

Name: Ms. Prarthi Hrishit Kothari

Class: M. Sc. Bioinformatics (Part I)

Roll Number: 115

Course: M. Sc. Bioinformatics

Department: Department of Bioinformatics

Paper: Elective Paper

**Paper Name
and Code:** Genomics & Proteomics in
Bioinformatics, NGS
(GNKPSBI3P502)

**Academic
Year:** 2023-24



SGCP's
Guru Nanak Khalsa College
of Arts, Science & Commerce (Autonomous)

DEPARTMENT OF BIOINFORMATICS

CERTIFICATE

This is to certify that Ms. Prarthi Hrishit Kothari (Roll No: 115) of M. Sc. Bioinformatics (Part I) has satisfactorily completed the practical for Elective Paper: Genomics & Proteomics in Bioinformatics, NGS (GNKPSBI3P502) for Semester II course prescribed by the University of Mumbai during the academic year 2023-2024.

Teacher-in-Charge
(Mrs. Aparna Patil Kose)

Head of Department
(Dr. Gursimran Kaur Uppal)

External Examiner

INDEX

SR. NO.	WEBLEM	DATE	PAGE NO.	SIGN
1.	Introduction to Gene Prediction and various elements in Prokaryotes and Eukaryotes using various tool on Softberry server	06/03/2024	001	
1(A)	To predict the gene structure for query ‘CDK1’ (GenBank ID: NM_001320918.1) using HMM-based algorithm via FGENESH tool.	06/03/2024	006	
1(B)	To predict bacterial operon and gene for query ‘ <i>Sulfobacillus thermosulfidooxidans</i> ’ (GenBank ID: NZ_MDZD01000033.1) using FGENESB tool.	06/03/2024	014	
1(C)	To predict bacterial promoter for query ‘ <i>Sulfobacillus thermosulfidooxidans</i> ’ (GenBank ID: NZ_MDZD01000033.1) using BPROM tool.	06/03/2024	022	
1(D)	To predict human promoter for query ‘JAK2’ (GenBank ID: NM_001322195) using FPROM tool.	06/03/2024	030	
1(E)	To predict plant promoters for query ‘Glutenin’ (GenBank ID: X03346.1) using TSSP tool.	06/03/2024	037	
2.	Introduction Microarray & GEO Database	07/03/2024	044	
2(A)	To analyze \gene expression data from the Gene Expression Omnibus (GEO) database to identify differentially expressed genes associated with ‘Bladder tumor’.	07/03/2024	047	

3.	Introduction GPCR Database	17/03/2024	058	
3(A)	To explore the structural and functional characteristics of query 'PTH1' using GPCR database.	17/03/2024	064	
4.	Introduction to EST Database	16/03/2024	070	
4(A)	To identify and characterize gene expression patterns for query 'DDIT3' (Accession ID: DN990078.1) using an EST database.	16/03/2024	073	
5.	Introduction to BLAST2GO: A tool for annotation, visualization and analysis in functional genomics research	15/03/2024	078	
5(A)	To annotate and analyze genomic or transcriptomic sequences for query 'LEPR' to understand their functional significance by using BLAST2GO tool.	15/03/2024	084	
6.	Introduction to SNP Database (dbSNP)	16/03/2024	100	
6(A)	To identify and analyze genetic variations (mutational gene) in the query 'DDIT3' (Reference SNP Report: rs28382352) using dbSNP (SNP database).	16/03/2024	103	
7.	Introduction to Whole Genome Sequencing	18/03/2024	111	
7(A)	To compare genomic sequences, identify conserved regions, and visualize sequence alignments for query 'ADD1' (Accession ID: NM_001354759.2 and NM_001354756.2) using PipMaker Tool.	18/03/2024	115	
7(B)	To perform comparative genomics analysis to identify conserved regions and its functional elements for query 'ADD1' (Accession ID: NM_001354759.2) using VISTA tool.	18/03/2024	123	

8.	Introduction to Identification of repetitive elements by using Repeat Masker Tool	20/03/2024	132	
8(A)	To mask repeats using RepeatMasker tool on dataset imported from Zenodo.	20/03/2024	136	
9.	Introduction to Next Generation Sequencing	21/03/2024	149	
9 (A)	To effectively manage, analyze, and interpret Next-Generation Sequencing (NGS) data from databases using appropriate file formats.	23/03/2024	164	
9 (B)	To remove adapters, trim low-quality bases, and filter out short or low-quality reads from raw sequencing data using Trimmomatic for data preprocessing.	10/04/2024	174	
9 (C)	To assess the quality of raw sequencing data and identify potential issues or biases using FASTQC for quality control.	04/04/2024	181	
9 (D)	To map sequencing reads to a reference genome and generate aligned BAM/SAM files using Bowtie2 for downstream analysis.	10/04/2024	194	
9 (E)	To perform de novo assembly of sequencing reads into contigs using Velvet for genome assembly.	13/04/2024	200	
9 (F)	To assess the quality of a genome assembly using QUAST for comprehensive evaluation and comparison of assembly metrics.	21/04/2024	206	

DATE: 06/03/2024

WEBLEM 1

Introduction to Softberry Server

(URL: <http://www.softberry.com/>)

The Softberry server is a collection of software tools for genomic research that focuses on computational methods for high-throughput biomedical data analysis. Softberry provides services such as genome and transcript assembling, reads mapping, alternative transcripts, SNP discovery, bacterial gene identification, metagenomics, microbiome sequence analysis, gene identification using HMM FGENESH gene finder, genome annotation pipeline, RNA folding, miRNA identification, and more. The company works closely with clients to meet their computational genomics needs and offers custom genome annotation services, data set customization for gene prediction programs, expression data analysis tools, software customization, and more. Softberry is a leading developer of software tools for genomic research with a focus on computational methods for high-throughput biomedical data analysis. Their software supports next-generation sequencing technologies, transcriptome analysis, SNP detection, and disease-specific SNP subsets selection. The company's programs have been widely used in numerous research projects and publications.

Hundreds software applications developed in Softberry can be executed on-line at www.softberry.com or downloaded for free academic usage. Softberry started its commercial activities on May 1997 and is headquartered in Mount Kisco, NY, USA, with a satellite office in Novosibirsk, Russia. It participated in a number of academic collaborations on various research projects around the world.

Softberry's bioinformatics tools cover a wide range of functionalities beyond promoter prediction, including genome and transcript assembly, reads mapping, alternative transcript analysis, SNP discovery, metagenomics, microbiome sequence analysis, gene identification, RNA folding, miRNA target prediction, protein structure analysis, molecular docking, and much more. These tools are essential for comprehensive genomic analysis and functional annotation across various biological contexts. Softberry Programs available to academic users at no charge for occasional use in research projects. Below are some programs offered by Softberry.

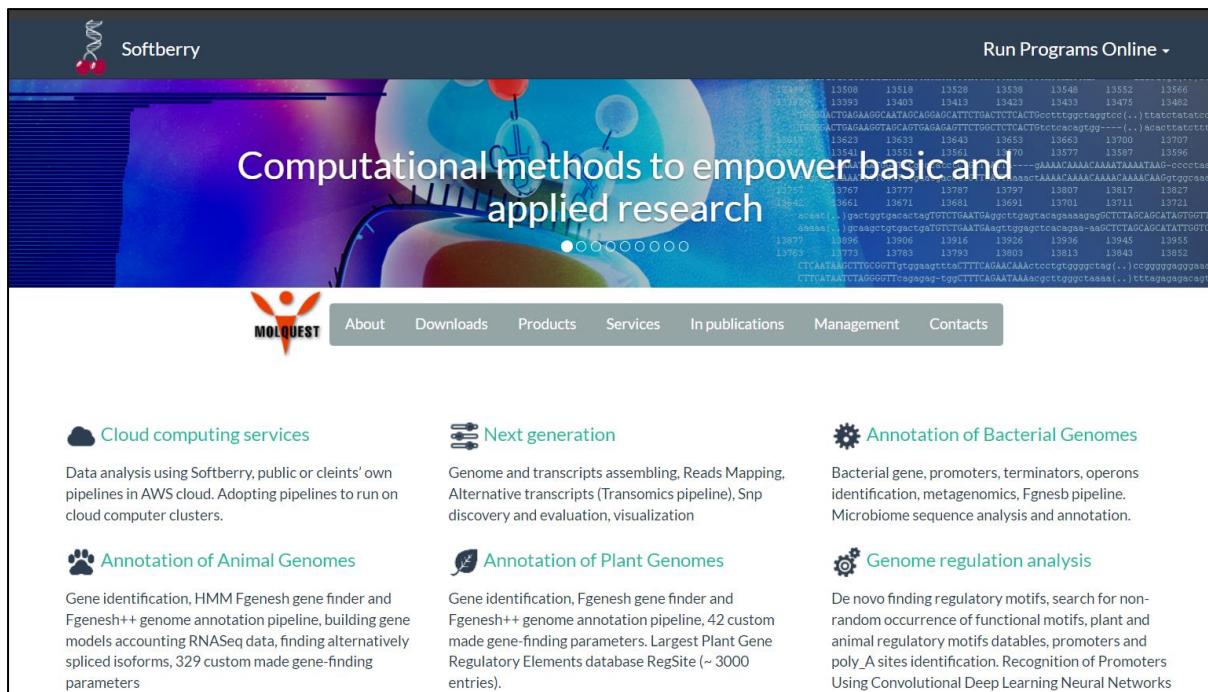


Fig 1: Homepage of Softberry server

Program 1: Gene Finding in Eukaryota

FGENESH - HMM-based gene structure prediction (multiple genes, both chains)

FGENESH is a highly accurate and fast gene prediction tool for eukaryotic genomes, outperforming other popular programs like GenScan and HMMGene in terms of sensitivity and specificity. Its efficacy has been notably demonstrated in rice genome sequencing projects, where it contributed significantly to the identification of high-confidence predicted genes. The comprehensive FGENESH++ package incorporates FGENESH and FGENESH+ programs, leveraging hidden Markov models (HMMs) and protein similarity to automate gene prediction in diverse eukaryotic genomes. This multifaceted tool enables users to predict genes through homology to known proteins, map existing mRNAs onto genomic sequences, and conduct ab-initio gene prediction in regions devoid of mapped mRNAs or homologous proteins.

In addition to its core functionalities, FGENESH offers a pseudogene annotation program (PSF) capable of identifying potential pseudogenes by analyzing exon-intron structures of annotated genes or known proteins. This feature enhances the comprehensiveness of gene discovery processes in eukaryotic genomes. Designed for versatility, FGENESH supports various organisms including humans, mice, *Drosophila*, nematodes, dicot plants, monocot plants, yeast (*S. pombe*), and *Neurospora*. Its adaptability across different species provides researchers with a robust platform for gene prediction and analysis.

The superiority of FGENESH is underscored by its performance in comparative evaluations. For instance, in a study assessing gene prediction accuracy in maize, FGENESH emerged as the top performer, outperforming other ab initio programs by a significant margin. Notably, it identified 11% more correct gene models compared to its closest competitor, GeneMark, on a set of 1353 test genes. FGENESH stands as a premier gene prediction tool renowned for its precision, speed, and adaptability across a wide range of eukaryotic genomes. Its advanced

features, including pseudogene annotation capabilities, make it an indispensable asset for researchers in the field of genomics and molecular biology.

Program 2: Operon and Gene Finding in Bacteria

A. FGENESB - Pattern/Markov chain-based bacterial operon and gene prediction:

The FGENESB program by Softberry is a suite of bacterial operon and gene finding programs known for its accuracy in prokaryotic gene prediction. It offers features such as automatic annotation of predicted genes by homology with COG and NR databases, prediction of promoters and terminators, and operon prediction based on distances between ORFs and frequencies of different genes neighboring each other. The FGENESB package includes options to work with a set of sequences like scaffolds of bacterial genomes or short sequencing data, making it a valuable tool for community sequence annotation.

The program follows a series of steps for gene prediction, including finding potential ribosomal RNA genes, predicting tRNA genes, and making initial predictions of long ORFs used as a starting point for calculating parameters for gene prediction. FGENESB utilizes highly accurate Markov chains-based gene prediction algorithms, based on coding regions and translation and termination sites, ensuring precise gene predictions in bacterial genomes. FGENESB automatically trains gene finding parameters for new bacterial genomes using genomic DNA as input, with the option to use pre-learned parameters from related organisms. FGENESB includes operon prediction based on distances between Open Reading Frames (ORFs) and frequencies of different genes neighboring each other, enhancing the understanding of gene organization in bacterial genomes.

FGENESB has been used in various projects and studies, including the first-ever published bacterial community annotation project by Tyson et al. in Nature 2004, showcasing its effectiveness in genome annotation and gene prediction tasks. The program's accuracy in gene prediction has been compared with other popular gene prediction programs like Glimmer and GeneMarkS, demonstrating high levels of accuracy in prokaryotic gene prediction. The final annotation can be presented in GeneBank format to be readable by visualization software such as Artemis or Softberry Bacterial Genome Explorer.

FGENESB from Softberry is a powerful tool for automatic annotation of bacterial genomes, offering accurate gene predictions, operon predictions, and various features that contribute to comprehensive genome annotation and analysis.

B. BPROM - Prediction of bacterial promoters

The BPROM program by Softberry is a bacterial promoter prediction tool that predicts potential transcription start positions of bacterial genes regulated by sigma70 promoters, a major E. coli promoter class. BPROM is bacterial sigma70 promoter recognition program with about 80% accuracy and specificity. It is best used in regions immediately upstream from ORF start for improved gene and operon prediction in bacteria. It uses a linear discriminant function (LDF) to combine characteristics describing functional motifs and oligonucleotide composition of these sites, achieving an accuracy of about 80% in E. coli promoter recognition. It is recommended to run BPROM on a region between two neighboring ORFs located on the same

strand or on a sequence upstream from an ORF, as most promoters are located within 150 bp from the protein coding sequence.

B PROM is part of the bacterial genome analysis suite of programs and complements operon and gene prediction by the FGENESB program. FGENESB is an automatic annotation tool for bacterial genomes that includes features like gene prediction, mapping of tRNA and rRNA genes, prediction of promoters and terminators, and operon prediction based on distances between predicted genes. The FGENESB package is known for its high accuracy in prokaryotic gene prediction and operon refinement using predicted promoters and terminators as additional evidence. B PROM is a specialized tool for predicting bacterial promoters, while FGENESB offers a comprehensive solution for gene prediction, operon identification, and annotation in bacterial genomes.

Program 3: Search for promoters/functional motifs

A. FPROM - Human promoter prediction

The FPROM program by Softberry is a bioinformatics tool specifically designed for human promoter prediction and the search for functional motifs within the human genome. This program is part of Softberry's suite of bioinformatics tools that aim to identify genes, functional signals, decipher gene expression data, and select disease-specific genes and drug target candidates. FPROM plays a crucial role in understanding gene regulation, transcriptional control, and gene expression patterns in humans.

FPROM focuses on predicting human promoters and functional motifs, providing valuable insights into the regulatory elements that govern gene expression in the human genome. It complements other tools within the Softberry suite, such as FGENESH for gene structure prediction and FGENESB for operon and gene finding in bacteria. By utilizing FPROM, researchers can gain a deeper understanding of the mechanisms underlying gene regulation in humans. FPROM is a specialized program developed by Softberry for human promoter prediction and the search for functional motifs within the human genome. It forms part of a comprehensive suite of bioinformatics tools that facilitate advanced genomic analysis and interpretation in the context of gene regulation and expression in humans.

B. TSSP - Prediction of PLANT Promoters

The TSSP program by Softberry is a tool designed for predicting potential transcription start positions in DNA sequences. It utilizes a linear discriminant function to combine characteristics and predict these positions. TSSP uses a file with selected factor binding sites from the RegSite DB (Plants) developed by Softberry Inc.

The TSSP program by Softberry has shown a high level of accuracy in predicting promoter regions and functional motifs. The TSSG program, which is closely related to TSSP, has been reported to achieve approximately 50-55% accuracy in recognizing true promoter regions with one false positive prediction for about 5000 bp. Additionally, TSSG is highlighted as the most accurate mammalian promoter prediction program, with significantly fewer false positive predictions compared to other programs like PROSCAN1.7 and NNPP2.0.

While specific accuracy metrics for TSSP are not explicitly mentioned in the search results provided, the overall performance of Softberry's promoter prediction programs, including

TSSP, is notable. These tools are designed to predict transcription start positions and functional motifs effectively, making them valuable resources for bioinformatics analysis, particularly in plant gene research.

The algorithm predicts transcription start positions and functional motifs, providing information on the first nucleotide of the transcript, TATA box positions, and functional motifs for each predicted region. TSSP is specifically used for searching for plant RNA Polymerase II TATA and TATA-less promoters (transcription start sites, TSSs). It takes single or multiple sequences in FASTA format as input, with an allowed sequence length of 251 - 100,000 bp. The output file is in classic Text format and can run on Linux or Unix systems. The program requires setting up an environmental variable and offers various options for customization such as selecting search criteria, neural network thresholds for promoters, and more. TSSP program is a valuable tool for bioinformatics analysis, particularly in predicting transcription start positions and functional motifs in DNA sequences, with a specific focus on plant RNA Polymerase II promoters.

