenotypic changes in cancer biology.

Experiment 5: Data analysis of RNA-seq, ChIP-seq, and Proteome-MS

Data Analysis of RNA-seq Data and ChIp-seq Data using Geo2R. Both RNA-Seq and ChIp-Seq can be analysed using the Geo2R tool. The workflow remains the same.

Aim: To analyse the RNASeq Data using Geo2R tools and GEO database

Introduction:

RNA sequencing (RNA-Seq) is a widely used technique for transcriptome profiling, enabling the analysis of gene expression patterns in various biological conditions. The Gene Expression Omnibus (GEO) database serves as a valuable resource for accessing publicly available RNA-Seq datasets. GEO2R, a web-based tool provided by NCBI, offers an accessible platform for analyzing these datasets without requiring advanced programming skills. It is particularly useful for identifying differentially expressed genes (DEGs) between experimental groups.

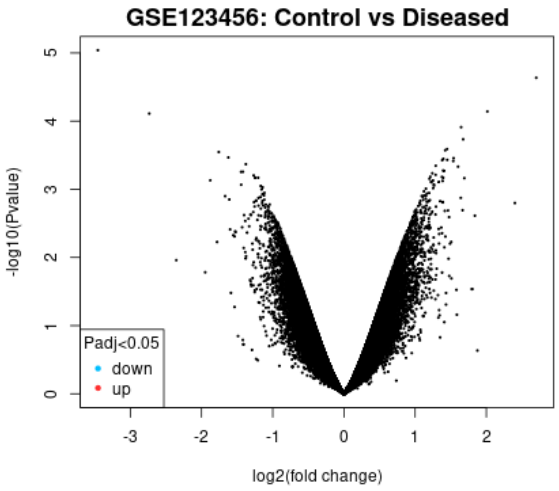
GEO2R enables users to perform statistical comparisons of expression levels between conditions using pre-processed data. The tool utilizes the Limma (Linear Models for Microarray and RNA-Seq Data) package to calculate log2 fold changes and p-values, which help in identifying upregulated and downregulated genes. Results are provided in a tabular format, which can be visualized further using tools like volcano plots or heatmaps for better interpretation. By leveraging GEO2R, researchers can efficiently extract meaningful insights from RNA-Seq data, such as identifying gene expression changes associated with diseases, treatments, or experimental conditions. This user-friendly approach democratizes transcriptomic data analysis, making it accessible to a broader audience, including those with limited computational expertise. It serves as a foundation for hypothesis generation and further experimental validation.

Materials and Methods:

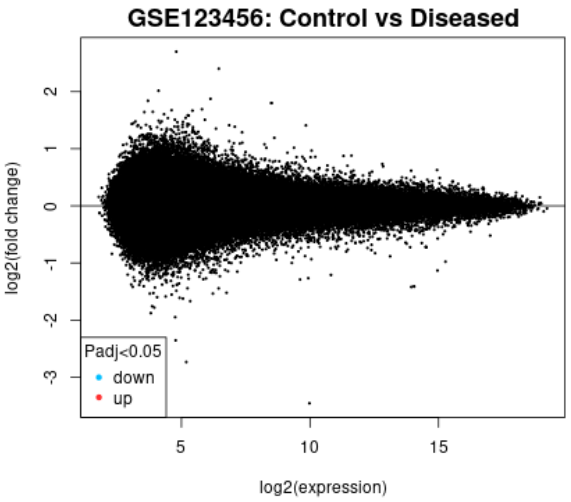
The RNA-Seq Dataset was obtained from Gene Expression Omnibus (GEO) database for accession number GSE123456 (https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE123456)
The paper was titled “Gene Level Expression Profiling in human tongue squamous carcinoma cell line (SAS): gene expression profile in SAS cells transfected with scrambled control or C6orf141 mimic”