REFERENCES:

1. Solovyev, V., Kosarev, P., Seledsov, I., & Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome biology*, 7 Suppl 1(Suppl 1), S10.1–S10.12. <https://doi.org/10.1186/gb-2006-7-s1-s10>
2. Softberry - FGENESB HELP. (n.d.). <http://www.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb>
3. Robert W. Li, pp., V. S. and A. S. (n.d.). AUTOMATIC ANNOTATION OF MICROBIAL GENOMES AND METAGENOMIC SEQUENCES. Researchgate. https://www.researchgate.net/publication/259450599_V_Solovyev_A_Salamov_2011_Automatic_Annotation_of_Microbial_Genomes_and_Metagenomic_Sequences_In_Metagenomics_and_its_Applications_in_Agriculture_Biomedicine_and_Environmental_Studies_Ed_RW_Li_Nova_Sc
4. Bacterial Promoter, Operon and Gene Finding. (n.d.). [Www.softberry.com](http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfindb). Retrieved March 10, 2024. <http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfindb>
5. Solovyev, V., & Salamov, A. (n.d.). The Gene-Finder computer tools for analysis of human and model organism's genome sequences. Retrieved March 10, 2024. <https://cdn.aaai.org/ISMB/1997/ISMB97-045.pdf>
6. Softberry - FPROM HELP. (n.d.). <http://www.softberry.com/berry.phtml?topic=fprom&group=help&subgroup=promoter>
7. Softberry - about. (n.d.). http://www.softberry.com/berry.phtml?topic=about&no_menu=on

DATE: 06/03/2024

WEBLEM 1(A)

Gene Prediction in Eukaryotes: FGENESH

(URL:<http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfind>)

AIM:

To predict the gene structure for query ‘CDK1’ (GenBank ID: NM_001320918.1) using HMM-based algorithm via FGENESH tool.

INTRODUCTION:

The FGENESH program by Softberry is a gene prediction tool used to predict multiple genes in genomic DNA sequences. It is known for its speed and accuracy, being 50-100 times faster than GenScan and highly precise in gene prediction. FGENESH has been recognized for its accuracy in various studies, outperforming other gene prediction programs like GenScan and HMMGene. It is widely used in different genomes including human, mouse, Drosophila, nematode, dicot plants, monocot plants, yeast, and Neurospora. The program offers features like predicting potential genes in genomic DNA, transcription start sites, coding sequences, and more. FGENESH++ is an advanced version of the FGENESH program that includes ab initio gene prediction using the FGENESH algorithm and additional steps for gene prediction. It is a pipeline for automatic gene prediction in eukaryotic genomes without human intervention. The FGENESH-C program is designed for predicting multiple genes in genomic DNA sequences using HMM gene models and similarity with known mRNA/EST sequences. It emphasizes the importance of using homologous mRNA information to improve the accuracy of gene finding.

FGENESH is the HMM-based gene-finding program with the algorithm similar to Genie (Kulp et al. 1996) and GENSCAN (Burge and Karlin 1997). The difference between FGENESH and analogous programs is that in the model of gene structure a signal term (such as splice site or start site score) has some advantage over a content term (such as coding potentials), reflecting the biological significance of the signals. Parameters of the program were trained on 1600 *D. melanogaster* entries from GenBank. The run time of FGENESH is approximately linear.

CDK1:

CDK1, a cyclin-dependent kinase, holds pivotal significance in cell cycle regulation, particularly governing the G2/M phase transition. Its heightened expression in various tumors signifies its involvement in tumorigenesis, implicating CDK1 as a potential target for cancer therapy. Pan-cancer analyses have illuminated CDK1's association with oncogenic signatures, immune cell infiltration, and survival rates across multiple cancers, suggesting its promise as a target for cancer immunotherapy. Additionally, CDK1's role extends beyond cell cycle regulation, encompassing the modulation of RNA polymerase activity, which may influence protein synthesis and cell cycle entry. In colorectal cancer, CDK1 emerges as a critical factor in oxaliplatin resistance, presenting a potential therapeutic target to overcome drug resistance.

and inhibit tumor progression. Moreover, CDK1's predictive value in immune checkpoint inhibitor therapy response and its diverse impact on tumor immunity underscore its multifaceted role in tumor progression and treatment response across various cancer types.

METHODOLOGY:

1. Visit the Softberry Server, homepage. URL: <http://www.softberry.com/>
2. Click on "Run Programs Online" and select "Gene Finding in Eukaryota" from the available programs.
3. Choose "FGENESH - HMM-based gene structure prediction (multiple genes, both chains)" as the program.
4. Paste the sequence to the window for the query 'CDK1' gene (GenBank ID: NM_001320918.1) provide the nucleotide sequence in FASTA format, obtained from GenBank.
5. If the organism's name is given in your study, select it from the options provided.
6. Click on "Search" to initiate the analysis.
7. The tool will display the prediction of potential genes in the genomic DNA.

OBSERVATIONS:

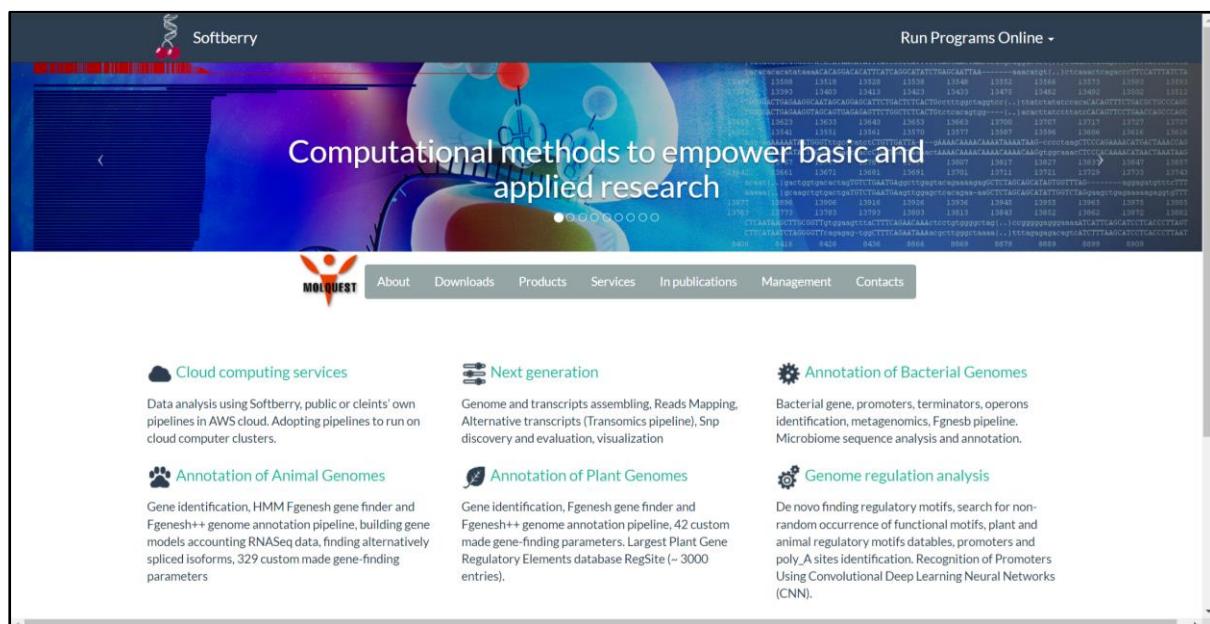


Fig 1: Homepage of Softberry server

The screenshot shows the Softberry website interface. At the top right, there is a dropdown menu labeled "Run Programs Online ▾". Below it, a sub-menu titled "Gene finding in Eukaryota" is expanded, listing various options such as "Gene finding with similarity", "Operon and Gene Finding in Bacteria", "Gene Finding in Viral Genomes", etc. The main content area features a banner with the text "Annotation of Bacterial Genomes" and a logo for "MOLQUEST". Below the banner, there are several sections with icons and descriptions:

- Cloud computing services**: Data analysis using Softberry, public or clients' own pipelines in AWS cloud. Adopting pipelines to run on cloud computer clusters.
- Annotation of Animal Genomes**: Gene identification, HMM Fgenesh gene finder and Fgenesh++ genome annotation pipeline, building gene models accounting RNASeq data, finding alternatively spliced isoforms, 329 custom made gene-finding parameters.
- Next generation**: Genome and transcripts assembling, Reads Mapping, Alternative transcripts (Transomics pipeline), Snp discovery and evaluation, visualization.
- Annotation of Plant Genomes**: Gene identification, Fgenesh gene finder and Fgenesh++ genome annotation pipeline, 42 custom made gene-finding parameters. Largest Plant Gene Regulatory Elements database RegSite (~ 3000 entries).
- Analysis**: Bacterial identification, Microbiology.
- Genetics**: De novo random or animal repeats, poly_A site, Using Cor.
- Proteomics**: Manipulations with sequences, Multiple alignments, Synteny from genome contigs, Analysis of gene expression data, Plant Promoter Database, Repeats, SNP.
- Gene identification pipelines**: Gene identification ninelines.

Fig 2: Select the program ‘Gene Finding in Eukaryota’

The screenshot shows the "Services Test Online" page. On the left, there is a sidebar with a "Home" button and a list of program categories. The "Gene finding in Eukaryota" category is highlighted with a red border. The main content area displays information about the FGENESH suite:

Gene Finding: Gene models construction, splice sites, protein-coding exons

Total 506 genome-specific parameters are available for genefinders of FGENESH suite
The programs usage in Scientific publications

FGENESH is the fastest and most accurate *ab initio* gene prediction program available - for more details, see [FGENESH help](#). Its variants that use similarity information: FGENESH+ (similar protein), FGENESH_C (similar cDNA), FGENESH-2 (two homologous genomic sequences) greatly improve accuracy of gene prediction when such similarity information is available. These programs can be accessed [here](#).

To find genes in Bacterial sequences click [here](#).

Our two best gene finders cannot be accessed at our site due to computing resources limitations. These two are FGENESH++ (automated version of FGENESH+) and FGENESH++C, which maps known mRNA/EST sequences from RefSeq and then performs FGENESH++-like gene prediction, resulting in fully automatic annotation of quality similar to that of manual annotation.

FGENES, FGENES-M, FGENESH_GC and SPLM can be used on human sequences only.
BESTORF and Fsplice can be used with 296 organisms sequences.
SPL can be used for human, Drosophila, nematode, *S.cerevisiae*, and dicots.

FGENESH - HMM-based gene structure prediction (multiple genes, both chains) [Help] [Example]

FGENES - Pattern based human gene structure prediction (multiple genes, both chains) [Help] [Example]

FGENES-M - Pattern-based human multiple variants of gene structure prediction [Help] [Example]

FGENESH-M - Prediction of multiple variants potential genes in genomic DNA [Example]

Fig 3: Selecting ‘FGENESH - HMM-based gene structure prediction (multiple genes, both chains)’

The screenshot shows the FGENESH service page on the Softberry website. The main title is "Services Test Online" followed by "FGENESH". Below the title, it says "Used in more than 2800 publications" and provides a reference: "Reference: Solovyev V, Kosarev P, Seledsov I, Vorobьев D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 2006;7, Suppl 1: P.10.1-10.12." A large text area is labeled "Paste nucleotide sequence here:" with a red box around the input field. Below this, there's a "Local file name:" input field with "Choose File" and "No file chosen". To the right, a note says "Total 506 genome-specific parameters are available for genefinders of FGENESH suite". On the left sidebar, there's a list of services including "Gene finding in Eukaryota" (which is highlighted in dark blue), "Gene finding with similarity", "Operon and Gene Finding in Bacteria", "Gene Finding in Viral Genomes", "Next Generation", "Alignment (sequences and genomes)", "Genome visualization tools", "Search for promoters/functional motifs", "Deep learning recognition", "Protein Location", "RNA structures", and "Protein structure".

Fig 4: FGENESH Homepage where the query is to be pasted

The screenshot shows the GenBank homepage from the National Library of Medicine. The search bar at the top contains the query "Homo sapiens CDK1", which is highlighted with a red box. The results page includes sections for "GenBank Overview", "What is GenBank?", "Access to GenBank", "GenBank Data Usage", and "Data Processing, Status and Release". The "GenBank Overview" section provides a brief introduction to the database. The "What is GenBank?" section details its history and collaboration with other databases. The "Access to GenBank" section offers search tips. The "GenBank Data Usage" section discusses the terms of use. The "Data Processing, Status and Release" section provides information about submission guidelines. On the right side, there's a sidebar titled "GenBank Resources" with links to "GenBank Home", "Submission Types", "Submission Tools", "Search GenBank", and "Update GenBank Records". A "Log in" button is also visible in the top right corner.

**Fig 5: Homepage of GenBank database
Query searched for '*Homo sapiens CDK1*' Gene**

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide

Species: Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: [Manage Filters](#)

Results by taxon
Top Organisms [Tree]
Homo sapiens (612)
Clavospora lusitaniae (5)
Cryptosporidium muris RN66 (4)
Cervus elaphus hippelaphus (4)
Galemys pyrenaicus (4)
All other taxa (51)
More...

Find related data
Database:

Search details
("Homo sapiens"[Organism] OR Homo sapiens[All Fields]) AND CDK1[A11 Fields]

Recent activity
Turn Off Clear
Homo sapiens CDK1 (681)
(ADD1) AND "Homo sapiens"[orgn] (57)

Fig 6: Result page

GenBank FASTA Graphics

[Synthetic construct Homo sapiens clone ccsbBroadEn_00270 CDK1 gene encodes complete protein](#)

1. 1,023 bp linear other-genetic
Accession: KJ905162.1 GI: 649150780
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Synthetic construct Homo sapiens clone ccsbBroadEn_00270 CDK1 gene encodes complete protein](#)

2. 1,023 bp linear other-genetic
Accession: KJ90876.1 GI: 649099068
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) transcript variant 5. mRNA](#)

3. 1,718 bp linear mRNA
Accession: NM_001170407.2 GI: 1889675079
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) transcript variant 1. mRNA](#)

4. 1,889 bp linear mRNA
Accession: NM_001786.5 GI: 1653961008
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) transcript variant 6. mRNA](#)

5. 1,935 bp linear mRNA
Accession: NM_001320918.1 GI: 1004170671
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) transcript variant 2. mRNA](#)

6. 1,718 bp linear mRNA
Accession: NM_033379.5 GI: 1890275799
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) transcript variant 4. mRNA](#)

7. 1,754 bp linear mRNA
Accession: NM_001170406.1 GI: 281427277
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens cyclin dependent kinase 1 \(CDK1\) RefSeqGene on chromosome 10](#)

8. 23,522 bp linear DNA
Accession: NG_029877.1 GI: 345478661
[Protein](#) [PubMed](#) [Taxonomy](#)

See more...

Fig 6.a: Query selected

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide Advanced Search Help

GenBank ▾ Send to: ▾ Change region shown

Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA

NCBI Reference Sequence: NM_001320918.1

FASTA Graphics

Go to: ▾

Locus NM_001320918 1935 bp mRNA linear PRI 04-APR-2024

Definition Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA.

Accession NM_001320918 XM_006718082

Version NM_001320918.1

Keywords RefSeq,

Source Homo sapiens (human)

Organism Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference 1 (bases 1 to 1935)

Authors Ci,M., Zhao,G., Li,C., Liu,R., Hu,X., Pan,J., Shen,Y., Zhang,G., Li,Y., Zhang,L., Liang,P. and Cui,H.

Title OTUD4 promotes the progression of glioblastoma by deubiquitinating CDK1 and activating MAPK signaling pathway

Journal Cell Death Dis 15 (3), 179 (2024)

PubMed 38429268

Remark GenetIF: OTUD4 promotes the progression of glioblastoma by deubiquitinating CDK1 and activating MAPK signaling pathway. Publication Status: Online-Only

Reference 2 (bases 1 to 1935)

Authors Kang,X., Chen,H., Zhou,Z., Tu,S., Cui,B., Li,Y., Dong,S., Zhang,Q. and Xu,Y.

Title Targeting Cyclin-Dependent Kinase 1 Induces Apoptosis and Cell Cycle Arrest of Activated Hepatic Stellate Cells

Journal Adv Biol (Weinh) 8 (3), e2300403 (2024)

PubMed 38103005

Remark GenetIF: Targeting Cyclin-Dependent Kinase 1 Induces Apoptosis and

Send to: ▾ Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Articles about the CDK1 gene

Multisite phosphorylation dictates selective E2-E3 pairing as revealed by Ubc8/L [Mol Cell. 2024]

Cancer-associated FBXW7 loss is synthetic lethal with pharmacological target [Mol Oncol. 2024]

Antagonistic roles of canonical and Alternative-RPA in disease-associated tandem C. [Cell. 2023]

See all..

Reference sequence information

RefSeq alternative splicing

See 7 reference mRNA sequence splice variants for the CDK1 gene.

RefSeq protein product

See the reference protein sequence for cyclin-dependent kinase 1 isoform 1.

Fig 7: Entry Selected for the Query ‘*Homo sapiens* cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA’

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Log in

Nucleotide Help

FASTA ▾

Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA

NCBI Reference Sequence: NM_001320918.1

[GenBank](#) [Graphics](#)

>NM_001320918.1 Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA

```
ATCGGGGATTGTCTTGTCACTAGAGAAGTTCTCCACAGAGACTTAAACCTCAAATCTCTGA  
TTGATGCACAAAAGAACATAACTTCGATTTGGCTTGGCCAGCTTGGAAATCATCTCAGAT  
ATATAACATACAGGTAGTAAACACTTGTCACTAGAGACTTATGGGGTGCAGCTCGTTACTCA  
ACTCCAGTTGGATTTGGATGATAGGCCACATATTGCTGAATCAAGAACCTTCCCCATG  
GGGATTCAAGAAAATTGTCAACTCTCAGGATTTTCAAGCTTGGCACTTCCAAATAATGAAAGTGTGCC  
AGAAAGTGAGTCTTACAGGACTTAACAGGATACATTCTCCAAATGGAAACCCAGGGAGGCTTACGATCCCCT  
GTCAAAACATTGGATGAAATGGCTTGGATTTGCTCTGAAAATTGTAATCTATGTCAGGCCAACGAA  
TTCTGACAAAGGGACTGAATCTCATATAATGATTGGACAATCAGATTGAAAGATGTAGCT  
TTCTGACAAAGGGCTTGGATTTGCTCATATGAACTGATGTTGTTTATTGTAACCTCTGTATTT  
TTGCTTCAATATCTTCTGATCACTTGTCAAGCTGACTTCTGCTCTTAAATTCAAATAATAACT  
TAAAAAAATGTAATATCTATGTAATGAAATAATATCTGTGATTTGTTGAGCTTCAAGTCTGTTAC  
TATTTGTTACATAATAAACTATAATTGATGTCAGGAATCAGGGAAAAATTGAGTTGAGTTGCCCTAAATC  
ATCTCAGGATGATCTGGCTTGGATTTGAGCTTCAAGTCTTAAATTGAGAAATGCTAAGTT  
AAGTTGCTGATGTCTGGATTTGAGTCTTGTGCTGATTTGAGTCTTCACTGAGTTCTGGCATG  
TTGTTGAGTCTTACACAGGGCTTGGAGTATTTCTACTGGTATTTTAAATTGAGCTTAAATGTT  
TAAGCTTGGCTAGGAAACTACATGACGATTTGGAGAAATGATGCTAAATTGAGGAGTTTCTGAAAC  
TAAAGGAGCTTACATGAGGACGCAACCAAAATTGAGCTTCAAGTCTTAAAGGATCAAGGGCTTGGCGAACAG  
GGAAAGACGTTGGATTTGAGTCTTACATGTTTGTGAGTTTGGAAAGCTTGTGCTAAGT  
GAATTCTTACCTGGCTGAGGAACTAAGTCAAGGGAGTTGCTTATCTTGGCTGAGCTGAGTTAA  
AACTCACACATTGGTACACTGTTGAGTCAAGGAGCTTGGCTTAAAGGCTTAAATGATATTAACTAA  
TACTGAGTTGGGAAATTGAGTCTTAACTGTTGAAACAAAAAA
```

Send to: ▾

Analyze this sequence

Run BLAST

Pick Primers

Show in Genome Data Viewer

Articles about the CDK1 gene

Multisite phosphorylation dictates selective E2-E3 pairing as revealed by Ubc8L [Mol Cell. 2024]

Cancer-associated FBXW7 loss is synthetic lethal with pharmacological tarç [Mol Oncol. 2024]

Antagonistic roles of canonical and Alternative-RPA in disease-associated tandem C [Cell. 2023]

[See all...](#)

Reference sequence information

RefSeq alternative splicing

See 7 reference mRNA sequence splice variants for the CDK1 gene.

RefSeq protein product

See the reference protein sequence for cyclin-dependent kinase 1 isoform 1 (NP_001307847.1)

Fig 8: FASTA Format of the entry selected with header files

Fig 9: Paste the FASTA sequence in the Query box and click on search

Fig 10: FGENESH Output for the query '*Homo sapiens CDK1*' (GenBank ID: NM_001320918.1) gene

RESULTS:

The FGENESH program predicted one gene within the genomic DNA sequence of the CDK1 (GenBank ID: NM_001320918.1) gene. This gene is composed of one exon contributing to the protein sequence. The predicted protein sequence consists of 297 amino acids, translated from the exons, and exhibits characteristic features of the DDI73 transcripts. The prediction is supported by high scores assigned to the exon, indicating a high degree of confidence in the gene prediction.

CONCLUSION:

FGENESH program is the fastest and most accurate, ab initio - HMM-based eukaryotic gene prediction program. The *Homo sapiens* CDK1 (GenBank ID: NM_001320918.1) likely encodes for a protein consisting of 297 amino acids. The predicted protein exhibits feature consistent with its role in cell cycle regulation and modulation of RNA polymerase activity, suggesting its potential involvement in processes such as protein synthesis and cell cycle entry. Further experiments are necessary for verification.

Abbreviation:

- G - predicted gene number, starting from start of sequence;
- Str - DNA strand (+ for direct or - for complementary);
- Feature - type of coding sequence:
- CDSf - First (Starting with Start codon);
- CDSi - internal (internal exon);
- CDSL - last coding segment, ending with stop codon);
- TSS - Position of transcription start (TATA-box position and score);
- Start and End - Position of the Feature;
- Weight - Log likelihood*10 score for the feature;
- ORF - start/end positions where the first complete codon starts and the last codon ends.

REFERENCES:

1. Softberry - FGENESH HELP. (n.d.). [Www.softberry.com](http://www.softberry.com/berry.phtml?group=help&subgroup=gfind&topic=fgenesh). Retrieved March 9, 2024.
<http://www.softberry.com/berry.phtml?group=help&subgroup=gfind&topic=fgenesh>
2. FGENESH - HMM-based gene structure prediction. (n.d.).
<http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind&advanced=on>
3. Yang, Y., Liu, Q., Guo, X., Yuan, Q., Nian, S., Kang, P., Xu, Z., Li, L., & Ye, Y. (2022, August 31). Systematic Pan-Cancer Analysis Identifies CDK1 as an Immunological and Prognostic Biomarker. *Journal of Oncology*, 2022, 1–24.
<https://doi.org/10.1155/2022/8115474>
4. Enserink, J. M., & Chymkowitch, P. (2022, January 24). Cell Cycle-Dependent Transcription: The Cyclin Dependent Kinase Cdk1 Is a Direct Regulator of Basal Transcription Machineries. *International Journal of Molecular Sciences*, 23(3), 1293.
<https://doi.org/10.3390/ijms23031293>

DATE: 06/03/2024

WEBLEM 1(B)

FGENESB: Bacterial Operon and Gene Prediction

(URL:<http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb>)

AIM:

To predict bacterial operon and gene for query '*Sulfobacillus thermosulfidooxidans*' (GenBank ID: NZ_MDZD01000033.1) using FGENESB tool.

INTRODUCTION:

FGENESB proposes an assortment of algorithms which can assign operon as well as to determine promoters, terminators and protein coding gene. Besides, the program can be used to determine both tRNA and rRNA genes. FGENESB pipeline that provides completely automatic and comprehensive annotation of microbial genomic sequence. The pipeline identifies protein, tRNA and rRNA genes, finds potential promoters, terminators and operons. Finally potential functions are assigned to predicted proteins using comparison with a set of databases such as COG, KEGG and NR. The package provides options to work with a set of sequences such as scaffolds of bacterial genomes or short reads of DNA extracted from an environmental sample of entire bacterial community. The gene prediction algorithm is based on Markov chain models of coding regions, start of translation and termination sites. Operon models are derived using distances between ORFs, frequencies of neighboring genes in known bacterial genomes and positions of predicted promoters and terminators. The parameters of gene prediction are automatically learned on initial steps of sequence analysis, so the only input necessary for annotation of a new genome is its sequence. The pipeline is actively used in many new genome-sequencing projects and provides a superior accuracy of gene finding (comparing with the other popular bacterial gene finding software) in bacterial contigs and especially in short sequences analyzed in metagenomic projects.

***Sulfobacillus thermosulfidooxidans*:**

Sulfobacillus thermosulfidooxidans stands out as a thermophilic, acidophilic bacterium renowned for its capability to thrive in harsh environments such as acid hot springs, demonstrating a remarkable ability to biosorb heavy metal ions. Its potential in biotechnological recycling, particularly in the context of hazardous waste PCBs, has garnered attention, with ongoing efforts focused on optimizing processes and studying kinetics to enhance efficiency.

Furthermore, investigations into its interactions with other microorganisms, such as *Leptospirillum ferriphilum*, shed light on its biofilm formation and leaching performance dynamics. Coaggregation phenomena during dual-species biofilm formation contribute to its bioleaching efficacy, although outcomes vary depending on *L. ferriphilum's* preculture state. Additionally, studies exploring its tolerance mechanisms to nickel ions and the impact of iron

and media composition on arsenopyrite bioleaching highlight its versatility and potential for diverse industrial and environmental applications.

In essence, *Sulfobacillus thermosulfidooxidans* emerges as a promising candidate for biotechnological endeavors, offering insights into heavy metal biosorption, bioleaching, and microbial interactions, with implications spanning from waste management to resource recovery.

METHODOLOGY:

1. Visit the website <http://www.softberry.com/>
2. Go to the homepage, click on the 'Run Programs online' section and choose the 'Operon and Gene finding in bacteria' in it.
3. Within the "Services Test Online" section, select " FGENESB."
4. Proceed to the GenBank website and enter the query like '*Sulfobacillus thermosulfidooxidans*' and click on search.
5. Identify a relevant entry, extract its FASTA sequence.
6. Paste the acquired sequence into the nucleotide sequence box on the FGENESB page.
7. Pick a specific organism from the 'Choose closest organism' section and proceed by clicking on the process button.

OBSERVATIONS:

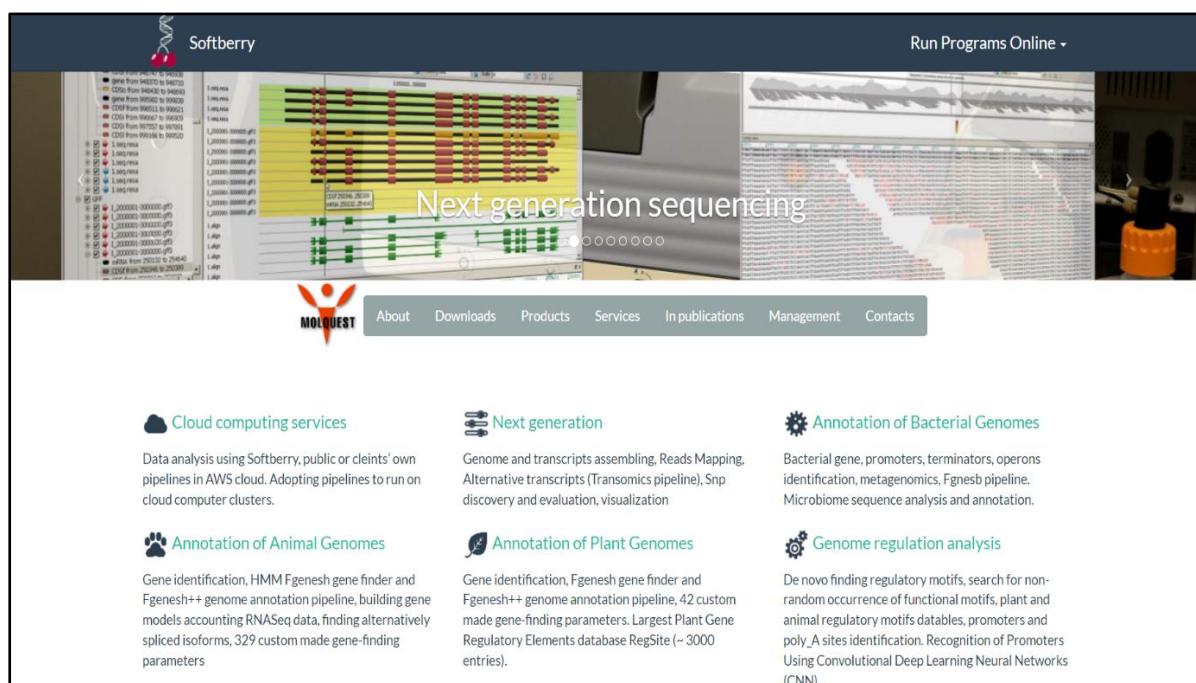


Fig 1: Homepage of Softberry server

The screenshot shows the Softberry MOLQUEST website. At the top right, there is a button labeled "Run Programs Online". Below it, a sidebar menu is open, showing various bioinformatics services. The "Operon and Gene Finding in Bacteria" option is highlighted with a red box. Other options listed in the sidebar include: Gene finding in Eukaryota, Gene finding with similarity, Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs, Deep Learning Recognition, Protein Location, RNA structures, Protein structure, Pathway prediction, Protein/DNA 3D-Visual Works, Manipulations with sequences, Multiple alignments, Synteny from genome contigs, Analysis of gene expression data, Plant Promoter Database, Repeats, SNP, Proteomics, and Gene identification pipelines.

Fig 2: Select the option ‘Operon and Gene Finding in Bacteria’

The screenshot shows the "Services Test Online" page. On the left, a sidebar lists various services: Home, Gene finding in Eukaryota, Gene finding with similarity, **Operon and Gene Finding in Bacteria** (highlighted with a red box), Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs, Deep learning recognition, Protein Location, RNA structures, and Protein structure. The main content area is titled "Bacterial Promoter, Operon and Gene Finding". It includes a section for "The programs usage in Scientific publications" with a red box around the "fgenesB" entry. Other entries in this section include: BPROM - Prediction of bacterial promoters, AbSplit - Separating archaea and bacterial genomic sequences, FindTerm - Finding Terminators in bacterial genomes, Visualization of fgenesB annotation using CGView, Bacterial GenomeSequence Explorer - Visualization of Bacterial genomes information, All bacteria genomes annotations, General scheme of bacterial genome annotation - (automatic pipeline - FgenesB_annotator), Human Microbiome gene prediction, and Softberry Microbiome Annotation Database. A detailed description of the FGENESB program follows, mentioning its speed, accuracy, and requirements.

Fig 3: Select the ‘FGENESB’ option from the list of software

FGENESB: Bacterial Operon and Gene Prediction

Used in more than 330 publications

Reference: V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

FGENESB is a suite of bacterial operon and gene prediction programs: its detailed description is given [here](#). Presented on this page is gene finding portion of FGENESB, which is pattern/Markov chain-based and is the fastest (*E.coli* genome is annotated in appr. 14 sec) and most accurate *ab initio* bacterial gene prediction program available - for more details, see [FGENESB help](#). FGENESB uses genome-specific parameters learned by [FgenesB-train script](#), which requires only DNA sequence from genome of interest as an input. It automatically creates a file with gene prediction parameters for analyzed genome. It took only a few minutes to create such file for *E.coli* genome using its sequence. If you need parameters for your new bacteria, please contact Softberry - we can include them in the web list.

Annotation portion of FGENESB consumes a lot computer resources and is therefore not available at our web site.

Paste nucleotide sequence here (plain or in fasta format):

Alternatively, load a local file with sequence:

Local file name: No file chosen

Table of Genetic code

Choose closest organism: ARCHAE generic

[Help] [Example]

EXAMPLE: Annotation of *Escherichia coli* K12 genome by [FgenesB-Annotator script](#).
[Annotation in GenBank format](#)

Fig 4: FGENESB software main page where the sequence is to be pasted and processed

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

GenBank

[GenBank](#) [Submit](#) [Genomes](#) [WGS](#) [Metagenomes](#) [TPA](#) [TSA](#) [INSDC](#) [Documentation](#) [Other](#)

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research](#) 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Resources

[GenBank Home](#)
[Submission Types](#)
[Submission Tools](#)
[Search GenBank](#)
[Update GenBank Records](#)

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

Data Processing, Status and Release

The most important source of new data for GenBank is direct submissions from a variety of individuals, including researchers, using one of

Fig 5: Homepage of GenBank Database (with the query ‘*Sulfobacillus thermosulfidooxidans*’)

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide Search Help

Species Summary 20 per page Sort by Default order
Bacteria (1,081) Create alert Advanced

Molecule types
genomic DNA/RNA (1,078)
rRNA (4)
Customize ...

Source databases
INSDC (GenBank) (901)
RefSeq (181)
Customize ...

Sequence Type
Nucleotide (1,084)

Genetic compartments
Plasmid (2)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Items: 1 to 20 of 1084

1. [Sulfolobus thermophilic oxidans strain ZBY, whole genome shotgun sequencing project](#)
3,180,484 bp other DNA
This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.
Accession: NZ_MDZ00000000.1 GI: 1133549375
BioProject BioSample PubMed Taxonomy
GenBank

2. [Sulfolobus thermophilic oxidans strain ZBY contig9, whole genome shotgun sequence](#)
404,938 bp linear DNA
Accession: NZ_MDZ01000024.1 GI: 1133549374
Assembly BioProject BioSample Protein PubMed Taxonomy

FILTERS: Manage Filters

RESULTS BY TAXON
Top Organisms [Tree]
More...

Find related data
Database: Select Find items

SEARCH DETAILS
"Sulfolobus thermophilic oxidans"
[Organism] OR Sulfolobus thermophilic oxidans[All Fields]

RECENT ACTIVITY
Turn Off Clear
Sulfolobus thermophilic oxidans (1084) Nucleotide
Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA Nucleotide
Homo sapiens CDK1 (681) Nucleotide
(ADD1) AND "Homo sapiens"[prgn] (57) Nucleotide

Fig 6: Result page obtained

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide Search Help

GenBank Advanced

Sulfolobus thermophilic oxidans strain DX contig9, whole genome shotgun sequence
NCBI Reference Sequence: NZ_MDZ01000033.1
FASTA Graphics

Go to:

LOCUS NZ_MDZ01000033 11198 bp DNA linear CON 13-APR-2023
DEFINITION Sulfolobus thermophilic oxidans strain DX contig9, whole genome shotgun sequence.
ACCESSION NZ_MDZ01000033 NZ_MDZ01000000
VERSION NZ_MDZ01000033.1
DBLINK BioProject: PRJNA224116
BioSample: SAMW5598279
Assembly: GCF_001953285.1
KEYWORDS WGS; RefSeq;
SOURCE Sulfolobus thermophilic oxidans
ORGANISM Sulfolobus thermophilic oxidans
Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiales Family XVII. Incertae Sedis; Sulfolobacillus.
REFERENCE 1 (bases 1 to 11198)
AUTHORS Zhang,X., Liu,X., Liang,Y., Guo,X., Xiao,Y., Ma,L., Miao,B., Liu,H., Peng,D., Huang,W., Zhang,Y. and Yin,H.
TITLE Adaptive Evolution of Extreme Acidophile Sulfolobus thermophilic oxidans Potentially Driven by Horizontal Gene Transfer and Gene Loss
JOURNAL Appl Environ Microbiol 83 (7), e03098-16 (2017)
PUBMED 28115381
REMARK Publication Status: Online-Only

CHANGE REGION SHOWN
CUSTOMIZE VIEW
ANALYZE THIS SEQUENCE
Run BLAST
Pick Primers
Highlight Sequence Features
Find in This Sequence

RELATED INFORMATION
Assembly
BioProject
BioSample
Protein
PubMed
Taxonomy
Components (Core)
Full text in PMC
Identical GenBank Sequence

Fig 7: Query selected for study

Fig 8: FASTA Sequence of the selected query

Fig 9: Paste sequence in the nucleotide sequence section, choose a closest organism and click on the process button