1. Searched ‘GSE123456’ in GEO database
2. Click on Analyse with GEO2R tool option
3. Analyse the results

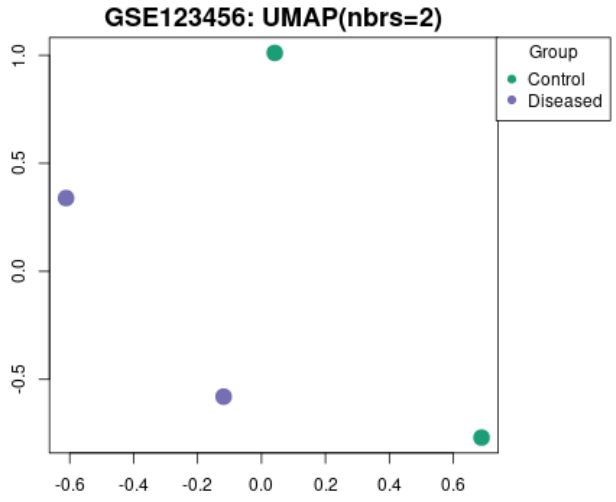
Results:



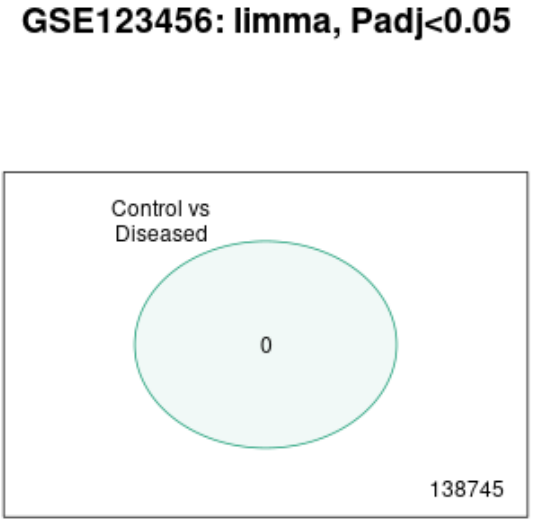
Volcano Plot



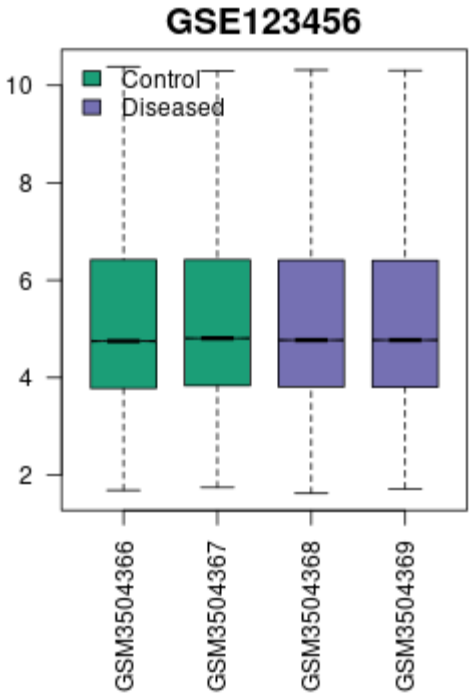
Mean-Difference Plot



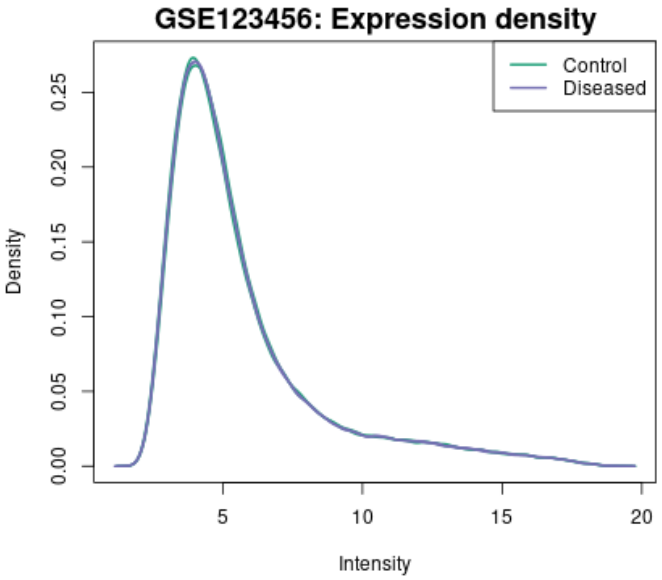
UMAP Plot



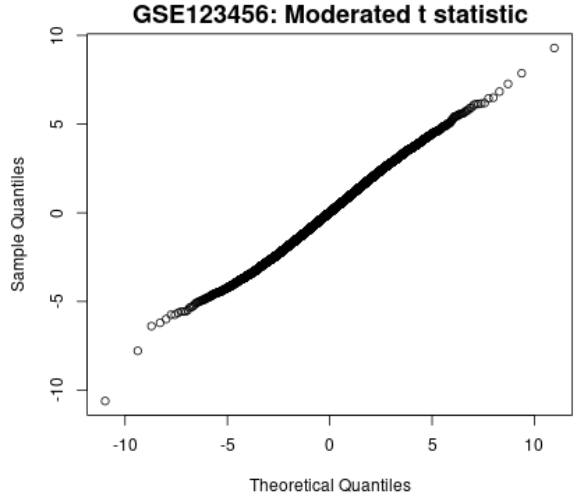
Venn Diagram



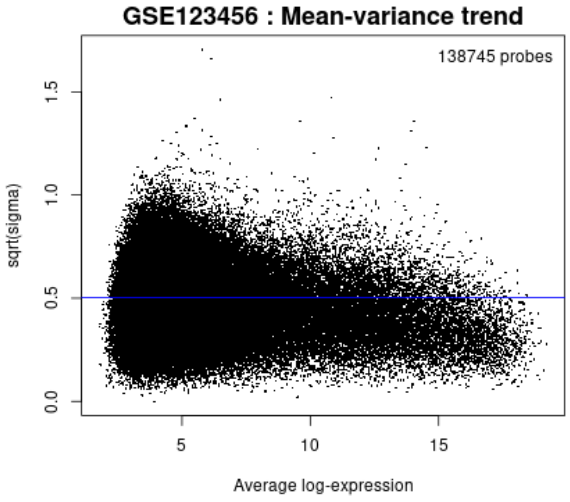
Box Plot



Expression Density Plot



t-statistic quantile quantile plot



Mean Variance trend plot

The RNA-Seq data analysis using GEO2R identified genes with significant changes in expression between the experimental groups (SAS cells transfected with scrambled control vs. C6orf141 mimic). **Upregulated genes** (red points in the volcano plot) were distinctly visible on one side of the x-axis, indicating genes with a positive fold change and high significance. **Downregulated genes** (blue points) were observed on the opposite side, showing negative fold changes and strong significance. Genes near the center of the volcano plot represented no significant change in expression. Upregulated genes likely correspond to pathways or mechanisms activated by the C6orf141 mimic. Downregulated genes might represent processes suppressed during the experimental condition. These

findings provide a strong basis for further functional studies to understand the role of C6orf141 in tumor biology.

- Volcano Plot: Highlighted significant DEGs with clear upregulated and downregulated categories.
- Box Plot: Showed the normalized expression distributions across samples.
- Expression Density Plot: Indicated uniform distribution of expression values after normalization.
- UMAP Plot: Represented clustering of samples based on transcriptomic profiles, confirming clear separation between experimental conditions.
- Venn Diagram: Overlap of DEGs across conditions or datasets, if applicable.
- Mean-Difference Plot (MA Plot): Showed log2 fold changes against average expression values, confirming data trends.
- t-Statistic Q-Q Plot: Validated the statistical assumptions of the test.
- Mean Variance Trend Plot: Highlighted the variance across expression levels.

Conclusion:

The analysis of RNA-Seq data using GEO2R successfully identified differentially expressed genes in the SAS cell line transfected with the C6orf141 mimic compared to the scrambled control. GEO2R, coupled with visualizations such as volcano plots and UMAP clustering, provided a clear distinction between significant DEGs and non-significant genes. These results highlight potential genes and pathways influenced by C6orf141, offering new insights into its role in gene regulation and cancer biology. Further experimental validation is essential to confirm the molecular mechanisms and explore therapeutic implications.