Prediction of potential genes in microbial genomes						
Time: Tue Jan 1 00:00:00 2005						
Seq name: NZ_M2D01000031 Sulfolobus thermosulfidooxidans strain DX contig9, whole						
Length of sequence - 11198 bp						
Number of predicted genes - 9						
Number of transcription units - 7, operons - 2						
N	Tu/Op	Conserved S	Start	End	Score	
1	1 Op 1	.	-	CDS	884 -	1069 101
2	1 Op 2	.	-	CDS	1080 -	1169 101
3	2 Tu 1	.	+	CDS	2056 -	3453 445
4	3 Tu 1	.	+	CDS	3780 -	3902 135
5	4 Tu 1	.	-	CDS	5375 -	5896 -198
6	5 Tu 1	.	+	CDS	6672 -	7196 -64
7	6 Tu 1	.	+	CDS	7543 -	7764 130
8	7 Op 1	.	+	CDS	8575 -	10212 378
9	7 Op 2	.	+	CDS	10124 -	11077 104
Predicted protein(s):						
>GENE 1	884 -	1069 101	61 aa, chain -			
MESVAASPLPGIVTVRSSAAANEHHCPFSMARYIEATGACGRWATYGMITPGLNSVPTC						
P						
>GENE 2	1080 -	1169 74	29 aa, chain -			
VTVGCGHEGLGNPAISAEARALGELVSREVP						
>GENE 3	2056 -	3453 445	465 aa, chain +			
MESYVSHSVDLQGVKVFQDAGWFGTQPTCRNPIFSLRPLPRTRIAQGIIKWAIAANANA						
EVGALPEHNTQIAJIMADEVISGMSGHASNWTTRPGSLAATTGFDADSTVQLRCFSFSLMTRFRPVA						
KIGQVHLVPHNDVHMNAQSINTCHVIAHIAEAIHVHDLPAISRAEETLRTKQHLMWN						
VVKSGSRTHLQAVAMQFQDQVQVYDQVYDQVYDQVYDQVYDQVYDQVYDQVYDQVYDQV						
EYKQPAATQHVAEAKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 4	3780 -	3902 135	40 aa, chain +			
VSDRAVSGGYPAERGNPFAVYPPGNKKAWQAGLTCGEELAAN						
>GENE 5	5375 -	5896 -198	173 aa, chain -			
VTDGEESPRRSANAGMSSEKARENPKRPFQGSRRVPSLGSVGLVKGKARLRSRWHRGGD						
SVTSRSPFREQUDVVRGKGPGRGPARGETEFQVVKRRAEETRKSAATGVYICQYKPTQV						
GENALRATCHABEKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 6	6672 -	7196 -64	174 aa, chain +			
VRGTVRGRFPRRSANAGMSSEKARENPKRPFQGSRRVPSLGSVGLVKGKARLRSRWHRGGD						
SVTSRSPFREQUDVVRGKGPGRGPARGETEFQVVKRRAEETRKSAATGVYICQYKPTQV						
GENALRATCHABEKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 7	7543 -	7764 130	73 aa, chain +			
LGYHPFGWHDLTAAATRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 8	8575 -	10212 378	545 aa, chain +			
LEYPHRPEKSTKEMTMSKTTMSTFSSWWNPHTDEMTHGDNHCORNGTJIMDPOOLST						

Fig 10: FGENESB Output for the query ‘*Sulfolobus thermosulfidooxidans*’

Key: ‘Tu’ denotes Transcription units

‘Op’ denotes Operons

N	Tu/Op	Conserved S	Start	End	Score	
8	7 Op 1	.	+	CDS	8575 -	10212 378
9	7 Op 2	.	+	CDS	10124 -	11077 104
Predicted protein(s):						
>GENE 1	884 -	1069 101	61 aa, chain -			
MESVAASPLPGIVTVRSSAAANEHHCPFSMARYIEATGACGRWATYGMITPGLNSVPTC						
P						
>GENE 2	1080 -	1169 74	29 aa, chain -			
VTVGCGHEGLGNPAISAEARALGELVSREVP						
>GENE 3	2056 -	3453 445	465 aa, chain +			
MESYVSHSVDLQGVKVFQDAGWFGTQPTCRNPIFSLRPLPRTRIAQGIIKWAIAANANA						
EVGALPEHNTQIAJIMADEVISGMSGHASNWTTRPGSLAATTGFDADSTVQLRCFSFSLMTRFRPVA						
KIGQVHLVPHNDVHMNAQSINTCHVIAHIAEAIHVHDLPAISRAEETLRTKQHLMWN						
VVKSGSRTHLQAVAMQFQDQVQVYDQVYDQVYDQVYDQVYDQVYDQVYDQVYDQVYDQV						
EYKQPAATQHVAEAKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 4	3780 -	3902 135	40 aa, chain +			
VEDRAVSGGYPAERGNPFAVYPPGNKKAWQAGLTCGEELAAN						
>GENE 5	5375 -	5896 -198	173 aa, chain -			
VTDGEESPRRSANAGMSSEKARENPKRPFQGSRRVPSLGSVGLVKGKARLRSRWHRGGD						
SVTSRSPFREQUDVVRGKGPGRGPARGETEFQVVKRRAEETRKSAATGVYICQYKPTQV						
GENALRATCHABEKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 6	6672 -	7196 -64	174 aa, chain +			
VRGTVRGRFPRRSANAGMSSEKARENPKRPFQGSRRVPSLGSVGLVKGKARLRSRWHRGGD						
SVTSRSPFREQUDVVRGKGPGRGPARGETEFQVVKRRAEETRKSAATGVYICQYKPTQV						
GENALRATCHABEKTOLFNFQENMFITQNLDAVLEVSGLVRLATAVGKIANDLRLSS						
GPTCLAEKLCPAVCPQGAGMVKNPVMAEADICQYVINGDTUTQQAVAGAGLELNV						
MMPVIAINYFLAH1HLSNQGKAPTMKALAEADWITGVEVENTALATALNFY1GVDQ						
AARVARTAYRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 7	7543 -	7764 130	73 aa, chain +			
LGYHPFGWHDLTAAATRERGKTVRVEENRNLLSEQEINNEALLERIRTHQES						
>GENE 8	8575 -	10212 378	545 aa, chain +			
LEYPHRPEKSTKEMTMSKTTMSTFSSWWNPHTDEMTHGDNHCORNGTJIMDPOOLST						

Fig 10.1: FGENESB Output for the query ‘*Sulfolobus thermosulfidooxidans*’

RESULTS:

Using FGENESB, the *Sulfobacillus thermosulfidooxidans* strain (GenBank ID: NZ_MDZD01000033.1) was analyzed for potential genes, revealing 9 predicted genes within a sequence length of 11198 bp. Among these, 7 transcription units and 2 operons were present. The ‘tu’ in the results denotes the Transcription units and ‘op’ in the results denotes the Operon. The analysis provided details on the start and end points of the genes, their scores, chain affiliations, and gene sequences.

CONCLUSION:

FGENESB gene prediction algorithm is based on Markov chain models and program uses genome-specific parameters learned by FGENESB-train script, which requires only DNA sequence from genome of interest as an input to give its prediction of termination units and operon for the given sequence. The required predictive result for the query *Sulfobacillus thermosulfidooxidans* strain (GenBank ID: NZ_MDZD01000033.1) was obtained.

REFERENCES:

1. FGENESB description. (n.d.).
<https://www.molquest.com/help/2.3/programs/FGENESB/description.html>
 2. Huang, Y., Li, M., Yang, Y., Zeng, Q., Loganathan, P., Hu, L., Zhong, H., & He, Z. (2020, February 12). *Sulfobacillus thermosulfidooxidans*: an acidophile isolated from acid hot spring for the biosorption of heavy metal ions. *International Journal of Environmental Science and Technology*, 17(5), 2655–2666. <https://doi.org/10.1007/s13762-020-02669-1>
 3. Li, Q., Zhu, J., Li, S., Zhang, R., Xiao, T., & Sand, W. (2020, January 29). Interactions Between Cells of *Sulfobacillus thermosulfidooxidans* and *Leptospirillum ferriphilum* During Pyrite Bioleaching. *Frontiers in Microbiology*, 11. <https://doi.org/10.3389/fmicb.2020.00044>
-

Date: 06/03/2024

WEBLEM 1(C)

BPROM: Operon and Gene finding Bacteria

(URL:<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>)

AIM:

To predict bacterial promoter for query '*Sulfbacillus thermosulfidooxidans*' (GenBank ID: NZ_MDZD01000033.1) using BPROM tool.

INTRODUCTION:

BPROM is a bacterial promoter prediction program developed by Softberry to identify potential transcription start positions of bacterial genes regulated by sigma70 promoters, particularly in the major E. coli promoter class. The algorithm utilizes a linear discriminant function (LDF) that combines characteristics describing functional motifs and oligonucleotide composition of these sites. BPROM demonstrates an accuracy of about 80% in recognizing E. coli promoters, with a similar specificity when tested on sets containing equal numbers of promoter and non-promoter sequences. To enhance specificity, it is recommended to run BPROM on a region between two neighboring ORFs located on the same strand or on a sequence upstream from an ORF, considering that most promoters are situated within a 150 bp region from the protein coding sequence. The output of BPROM includes the name of the sequence, LDF threshold, length of the presented sequence, number of predicted promoters, positions of predicted promoters, their scores, and the weights of two conserved promoter boxes. Promoter positions are assigned to the first nucleotide of the transcript (Transcription Start Site position). Additionally, for each predicted promoter, elements of Transcriptional factor binding sites are presented if found. This tool serves as a valuable resource for researchers involved in bacterial genome analysis and gene prediction, contributing to the understanding of gene regulation mechanisms in bacteria.

***Sulfbacillus thermosulfidooxidans*:**

Sulfbacillus thermosulfidooxidans stands out as a thermophilic, acidophilic bacterium renowned for its capability to thrive in harsh environments such as acid hot springs, demonstrating a remarkable ability to biosorb heavy metal ions. Its potential in biotechnological recycling, particularly in the context of hazardous waste PCBs, has garnered attention, with ongoing efforts focused on optimizing processes and studying kinetics to enhance efficiency.

Furthermore, investigations into its interactions with other microorganisms, such as *Leptospirillum ferriphilum*, shed light on its biofilm formation and leaching performance dynamics. Coaggregation phenomena during dual-species biofilm formation contribute to its bioleaching efficacy, although outcomes vary depending on *L. ferriphilum*'s preculture state. Additionally, studies exploring its tolerance mechanisms to nickel ions and the impact of iron

and media composition on arsenopyrite bioleaching highlight its versatility and potential for diverse industrial and environmental applications.

In essence, *Sulfobacillus thermosulfidooxidans* emerges as a promising candidate for biotechnological endeavors, offering insights into heavy metal biosorption, bioleaching, and microbial interactions, with implications spanning from waste management to resource recovery.

METHODOLOGY:

1. Visit the website <http://www.softberry.com/>.
2. Go to the Homepage, click on the ‘Run Program Online’ and choose the ‘Operon and Gene finding Bacteria’ in it.
3. Choose BPROM within the ‘services test online’.
4. Obtain the FASTA sequence for query ‘*Sulfobacillus thermosulfidooxidans*’ from GenBank and paste it into the nucleotide sequence box on the BPROM page.
5. Proceed by clicking on the process button.

OBSERVATIONS:

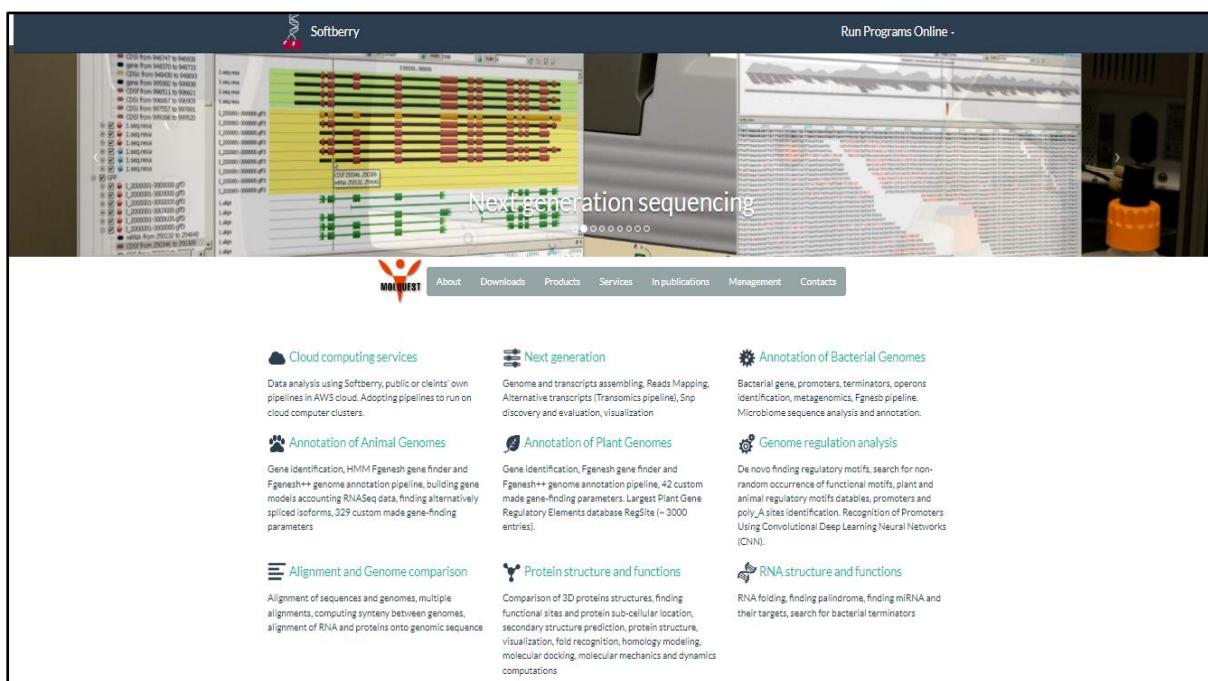


Fig 1: Homepage of Softberry server

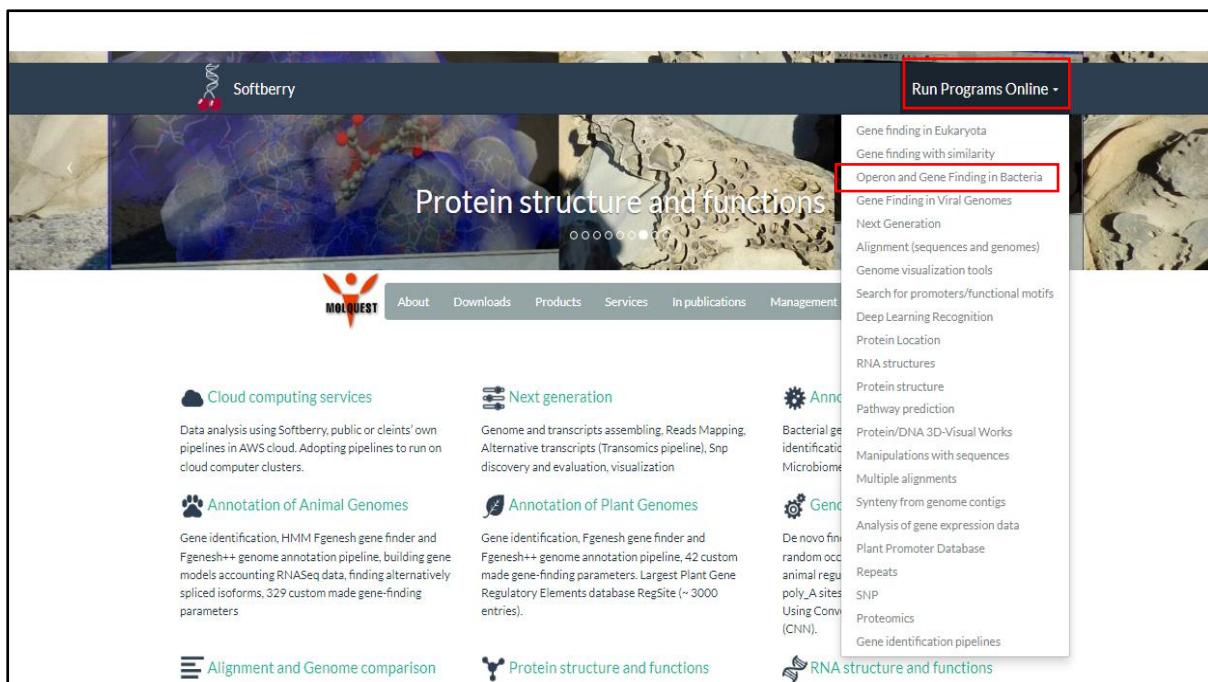


Fig 2: Select the option ‘Operons and Gene finding in Bacteria’

The screenshot shows the 'Services Test Online' page under the 'Bacterial Promoter, Operon and Gene Finding' section. A list of programs is provided, with 'BPROM' highlighted with a red box.

- fgenesB - Pattern/Markov chain-based bacterial operon and gene prediction [Help] [Example]
- BPROM** - Prediction of bacterial promoters [Help] [Example]
- AbSplit - Separating archaea and bacterial genomic sequences [Help] [Example]
- FindTerm - Finding Terminators in bacterial genomes [Help] [Example]
- Visualization of fgenesB annotation using CGView [Example]
- Bacterial GenomeSequence Explorer - Visualization of Bacterial genomes information [Help]
- All bacteria genomes annotations
- General scheme of bacterial genome annotation -(automatic pipeline - Fgenesb_annotator)
- Human Microbiome gene prediction
- Softberry Microbiome Annotation Database

FGENESB is the fastest (*E.coli* genome is annotated in ~14 sec) and most accurate *ab initio* bacterial operon and gene prediction program available - for more details, see **FGENESB help**. It uses genome-specific parameters learned by **FGENESB-train script**, which requires only DNA sequence from genome of interest as an input. It automatically creates a file with gene prediction parameters for analyzed genome. It took only a few minutes to create such file for *E.coli* genome using its sequence. If you need parameters for your new bacteria, please contact Softberry - we can include them in the web list.

Fig 3: Select the BPROM option from the list of programs

The screenshot shows the Softberry Services Test Online interface. On the left, a sidebar lists various bioinformatics tools: Home, Gene finding in Eukaryota, Gene finding with similarity, Operon and Gene Finding in Bacteria (highlighted in dark blue), Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs, Deep learning recognition, Protein Location, RNA structures, Protein structure, and Pathway prediction. The main content area is titled 'BPROM' and describes it as a bacterial sigma70 promoter recognition program used in over 800 publications. It includes a red-bordered input field for pasting nucleotide sequences and a link to upload local files. Below the input field, there are 'Process' and 'Reset' buttons, and links for help and examples. A note at the bottom states that use of Softberry programs accepts their Terms of Use.

Fig 4: BPROM program main page where the sequence is to be pasted

The screenshot shows the National Library of Medicine GenBank homepage. The search bar at the top contains the query 'Sulfobacillus thermosulfidooxidans'. The main content area features sections for 'GenBank Overview', 'What is GenBank?', 'Access to GenBank', and 'GenBank Data Usage'. The 'GenBank Overview' section provides a brief history and details about releases. The 'Access to GenBank' section offers search and download options. The 'GenBank Data Usage' section discusses the database's purpose and usage guidelines. On the right side, there is a sidebar titled 'GenBank Resources' with links to GenBank Home, Submission Types, Submission Tools, Search GenBank, and Update GenBank Records. A 'Log in' button is located in the top right corner.

Fig 5: Homepage of GenBank Database (with the query ‘*Sulfobacillus thermosulfidooxidans*’)

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide | Sulfolacillus thermosulfidooxidans | Search | Help

Species: Summary | 20 per page | Sort by Default order | Send to: | Filters: Manage Filters

Molecule types: Molecule types | Genomic DNA/RNA (1,078) | rRNA (4) | Customize ...

Source databases: INSDC (GenBank) (901) | RefSeq (181) | Customize ...

Sequence Type: Nucleotide (1,084)

Genetic compartments: Plasmid (2)

Sequence length: Custom range...

Release date: Custom range...

Revision date: Custom range...

[Clear all](#)

[Show additional filters](#)

TAXONOMY: Sulfolacillus thermosulfidooxidans Was this helpful? [Like](#) [Unlike](#)

Sulfolacillus thermosulfidooxidans is a species of firmicute in the family Clostridiaceae family xvii. incertae sedis.

Taxonomy ID: 28034

Genomes

Items: 1 to 20 of 1084

1. Sulfolacillus thermosulfidooxidans strain ZBY whole genome shotgun sequencing project
3,180,484 bp other DNA
This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.
Accession: NZ_MDZD00000000.1 GI: 1133549375
BioProject BioSample PubMed Taxonomy
GenBank

2. Sulfolacillus thermosulfidooxidans strain ZBY contig9 whole genome shotgun sequence
404,938 bp linear DNA
Accession: NZ_MDZD01000024.1 GI: 1133549374
Assembly BioProject BioSample Protein PubMed Taxonomy

Send to: | Filters: Manage Filters

Results by taxon: Top Organisms [Tree] More...

Find related data: Database: Select | Find items...

Search details: "Sulfolacillus thermosulfidooxidans" [Organism] OR Sulfolacillus thermosulfidooxidans [All Fields]

Recent activity: Turn Off Clear

Sulfolacillus thermosulfidooxidans (1084) Nucleotide

Homo sapiens cyclin dependent kinase 1 (CDK1), transcript variant 6, mRNA Nucleotide

Homo sapiens CDK1 (681) Nucleotide

(ADD1) AND "Homo sapiens"[orgn] (57) Nucleotide

Fig 6: Result page obtained

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide | Nucleotide | Advanced | Search | Help

GenBank | Change region shown | Customize view

Sulfolacillus thermosulfidooxidans strain DX contig9, whole genome shotgun sequence

NCBI Reference Sequence: NZ_MDZD01000033.1

FASTA | Graphics

Go to: ▾

Locus: NZ_MDZD01000033 11198 bp DNA linear CON 13-APR-2023

Definition: Sulfolacillus thermosulfidooxidans strain DX contig9, whole genome shotgun sequence

Accession: NZ_MDZD01000033 NZ_MDZD01000000

Version: NZ_MDZD01000033.1

DBLink: BioProject: PRJNA224116
BioSample: SAMW5598279
Assembly: GCF_001953285.1

Keywords: WGS; RefSeq;

Source: Sulfolacillus thermosulfidooxidans

Organism: Sulfolacillus thermosulfidooxidans
Bacteria; Bacilli; Clostridia; Eubacteriales; Clostridiales
Family XVII. Incertae Sedis; Sulfolacillus.

Reference: 1 (bases 1 to 11198)

Authors: Zhang,X., Liu,X., Liang,Y., Guo,X., Xiao,Y., Ma,L., Miao,B., Liu,H., Peng,D., Huang,W., Zhang,Y. and Yin,H.

Title: Adaptive Evolution of Extreme Acidophile Sulfolacillus thermosulfidooxidans Potentially Driven by Horizontal Gene Transfer and Gene Loss

Journal: Appl Environ Microbiol 83 (7), e03098-16 (2017)

Pubmed: 28115381

Remark: Publication Status: Online-Only

Analyze this sequence: Run BLAST | Pick Primers | Highlight Sequence Features | Find in this Sequence

Related information: Assembly | BioProject | BioSample | Protein | PubMed | Taxonomy | Components (Core) | Full text in PMC | Identical GenBank Sequence

Fig 7: Query selected for study

Fig 8: FASTA Sequence of the selected query

Softberry

Run Programs Online ▾

Home

Gene finding in Eukaryota

Gene finding with similarity

Operon and Gene Finding in Bacteria

Gene Finding in Viral Genomes

Next Generation

Alignment (sequences and genomes)

Genome visualization tools

Search for promoters/functional motifs

Deep learning recognition

Protein Location

RNA structures

Protein structure

Pathway prediction

Services Test Online

BPROM

Used in more than 800 publications.

Reference: V. Solovyev, A. Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

BPROM - Prediction of bacterial promoters

BPROM is bacterial sigma70 promoter recognition program with about 80% accuracy and specificity. It is best used in regions immediately upstream from ORF start for improved gene and operon prediction in bacteria.

Paste nucleotide sequence here (plain or in fasta format):

```
>NZ_MDZD01000033.1 Sulfolobus thermophilic strain DX contig9,  
whole genome shotgun sequence  
GCCTTGGGGGTTTTCTTGGTCCCTTGGGCTTCAGCTAACGTTCACCGT
```

Alternatively, load a local file with sequence:

Local file name: Choose File No file chosen

Process **Reset**

[Help] [Example]

Return to page with other programs of group: [Operon and gene finding in bacteria](#)

Fig 9: Paste the sequence in the nucleotide sequence section and click on process

```

>NZ_MDZD01000033.1 Sulfobacillus thermosulfidooxidans strain DX contig9, whole
Length of sequence- 11198
Threshold for promoters - 0.20
Number of predicted promoters - 24
Promoter Pos: 3601 LDF- 5.93
-10 box at pos. 3584 TGTAAAT Score 86
-35 box at pos. 3564 TTGACT Score 61
Promoter Pos: 10241 LDF- 4.91
-10 box at pos. 10226 GCGTATAAT Score 77
-35 box at pos. 10206 TAGTTA Score 8
Promoter Pos: 2026 LDF- 4.73
-10 box at pos. 2011 GTGTACAT Score 68
-35 box at pos. 1988 TTTATG Score 33
Promoter Pos: 4894 LDF- 4.39
-10 box at pos. 4879 TGCTACAT Score 72
-35 box at pos. 4857 TTGTATG Score 52
Promoter Pos: 451 LDF- 4.06
-10 box at pos. 436 TGATATAAT Score 82
-35 box at pos. 415 GTGTC Score 20
Promoter Pos: 10964 LDF- 3.94
-10 box at pos. 10949 CAATATGAT Score 48
-35 box at pos. 10929 TTGCAT Score 50
Promoter Pos: 3230 LDF- 3.47
-10 box at pos. 3215 CGTTAAAG Score 44
-35 box at pos. 3194 TTGTATG Score 52
Promoter Pos: 8483 LDF- 3.36
-10 box at pos. 8468 TTTPAAATT Score 59
-35 box at pos. 8444 GTGCC Score 6
Promoter Pos: 10616 LDF- 3.07
-10 box at pos. 10601 CGTTACCAT Score 55
-35 box at pos. 10583 TTGGCC Score 55
Promoter Pos: 2539 LDF- 2.82
-10 box at pos. 2524 GTTCATGAT Score 41
-35 box at pos. 2504 TTGGTG Score 47
Promoter Pos: 9426 LDF- 2.72
-10 box at pos. 9411 TGCTGTGCT Score 24
-35 box at pos. 9391 TTGGCT Score 56
Promoter Pos: 9828 LDF- 2.50
-10 box at pos. 9813 GGTTAACGT Score 44
-35 box at pos. 9790 TTGACT Score 61
Promoter Pos: 6171 LDF- 2.48
-10 box at pos. 6156 GCGTCAAT Score 31

```

Fig 10: BPROM Output for the query ‘*Sulfobacillus thermosulfidooxidans*’

```

Oligonucleotides from known TF binding sites:Berry
No such sites for promoter at 3601
For promoter at 10241:
    rpod17: ATTAGTTA at position 10204 Score - 15
    rpod16: CGTATAAT at position 10227 Score - 14
For promoter at 2026:
    lrp: TATTTTTT at position 1983 Score - 11
    argR: TTTTTTAT at position 1985 Score - 13
For promoter at 4094:
    rpod19: ACGTGCTA at position 4876 Score - 12
    rpod17: GCTACAAAT at position 4880 Score - 8
For promoter at 451:
    rpod16: TGATATAA at position 436 Score - 9
For promoter at 10964:
    lrp: TATTTCTTA at position 10936 Score - 8
For promoter at 3230:
    argR2: CATAATTT at position 3175 Score - 8
    nagC: ATATTTTA at position 3176 Score - 7
    rpod18: GATAGAAT at position 3241 Score - 9
For promoter at 8483:
    lrp: TATTTTTT at position 8436 Score - 11
    flhCD: GGCTCTTT at position 8464 Score - 9
No such sites for promoter at 10616
No such sites for promoter at 2539
For promoter at 9426:
    rpod17: CCAAATAG at position 9397 Score - 8
For promoter at 9828:
    rpod17: AATCTTTA at position 9782 Score - 7
For promoter at 6171:
    crp: TGTTGATCT at position 6137 Score - 13
    rpod19: GTGATCTA at position 6138 Score - 11
    hipB: CCCTTAAG at position 6166 Score - 18
For promoter at 8844:
    argR2: TTTTATT at position 8831 Score - 13
No such sites for promoter at 130
For promoter at 2026:
    glpR: TTCAAAAT at position 2058 Score - 6
For promoter at 7818:
    rpod17: ATACACAGG at position 7823 Score - 12
For promoter at 5345:
    rpod3: CTGATAAG at position 5310 Score - 17

```

Fig 10.1: BPROM Output for the query ‘*Sulfobacillus thermosulfidooxidans*’

RESULTS:

The query *Sulfobacillus thermosulfidooxidans* (GenBank ID: NZ_MDZD01000033.1) with the sequence length 11198 base pairs was provided for predicting the promoter regions using the BPROM tool. A total of 24 promoters were predicted, with subsequent lines detailing the positions of these promoters and their scores, including the weights assigned to two conserved promoter boxes. The promoter positions were assigned to the first nucleotide of the transcription start site position.

CONCLUSION:

The result of gene *Sulfobacillus thermosulfidooxidans* strain (GenBank ID: NZ_MDZD01000033.1) highlights the predictive capabilities of BPROM in identifying potential promoter regions within a sequence based on specific criteria and scoring mechanisms. The program's ability to pinpoint transcription start sites and evaluate promoter strength is essential for understanding gene expression regulation and functional genomic.

REFERENCES:

1. BPROM - Prediction of bacterial promoters. (n.d.). <http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>
 2. Huang, Y., Li, M., Yang, Y., Zeng, Q., Loganathan, P., Hu, L., Zhong, H., & He, Z. (2020, February 12). *Sulfobacillus thermosulfidooxidans*: an acidophile isolated from acid hot spring for the biosorption of heavy metal ions. *International Journal of Environmental Science and Technology*, 17(5), 2655–2666. <https://doi.org/10.1007/s13762-020-02669-1>
 3. Li, Q., Zhu, J., Li, S., Zhang, R., Xiao, T., & Sand, W. (2020, January 29). Interactions Between Cells of *Sulfobacillus thermosulfidooxidans* and *Leptospirillum ferriphilum* During Pyrite Bioleaching. *Frontiers in Microbiology*, 11. <https://doi.org/10.3389/fmicb.2020.00044>
-

DATE: 06/03/2024

WEBLEM 1(D)

FPROM: Search for promoters/functional motifs

(URL:<http://www.softberry.com/berry.phtml?topic=fprom&group=programs&subgroup=promoter>)

AIM:

To predict human promoter for query ‘JAK2’ (GenBank ID: NM_001322195) using FPROM tool.

INTRODUCTION:

FPROM by Softberry is a program that predicts potential transcription start positions by combining characteristics describing functional motifs and oligonucleotide composition of these sites. It uses a linear discriminant function to achieve this. For approximately 50-55% true promoter region recognition, FPROM gives one false positive prediction for about 4000 bp. The program's sensitivity and specificity vary based on the threshold used, with higher thresholds leading to increased specificity but decreased sensitivity. FPROM is part of the FGENESB annotator pipeline, which provides automatic annotation of bacterial sequences, identifying protein and RNA genes, potential promoters, terminators, and operon units. The gene prediction algorithm in FPROM is based on Markov chain models of coding regions and translation sites. Operon models consider distances between ORFs, gene frequencies in known genomes, and information from predicted promoters and terminators. The program is highly accurate in predicting bacterial genes and operons.

JAK2:

JAK2, a tyrosine kinase, holds pivotal roles in cellular processes like proliferation, differentiation, and apoptosis, chiefly as part of the JAK2/STAT3 signaling pathway. This pathway transduces signals from extracellular cytokines and growth factors to the nucleus, impacting critical functions such as embryonic development, hemopoiesis, and immune system regulation. However, dysregulation of JAK2/STAT3 signaling, frequently observed in various tumors, contributes to oncogenesis, angiogenesis, and metastasis, often rendering cancers refractory to standard chemotherapy. Consequently, targeting the JAK2/STAT3 pathway emerges as a promising strategy in solid malignancy treatment.

Moreover, acquired mutations in JAK2 underlie several myeloproliferative disorders, including polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Unique clonal mutations, such as those causing constitutive signaling, are implicated in polycythemia vera, while gain-of-function mutations drive clonal expansion of hematopoietic cells in these disorders. Notably, the V617F mutation in JAK2, a somatic mutation found in hematopoietic cells, is strongly associated with myeloproliferative disorders. These insights into JAK2 mutations shed light on their significance in disease pathogenesis and underscore their potential as diagnostic and therapeutic targets.

METHODOLOGY:

1. Visit the website <http://www.ssoftberry.com/>.
2. Go to the Homepage, click on the ‘Run Program Online’ and choose the ‘search for promoters/functional motifs’ in it.
3. Choose FPROM within the service test online.
4. Obtain the FASTA sequence for query ‘JAK2’ from GenBank and paste it into the nucleotide sequence box on the FPROM page.
5. Set the threshold at 0.80 for TATA-box less promoters and proceed by clicking on the process button.

OBSERVATIONS:

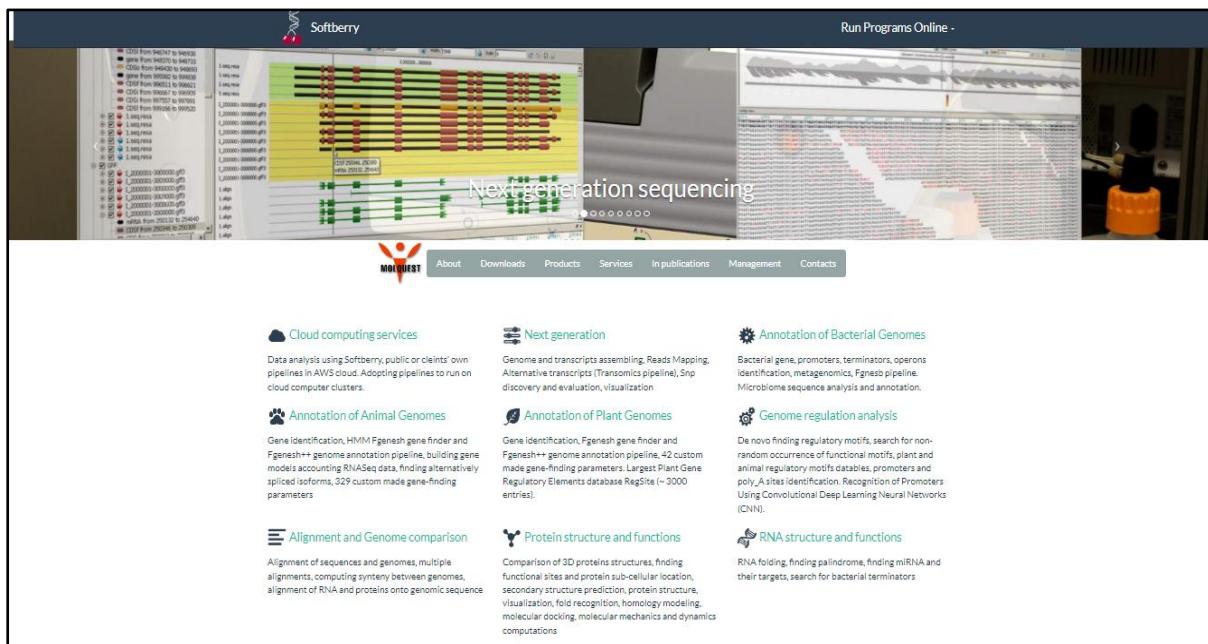


Fig 1: Homepage of Softberry server

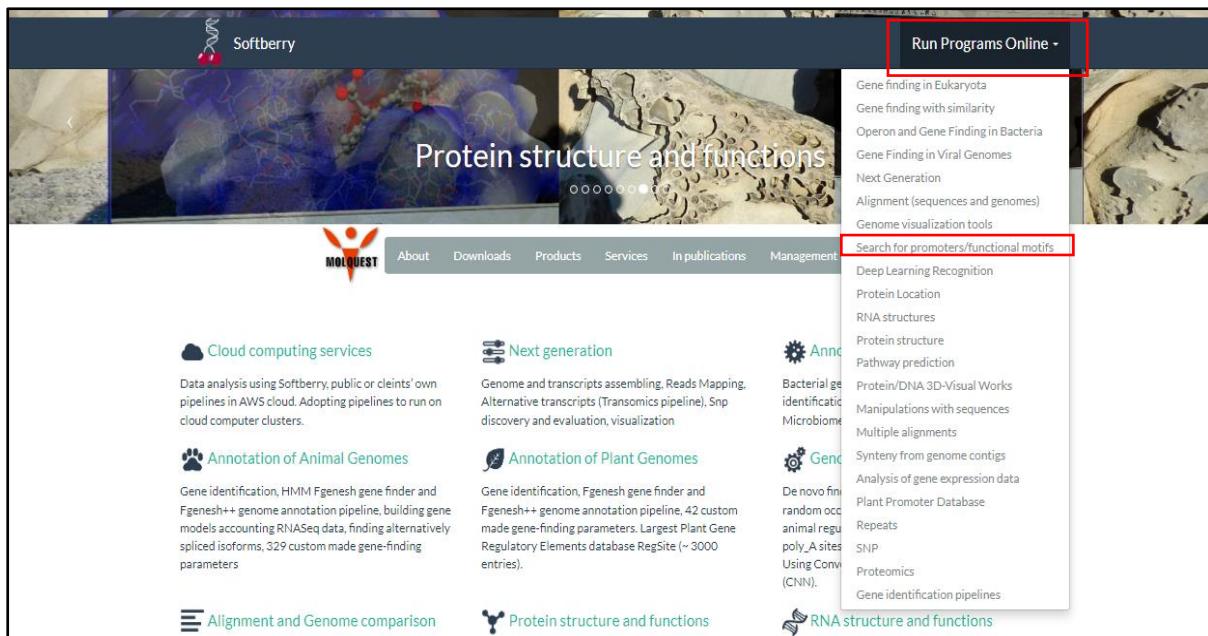


Fig 2: Select the option ‘Search for promoters/functional motifs’

The screenshot shows the 'Services Test Online' page. The left sidebar contains a list of services, with 'Search for promoters/functional motifs' highlighted with a red box. The main content area displays a list of software programs, each with a link to help and example pages. The 'FPROM / Human promoter prediction' option is highlighted with a red box.

Software	Description
FPROM / Human promoter prediction [Help] [Example]	Human promoter prediction
PATTERN / pattern search [Help] [Example]	Pattern search
TSSP / Prediction of PLANT Promoters (Using RegSite Plant DB, Softberry Inc.) [Help] [Example]	Prediction of PLANT Promoters
TSSPlant / Search for RNA polymerase II promoters (TSSs) in plant DNA sequences [Help] [Example]	Search for RNA polymerase II promoters
TSSG / Recognition of human PolII promoter region and start of transcription [Help] [Example]	Recognition of human PolII promoter region
TSSW / Recognition of human PolII promoter region and start of transcription (Transfac DB, Biobase GmbH, ONLY for academic use) [Help] [Example]	Recognition of human PolII promoter region
Nsite-PL / Recognition of PLANT Regulatory motifs with statistics (RegsitePL DB) [Help] [Example]	Recognition of PLANT Regulatory motifs
NsiteM-PL / Recognition of PLANT Regulatory motifs conserved in several sequences (RegsitePL DB) [Help] [Example]	Recognition of PLANT Regulatory motifs conserved in several sequences
PlantPromDB_Blast / BLAST search in sequences of PlantPromDB [Example]	BLAST search in PlantPromDB
Nsite / Recognition of Regulatory motifs (for RE Sets derived from ooTFD, RegsiteAN DB and RegsitePL DB) [Help] [Example]	Recognition of Regulatory motifs
NsiteM / Recognition of Conserved Regulatory motifs (for RE Sets derived from ooTFD, RegsiteAN DB and RegsitePL DB) [Help] [Example]	Recognition of Conserved Regulatory motifs
NsiteH / Search for functional motifs conserved in a pair of orthologous sequences (for RE Sets derived from ooTFD, Regsite AN DB and RegsitePL DB) [Help] [Example]	Search for functional motifs conserved in orthologous sequences
POLYAH / Recognition of 3'-end cleavage and polyadenylation region [Help] [Example]	Recognition of 3'-end cleavage and polyadenylation region

Fig 3: Select the FPROM option from the list of software

Softberry

Run Programs Online ▾

Services Test Online

FPROM

Reference: Solovyev VV, Shahmuradov IA, Salamov AA. (2010) Identification of promoter regions and regulatory sites. *Methods Mol Biol.* 674, 57-83.

FPROM / Human promoter prediction

Paste sequence here:

Alternatively, load a local file with sequence:

Local file name: Choose File No file chosen

Threshold for TATA-box containing promoters

Threshold for TATA-box less promoters 0.80

Process Reset

[Help] [Example of TATA-promoter prediction]

Your use of Softberry programs signifies that you accept Terms of Use

Last modification date: 24 Sep 2013

Fig 4: FPROM page where the sequence is to be pasted

An official website of the United States government [Here's how you know.](#)

National Library of Medicine
National Center for Biotechnology Information

Log in

GenBank Nucleotide JAK2

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Documentation Other

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research](#) 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/genbank/> and <ftp://ftp.ncbi.nlm.nih.gov/genbank/>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

Data Processing, Status and Release

The most important source of new data for GenBank is direct submissions from a variety of individuals, including researchers, using one of

Fig 5: Homepage of GenBank database with the query 'JAK2'

An official website of the United States government. Here's how you know: NIH

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Search Log in Help

Items: 1 to 20 of 527

Filters: Manage Filters

Find related data
Database: Select

Search details
JAK2[All Fields] AND "Homo sapiens"
[orgn]

Recent activity
Turn Off Clear

Q (JAK2) AND "Homo sapiens"[orgn] (527)
Nucleotide

Q JAK2 (6117)
Nucleotide

Q Homo sapiens heat shock transcription factor 1 (HSF1), mRNA
Nucleotide

Q (HSF1) AND "Homo sapiens"[orgn] (1513)
Nucleotide

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾

Species
Animals (495)
Customize ...

Molecule types
genomic DNA/RNA (156)
mRNA (311)
Customize ...

Source databases
INSDC (GenBank) (175)
RefSeq (320)
Customize ...

Sequence type
Nucleotide (487)
EST (7)
GS (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all
Show additional filters

Waiting for www.ncbi.nlm.nih.gov...

Fig 6: Results obtained

An official website of the United States government. Here's how you know: NIH

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Search Log in Help

GenBank Change region shown

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer

Articles about the JAK2 gene
Calreticulin and JAK2V617F driver mutations induce distinct mitotic defects in [Sci Rep. 2024]
Proinflammatory phenotype of iPS cell-derived JAK2 V617F megakary [Stem Cell Reports. 2024]
Identification and validation of the association of Janus kinase 2 mutations with [Inflamm Res. 2024]

See all...

Reference sequence information
RefSeq alternative splicing

Homo sapiens Janus kinase 2 (JAK2), transcript variant 3, mRNA

NCBI Reference Sequence: NM_001322195.2

FASTA Graphics

Go to: ▾

LOCUS NM_001322195 6669 bp mRNA linear PRI 26-FEB-2024

DEFINITION Homo sapiens Janus kinase 2 (JAK2), transcript variant 3, mRNA.

ACCESSION NM_001322195

VERSION NM_001322195.2

KEYWORDS RefSeq

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrates; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

REFERENCE 1 (bases 1 to 6669)

AUTHORS Puli'uaea,C., Emmanuel,T., Green,T.N., Tsai,P., Shepherd,P.R. and Kaley-Zylnska,M.L.

TITLE Insights into the role of JAK2-I724T variant in myeloproliferative neoplasms from a unique cohort of New Zealand patients

JOURNAL Hematology 29 (1), 2297597 (2024)

PUBLISHED 38197452

REMARK GeneIF: Insights into the role of JAK2-I724T variant in myeloproliferative neoplasms from a unique cohort of New Zealand patients.

REFERENCE 2 (bases 1 to 6669)

AUTHORS Flasdorf,N., Bohnke,J., de Toledo,M.A.S., Lutterbach,N., Lerma,V.G., Grasshoff,M., Olschok,K., Gupta,S., Tharmapalan,V.,

Fig 7: Entry selected for further study

Nucleotide Nucleotide Advanced

Send to: Change region shown

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Show in Genome Data Viewer

Articles about the JAK2 gene
Caineticulin and JAK2V617F driver mutations induce distinct mitotic defects in [Sci Rep. 2024]
Proinflammatory phenotype of iPS cell-derived JAK2 V617F megakary [Stem Cell Reports. 2024]
Identification and validation of the association of Janus kinase 2 mutations wif [Inflamm Res. 2024]

See all...

Reference sequence information
RefSeq alternative splicing
See 7 reference mRNA sequence splice variants for the JAK2 gene.
RefSeq protein product

Fig 8: Extract the FASTA sequence

Home
Gene finding in Eukaryota
Gene finding with similarity
Operon and Gene Finding in Bacteria
Gene Finding in Viral Genomes
Next Generation
Alignment (sequences and genomes)
Genome visualization tools
Search for promoters/functional motifs
Deep learning recognition
Protein Location
RNA structures
Protein structure
Pathway prediction

Services Test Online

FPROM

Reference: Solov'yev VV, Shahruradov IA, Salamov AA. (2010) Identification of promoter regions and regulatory sites. *Methods Mol Biol.* 674, 57-83.

FPROM / Human promoter prediction

Paste sequence here:
>NM_001322195.2 Homo sapiens Janus kinase 2 (JAK2), transcript variant 3, mRNA
ATTCGGGAGACTCGAGGCCAACGGGAGCTGAGTCAGCTAGCGAGCAAGGGCCAAAC
Alternatively, load a local file with sequence:

Local file name: Choose File No file chosen

Threshold for TATA-box containing promoters 0.80
Threshold for TATA-box less promoters 0.80

[Process] [Reset] [Help] [Example of TATA-promoter prediction]

Your use of Softberry programs signifies that you accept Terms of Use
Last modification date: 24 Sep 2013

Fig 9: Paste the FASTA sequence into the nucleotide sequence section and click on process

Sequence 1 of 1, Name: NM_001322195.2 Homo sapiens Janus kinase 2 (JAK2), transcript variant 3, mRNA
Length of sequence: 6669
1 promoter/enhancer(s) are predicted
Promoter Pos: 5441 LDF: -0.107 TATA box at 5413 +5.801 TATAAAAT

© 1999 - 2024 www.softberry.com

Fig 10: Result obtained for query 'JAK2'

RESULTS:

The results indicate the sequence name and its length of 6669 base pairs for the query ‘JAK2’ (GenBank ID: NM_001322195). Three promoters were predicted, with the TATA Box identified at positions 5413 within the promoter regions of 5441. This information highlights the presence of specific promoter elements within the sequence, aiding in the understanding of transcriptional regulation mechanisms in bacterial genomes.

CONCLUSION:

FPROM which is a robust tool used for query ‘JAK2’ (GenBank ID: NM_001322195) for predicting potential transcription start positions in bacterial genomes by utilizing a linear discriminant function that combines functional motifs and oligonucleotide composition.

REFERENCES:

1. Softberry - FPROM HELP. (n.d.).
<http://www.softberry.com/berry.phtml?group=help&subgroup=promoter&topic=fproom>
 2. Mengie Ayele, T., Tilahun Muche, Z., Behaile Teklemariam, A., Bogale, A., & Chekol Abebe, E. (2022, February). Role of JAK2/STAT3 Signaling Pathway in the Tumorigenesis, Chemotherapy Resistance, and Treatment of Solid Tumors: A Systemic Review. *Journal of Inflammation Research*, Volume 15, 1349–1364.
<https://doi.org/10.2147/jir.s353489>
 3. Levine, R. L., Wadleigh, M., Cools, J., Ebert, B. L., Wernig, G., Huntly, B. J., Boggon, T. J., Wlodarska, I., Clark, J. J., Moore, S., Adelsperger, J., Koo, S., Lee, J. C., Gabriel, S., Mercher, T., D’Andrea, A., Fröhling, S., Döhner, K., Marynen, P., . . . Gilliland, D. G. (2005, April). Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell*, 7(4), 387–397. <https://doi.org/10.1016/j.ccr.2005.03.023>
-

DATE: 06/03/2024

WEBLEM 1(E)

TSSP: Prediction of PLANT Promoters

(URL:<http://www.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter>)

AIM:

To predict plant promoters for query ‘Glutenin’ (GenBank ID: X03346.1) using TSSP tool.

INTRODUCTION:

The TSSP software by Softberry is a comprehensive tool for various genomic analyses and data processing. It is used for Recognition of Pol II promoter region, enhancers and start of transcription. The algorithm used in the TSSP predicts potential transcription start positions by linear discriminant function combining characteristics describing functional motifs and oligonucleotide composition of these sites. TSSP uses file with selected factor binding sites from RegSite DB (Plants) developed by Softberry Inc.

Glutenin:

Glutenin, a key component of gluten found in wheat endosperms, contributes significantly to the elastic properties crucial for the texture and structure of wheat-based products. Comprising high-molecular-weight glutenin subunits (HMW-GS) and low-molecular-weight glutenin subunits (LMW-GS), glutenin forms an elastic network within dough, providing strength and stability. Allelic variations at glutenin loci have been extensively studied, revealing associations between specific haplotypes and gluten strength, with observed changes in allelic frequencies over time coinciding with alterations in gluten strength. Additionally, the thermal stability and structural changes of glutenin and gliadin proteins in response to wheat bran dietary fiber (WBDF) incorporation have been investigated, highlighting modifications in protein characteristics and aggregation behavior induced by WBDF.

Moreover, nitrogen application during the wheat booting stage has been shown to enhance wheat gluten properties, including wet gluten content, sedimentation values, and dough strength, attributed to increased protein composition and modifications in glutenin polymerization and microstructure. Furthermore, the development of 3D porous scaffolds from wheat glutenin for cultured meat applications underscores the versatility of glutenin in novel food technologies. The gliadin/glutenin ratio has also been found to influence the pasting, thermal, and structural properties of wheat starch, offering insights into optimizing wheat-based product formulations. These studies collectively advance our understanding of glutenin's role in wheat quality and its potential applications in various food and biotechnological contexts.

METHODOLOGY:

1. Visit the website <http://www.softberry.com/>
2. Go to the homepage, click on the ‘Run Programs online’ section and choose the ‘Search for promoters/functional motifs’ in it.
3. Within the "Services Test Online" section, select "TSSP".
4. Proceed to the GenBank website and enter the query like ‘Glutenin’ and click on search.
5. Identify a relevant entry, extract its FASTA sequence.
6. Paste the acquired sequence into the nucleotide sequence box on the TSSP page.
7. Proceed by clicking on the process button.

OBSERVATIONS:

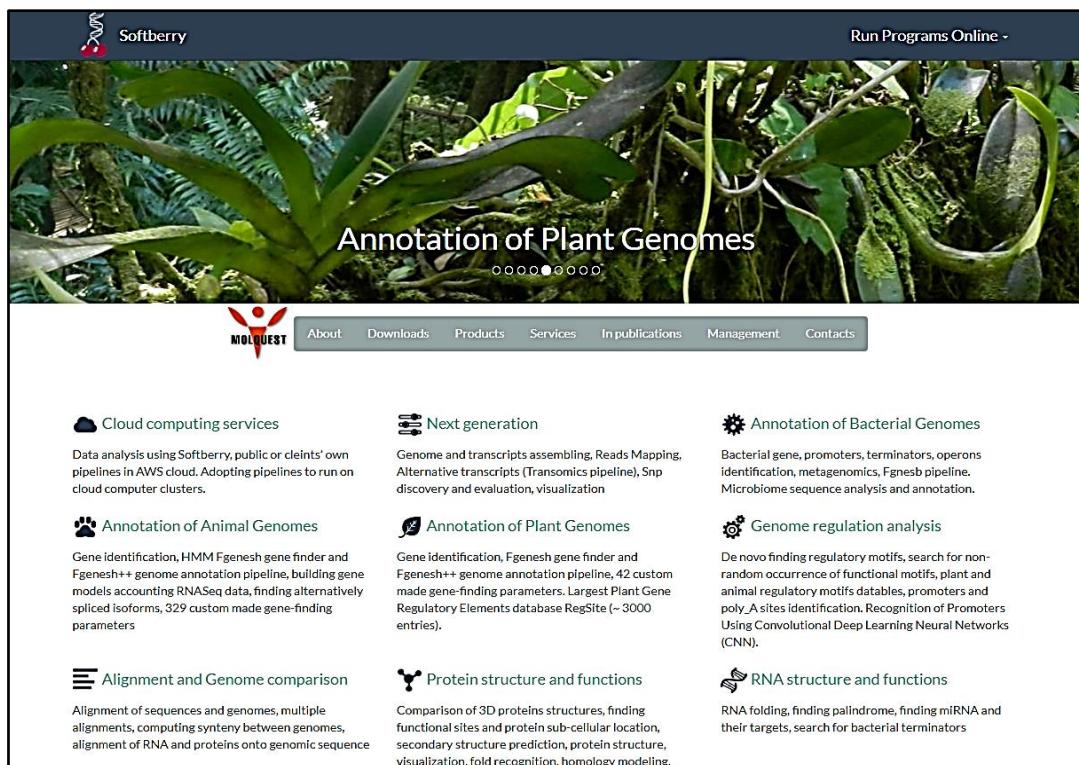


Fig 1: Homepage of Softberry Database

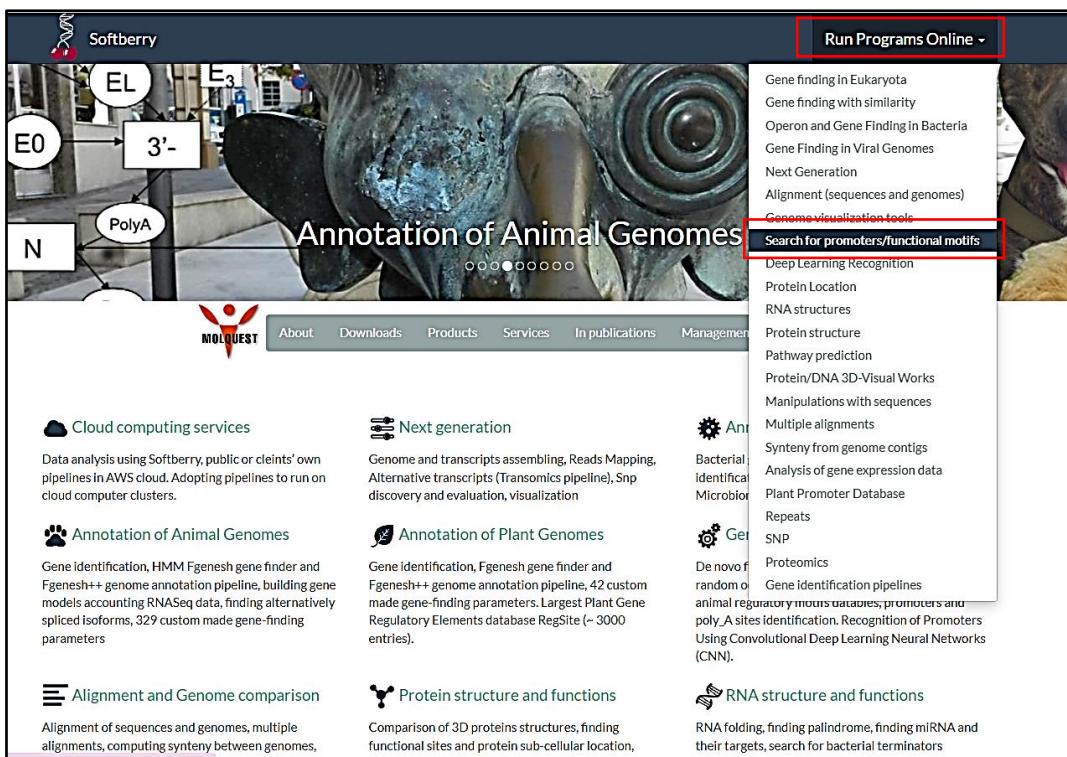


Fig 2: Select the option ‘Search for promoters/functional motifs’

The screenshot shows the Softberry website's 'Services Test Online' page. On the left, there is a sidebar with various links. A red box highlights the 'Search for promoters/functional motifs' link. In the main content area, there is a heading 'Search for promoters/functional motifs'. Below it, there is a list of software options. A red box highlights the 'TSSP / Prediction of PLANT Promoters (Using RegSite Plant DB, Softberry Inc.) [Help] [Example]' option.

Fig 3: Select the ‘TSSP’ option from the list of software

The screenshot shows the Softberry Services Test Online interface. On the left, there is a sidebar with various links: Home, Gene finding in Eukaryota, Gene finding with similarity, Operon and Gene Finding in Bacteria, Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs (which is highlighted in a dark blue box), Deep learning recognition, Protein Location, RNA structures, Protein structure, Pathway prediction, Protein/DNA 3D-Visual Works, Manipulations with sequences, and Multiple alignments.

The main content area is titled "Services Test Online" and "TSSP". It states "Used in more than 240 publications." and "Reference: Solovyev VV, Shahmuradov IA, Salamov AA. (2010) Identification of promoter regions and regulatory sites. Computational Biology of Transcription Factor Binding, Volume 674 of the series *Methods in Molecular Biology*, 57-83." Below this is a section titled "TSSP / Prediction of PLANT Promoters (Using RegSite Plant DB, Softberry Inc.)". A red box highlights the input field where users can "Paste nucleotide sequence here:" followed by a large empty text area. Below it, there is an alternative input method: "Alternatively, load a local file with sequence in Fasta format:" with a "Choose File" button and a message "No file chosen". There are also "Process" and "Reset" buttons, and links for "[Help]" and "[Example]". At the bottom, it says "Your use of Softberry programs signifies that you accept Terms of Use" and "Last modification date: 26 Oct 2016".

Fig 4: TSSP software main page where the sequence is to be pasted and processed

The screenshot shows the National Library of Medicine GenBank homepage. At the top, there is a banner with the NIH logo and the text "An official website of the United States government [Here's how you know](#)". The main navigation bar includes links for GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, Documentation, Other, and a "Search" button. A red box highlights the search input field which contains the query "glutenin". To the right of the search bar, there is a "Log in" button. The page content includes sections for "GenBank Overview", "What is GenBank?", "Access to GenBank", "GenBank Data Usage", and "Data Processing, Status and Release". The "GenBank Overview" section provides a brief description of the database and its history. The "Access to GenBank" section lists several ways to search and retrieve data. The "GenBank Data Usage" section discusses the database's design and access principles. The "Data Processing, Status and Release" section provides information about new data submissions and releases. On the right side, there is a "GenBank Resources" sidebar with links to GenBank Home, Submission Types, Submission Tools, Search GenBank, and Update GenBank Records.

Fig 5: Homepage of GenBank Database (with the query ‘glutenin’)

The screenshot shows the National Library of Medicine's Nucleotide search interface. The search term 'glutenin' is entered in the search bar. The results page displays a list of items found, starting with 'Wheat gene for HMW glutenin subunit' (X03346.1). Other entries include partial mRNA sequences from *Microsporum canis*. The right sidebar includes filters for managing search results by taxon, finding related data, and viewing search details.

Fig 6: Results obtained for the query ‘glutenin’

This screenshot shows the detailed sequence information for the wheat gene X03346.1. It includes the sequence definition (3193 bp DNA linear), source (Triticum aestivum), and various annotations such as keywords (glutenin; signal peptide; storage protein) and references (Sugiyama et al., 1985). The right sidebar provides links to related articles about the gene and its analysis.

Fig 7: Entry selected for further study

Fig 8: FASTA sequence of the selected entry

Fig 9: Paste sequence in the nucleotide sequence section and click on the process button



Fig 10: TSSP Output for the query ‘Glutenin’ (GenBank ID: X03346.1)

Key: Sequences in capitalized letters denote conserved domains

Sequences in small letters denote unconserved domains

RESULTS:

The TSSP was utilized to anticipate potential transcription start sites or promoters within the sequence of the ‘glutenin’ (GenBank ID: X03346.1) gene, which spans 3193 base pairs. The thresholds set for TATA+ promoters and TATA-enhancers were 0.02 and 0.04, respectively. The analysis identified a promoter at position 381 with a TATA box located at position 348. In the specific positions of the promoter, lowercase letters represent non-conserved sequences, while uppercase letters denote conserved sequences. The symbols + and – indicate the strands in the results.

CONCLUSION:

TSSP software by is used for Recognition of Pol II promoter region, enhancers and start of transcription. The predictions for the ‘glutenin’ (GenBank ID: X03346.1) gene were generated through the utilization of the TSSP software, yielding the intended outcomes.

REFERENCES:

1. Softberry - TSSP HELP. (n.d.). <http://www.softberry.com/berry.phtml?topic=tssp&group=help&subgroup=promoter>
2. Roncallo, P. F., Guzmán, C., Larsen, A. O., Achilli, A. L., Dreisigacker, S., Molfese, E., Astiz, V., & Echenique, V. (2021, November 18). Allelic Variation at Glutenin Loci (Glu-1, Glu-2 and Glu-3) in a Worldwide Durum Wheat Collection and Its Effect on Quality Attributes. *Foods*, 10(11), 2845. <https://doi.org/10.3390/foods10112845>
3. Li, M., Yue, Q., Liu, C., Zheng, X., Hong, J., Li, L., & Bian, K. (2020, May). Effect of gliadin/glutenin ratio on pasting, thermal, and structural properties of wheat starch. *Journal of Cereal Science*, 93, 102973. <https://doi.org/10.1016/j.jcs.2020.102973>

DATE: 07/03/2024

WEBLEM 2

INTRODUCTION TO GENE EXPRESSION OMNIBUS (GEO)
DATABASE

(URL: <https://www.ncbi.nlm.nih.gov/geo/>)

Gene Expression Omnibus (GEO) is a database supported by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) that accepts raw and processed data with written descriptions of experimental design, sample attributes, and methodology for studies of high-throughput gene expression and genomics. The introduction of DNA microarrays and the Serial Analysis of Gene Expression (SAGE) protocol as methods of simultaneously assaying gene expression of multiple genes in 1995 enabled scientists to study gene expression of hundreds to thousands of genes, thereby vastly increasing the experimental scale and providing a far more complete understanding of biological processes compared to earlier single-gene studies. Microarray technology quickly dominated the field of high-throughput gene expression studies and with the genome sequencing of humans and many model organisms, genome-wide gene expression and other functional genomic studies became commonplace by the early 2000s. The accelerating pace of genomic-level data production and the bulky raw and processed data files they generated created a challenge for individual labs or journals to make the data available to the research community.

In 2000, NCBI launched the GEO database as a repository for high-throughput gene expression data. In 2002, major journals started to require deposit of microarray data into public repositories, and consequently, the content of GEO grew quickly. Furthermore, the nature of high-throughput genomic experiments expanded rapidly since the first microarrays used to analyze gene expression, and thus the GEO database similarly evolved to keep pace with the changing technologies and applications. Today, GEO accepts data from a wide variety of technologies, including DNA microarrays, protein or tissue arrays, high-throughput nucleic acid sequencing, SAGE, and RT-PCR. And while the majority, approximately 90%, of the data in GEO are indeed gene expression data, the applications have also expanded to include studies on genome methylation, genome binding/occupancy, protein profiling, chromosome conformation studies, and genome variation/copy number.

There are three types of GEO submitter records:

1. A **Platform** record describes an array or sequencer and, for array-based platforms, a data table defining the array template. Sample records are linked from Platform records.
2. A **Sample** record describes the sample source, the protocols used in its analysis, and the expression data derived from it. Samples can only reference a single Platform
3. A **Series** record links together a group of related Samples and describes a whole study.

These three types of records are organized into two higher-level categories for querying and analysis:

1. A **Dataset** represents a curated collection of biologically and statistically comparable GEO Samples. All Samples in a Dataset reference the same Platform. Datasets can be searched using the GEO Datasets database.
2. A **Profile** consists of the expression measurements for an individual gene across all Samples in a Dataset. Profiles can be searched using the GEO Profiles database.

The word “geo” is a prefix meaning “earth” because not only does GEO primarily host global gene expression data, GEO itself is indeed a global resource; at the time of this writing GEO contains submissions from 72 nations. There are no fees to submit data to GEO, download data, or use GEO tools. Scientists submit to GEO in order to share their data with the research community and/or as a requirement of publication or grant directives. GEO supports the Minimum Information About a Microarray Experiment (MIAME) and Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE) guidelines set forth by the Functional Genomics Data Society for standardization of information about microarray and sequencing experiments that enable the data to be interpreted and replicated by the research community. The GEO database handles the majority of direct submissions from the research community and at the time of this writing holds 54,640 public studies, comprising over 1.3 million samples, derived from 2889 different organisms.

While the chief role of GEO is to serve as a public data archive, the database is not simply an online warehouse of data. GEO strives to make the data it contains accessible to the research community. Due to the complex nature of the data generated by genomic experiments most studies are analyzed by bioinformaticians and statisticians, or researchers with specialized analysis software. Researchers who lack these skills or software face a substantial challenge if they wish to analyze genomics experiments themselves. In order to make such data analysis accessible to all researchers, GEO has developed several tools for data query, visualization, and analysis that can be performed directly on the GEO website and do not require the download or manipulation of the data files.

NCBI Resources How To

GEO Home Documentation Query & Browse Email GEO Sign in to NCBI

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site
- ENCODE Data Listings and Tracks

Browse Content

Repository Browser	
DataSets:	4348
Series:	223658
Platforms:	25917
Samples:	7072199

Information for Submitters

Login to Submit	Submission Guidelines	MIAME Standards
	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		GEO Publications

Fig 1: Homepage of Gene Expression Omnibus (GEO) Database

REFERENCES:

1. Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. In Methods in molecular biology (pp. 93–110). https://doi.org/10.1007/978-1-4939-3578-9_5
 2. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207–210. <https://doi.org/10.1093/nar/30.1.207>
-

DATE: 07/03/2024

WEBLEM 2(A)

INTRODUCTION TO GENE EXPRESSION OMNIBUS (GEO)

DATABASE

(URL: <https://www.ncbi.nlm.nih.gov/geo/>)

AIM:

To analyze gene expression data from the Gene Expression Omnibus (GEO) database to identify differentially expressed genes associated with 'Bladder tumor'.

INTRODUCTION:

The Gene Expression Omnibus (GEO) is a publicly available, comprehensive database hosted by the National Centre for Biotechnology Information (NCBI). It serves as a repository for high-throughput gene expression and functional genomics data submitted by researchers worldwide. The GEO database facilitates the dissemination and sharing of these valuable datasets, enabling researchers to explore and analyze gene expression patterns across a wide range of experimental conditions, biological processes, and organisms. Established in 2000, GEO has become an indispensable resource for the scientific community, promoting data sharing and collaboration in the field of genomics. It adheres to the Minimum Information About a Microarray Experiment (MIAME) standard, ensuring that submitted data is consistently annotated and described, enabling reproducibility and facilitating meta-analyses.

There are three types of GEO submitter records:

1. A **Platform** record describes an array or sequencer and, for array-based platforms, a data table defining the array template. Sample records are linked from Platform records.
2. A **Sample** record describes the sample source, the protocols used in its analysis, and the expression data derived from it. Samples can only reference a single Platform
3. A **Series** record links together a group of related Samples and describes a whole study.

These three types of records are organized into two higher-level categories for querying and analysis:

1. A **Dataset** represents a curated collection of biologically and statistically comparable GEO Samples. All Samples in a Dataset reference the same Platform. Datasets can be searched using the GEO Datasets database.
2. A **Profile** consists of the expression measurements for an individual gene across all Samples in a Dataset. Profiles can be searched using the GEO Profiles database.

GEO DataSets contain raw and processed data from individual studies, including sample characteristics, experimental design, and raw data files. GEO Profiles, on the other hand, provide a curated and compact representation of the gene expression data, allowing users to quickly identify interesting genes or compare expression patterns across different experimental conditions or sample types.

The Gene Expression Omnibus is a valuable resource for researchers in the field of genomics, providing a centralized repository for gene expression and functional genomics data. Its user-friendly interface, adherence to data standards, and powerful data analysis tools make it an essential tool for exploring gene expression patterns, identifying biomarkers, and advancing our understanding of biological processes and disease mechanisms.

Bladder cancer:

Bladder cancer develops in the lining of the bladder, a storage organ for urine. The most prevalent type, urothelial carcinoma (previously called transitional cell carcinoma), originates in cells lining the bladder and also affects similar structures like the renal pelvis, ureters, and urethra. Less common types include squamous cell carcinoma, adenocarcinoma, and sarcoma. Risk factors include smoking and exposure to certain chemicals. The disease is more frequent in individuals over 55, with men being four times more likely to develop it than women. Blood in the urine is a common symptom, but other conditions can cause it as well. Urinalysis, cytology, cystoscopy, and imaging studies are used for diagnosis.

The stage of the cancer, determined by whether it has invaded the bladder muscle wall, is crucial for treatment decisions. Non-muscle-invasive cancer, the most common type, is typically treated before it spreads. Muscle-invasive cancer has penetrated the lining and requires different treatment approaches. Surgery, radiation therapy, chemotherapy, and immunotherapy are all options, with the specific choice depending on the cancer's stage and type.

METHODOLOGY:

1. Access the GEO database through the National Centre for Biotechnology Information (NCBI) website: (ncbi.nlm.nih.gov/geo).
2. On the GEO homepage, use the search bar to look for specific keywords or GEO accession numbers related to your research topic or study of interest.
3. To find a specific dataset, search by keyword or GEO accession number in the search bar. For example, here searching for "bladder tumor" will display relevant datasets and also profiles. Further, click on 'DataSets'.
4. Once found a dataset of interest, click on it to access its details. The dataset page will provide information such as the title, summary, organism, platform, citation, and sample count along with other information.
5. Under the "Data Analysis Tools" section, select the appropriate test (e.g. Two-tailed t-test) and set the desired significance level (e.g., 0.100) for comparing the sample groups.
6. Identify the sample groups you want to compare.
7. Assign the samples to their respective groups (Group A and Group B) based on the provided information. In this case,
 - Group A: GSM2519, GSM2524, GSM2525
 - Group B: GSM2544, GSM2506, GSM2509
8. Perform the selected statistical test (e.g., Two-tailed t-test) to compare the expression profiles between the two groups.

9. Analyze the results and visualize the data using various tools provided in the GEO Dataset Browser, such as hierarchical clustering, heatmaps, or correlation analysis.
10. Interpret the findings and draw conclusions based on the statistical analysis and visual representations.

OBSERVATIONS:

Fig 1: Homepage of Gene Expression Omnibus (GEO) Database

Fig 2: GEO Database search for the query 'Bladder tumor'

The screenshot shows the National Library of Medicine's GEO DataSets search interface. The search term 'bladder tumor' is entered in the search bar. The results page displays a list of 4678 datasets, with the first three entries detailed below:

- Rosiglitazone and Trametinib to treat basal bladder tumors**
 (Submitter supplied) This SuperSeries is composed of the SubSeries listed below.
 Organism: *Mus musculus*
 Type: Genome binding/occupancy profiling by high throughput sequencing; Expression profiling by high throughput sequencing
 Platforms: GPL19057 GPL24247 69 Samples
 Download date: BW
 Series Accession: GSE234327 ID: 200234327
- Rosiglitazone and Trametinib to treat basal bladder tumors [RNA-Seq]**
 (Submitter supplied) To investigate the anti-tumor effect of Rosiglitazone+Trametinib in basal bladder cancer, we tested the drug combination in BBN-induced basal tumor mouse model in vivo and BBN-derived tumor cell line BBN963 in vitro.
 Organism: *Mus musculus*
 Type: Expression profiling by high throughput sequencing
 Platforms: GPL19057 GPL24247 62 Samples
 Download date: XLSX
 Series Accession: GSE234326 ID: 200234326
- Rosiglitazone and Trametinib to treat basal bladder tumors [ATAC-Seq]**
 (Submitter supplied) To investigate the anti-tumor effect of Rosiglitazone+Trametinib in basal bladder cancer, we tested the drug combination in BBN-induced basal tumor mouse model in vivo and BBN-derived tumor cell line BBN963 in vitro.
 Organism: *Mus musculus*

On the right side of the results page, there are filters, a search bar, and links to related data and documentation.

Fig 3: GEO DataSets obtained for Query ‘Bladder tumor’

The screenshot shows the NCBI Dataset Browser interface for dataset GDS183. The dataset record includes the following details:

- Title:** Bladder tumor stage classification
- Summary:** Identification of clinically relevant subclasses of bladder carcinoma. Three major stages identified, Ta, T1 and T2-4. Ta tumors further classified into subgroups. A 32-gene molecular classifier was built.
- Organism:** *Homo sapiens*
- Platform:** GPL80: [Hu6800] Affymetrix Human Full Length HuGeneFL Array
- Citation:** Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL et al. Identifying distinct classes of bladder carcinoma using microarrays. *Nat Genet* 2003 Jan;33(1):90-6. PMID: 12469123
- Reference Series:** GSE89
- Value type:** count
- Sample count:** 40
- Series published:** 2002/12/08

The interface also features a 'Data Analysis Tools' section with options for cluster analysis, gene search, and disease/tissue filtering. A sidebar on the right provides download links for various file formats.

Fig 4: Results obtained for Dataset: ‘GDS183’

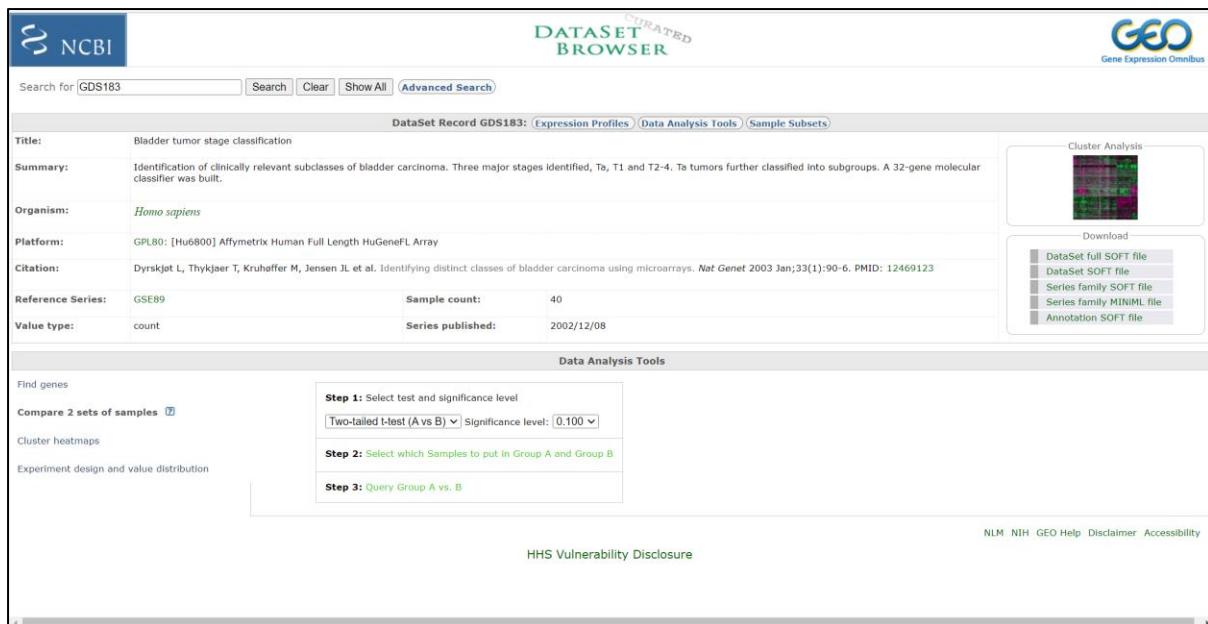


Fig 5: Data Analysis tool

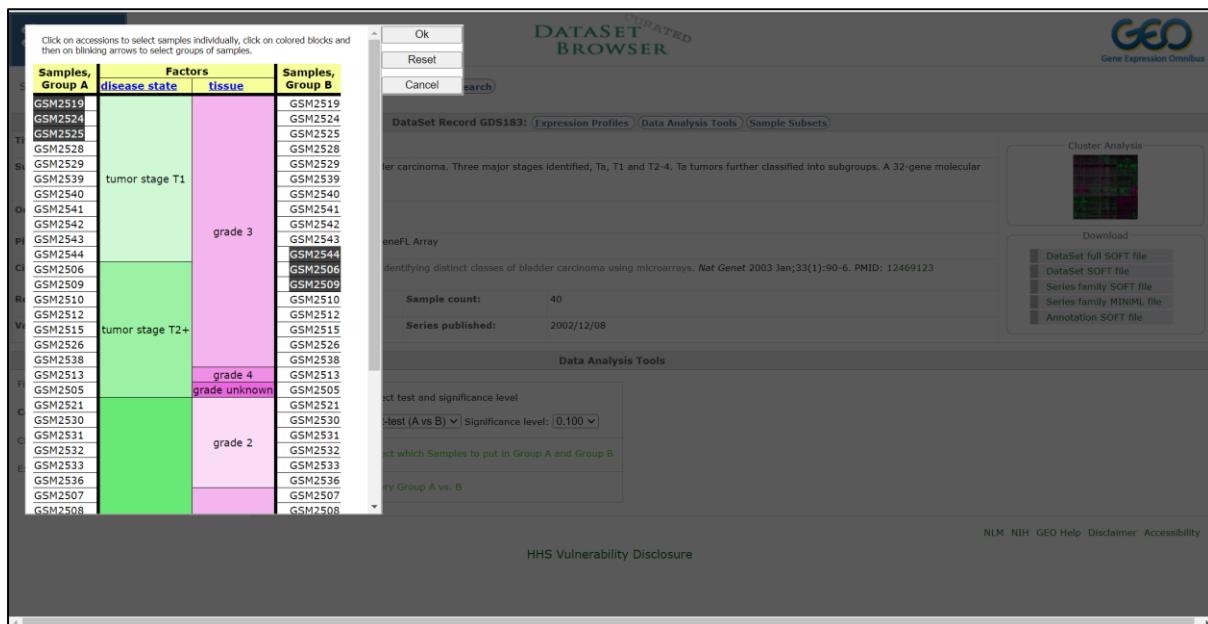


Fig 6: Selection of samples in Group A and Group B

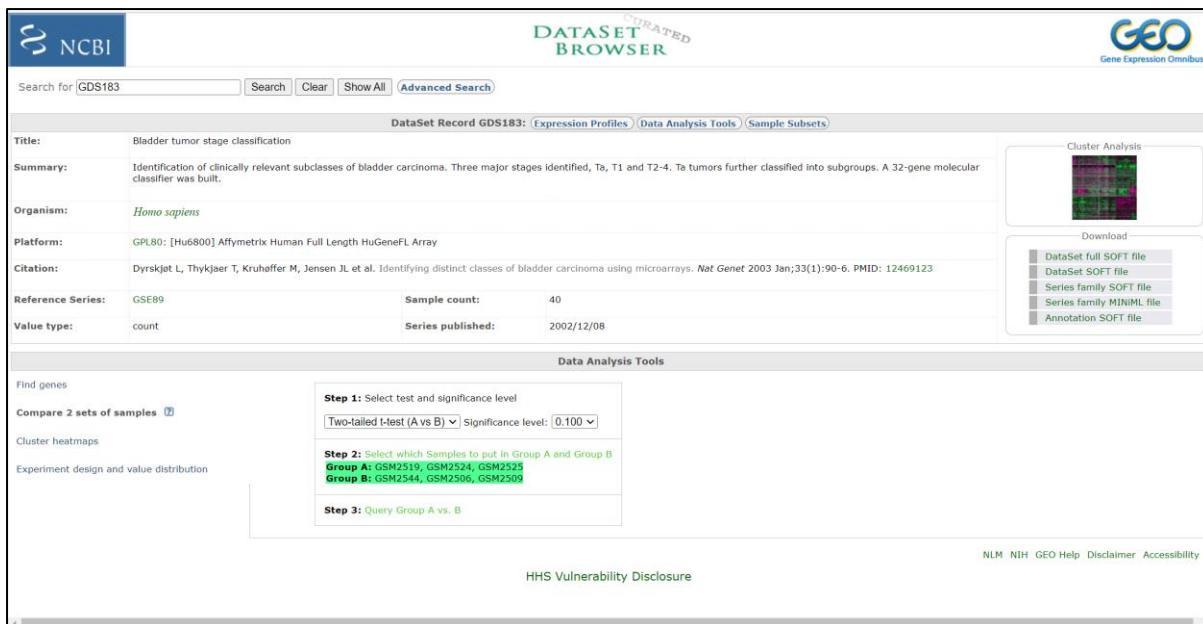


Fig 7: Data Analysis tool after selection of samples for comparison

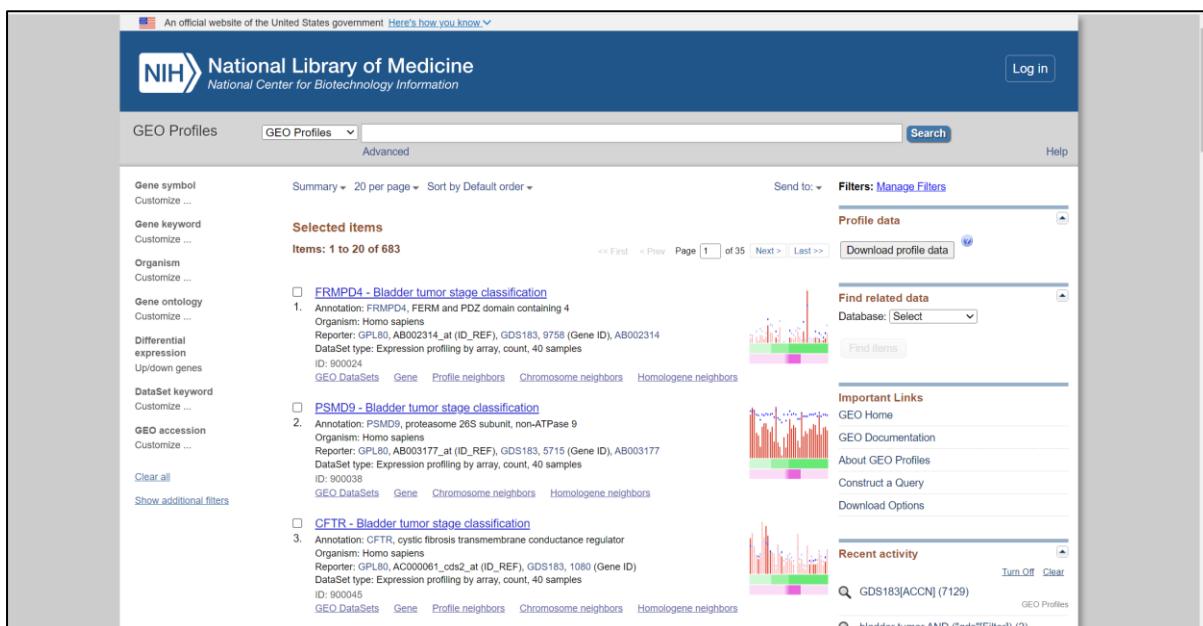


Fig 8: GEO Profiles page for selected samples

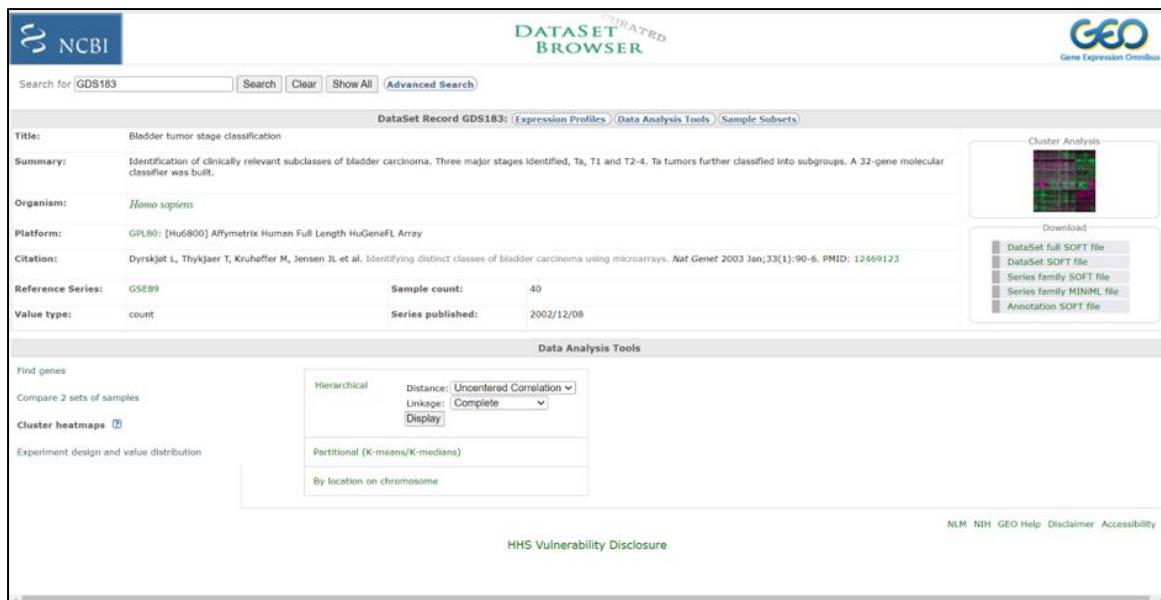


Fig 9: Option to analyze Query GDS183 using cluster maps

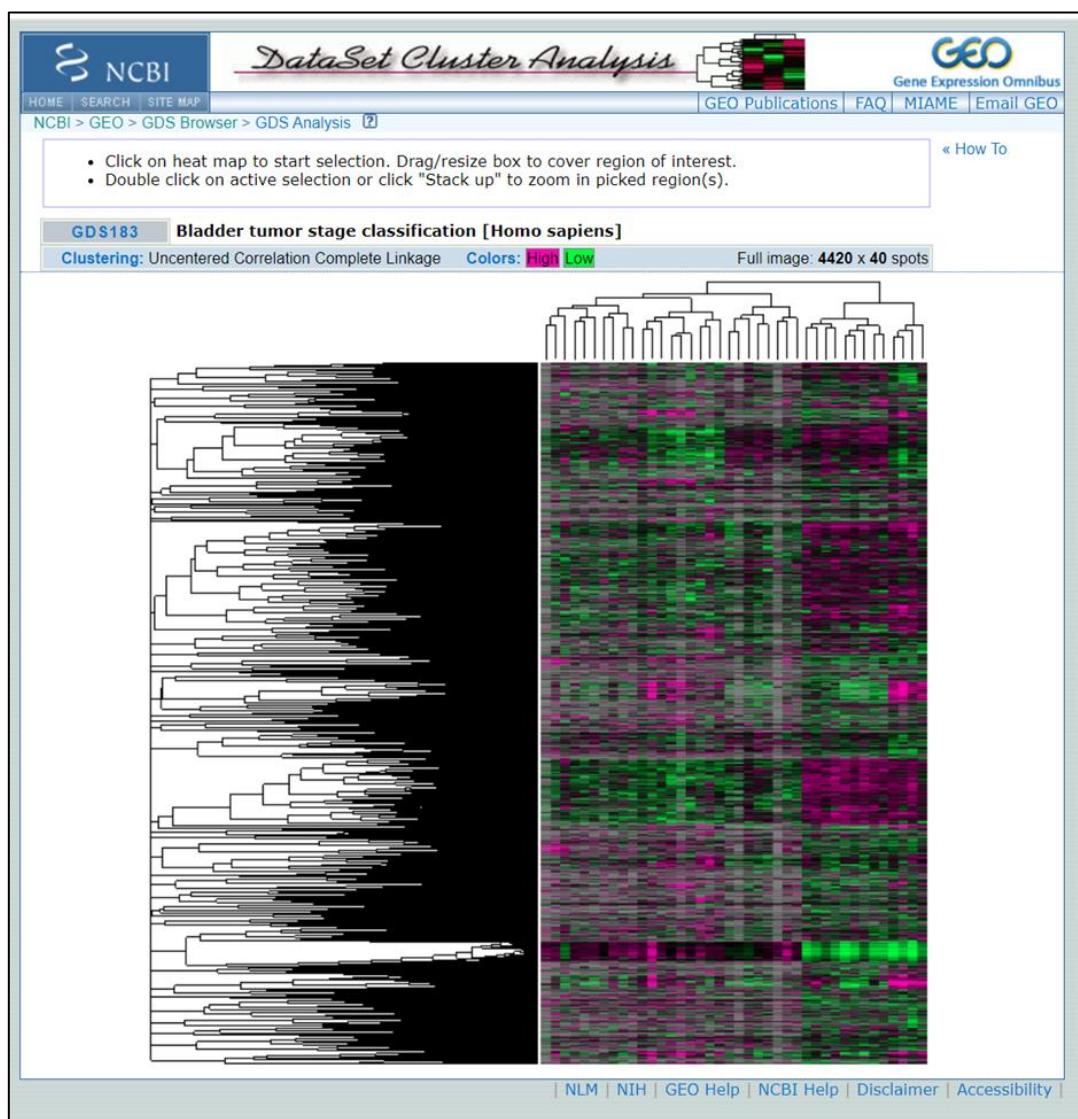


Fig 10: Cluster Heat maps for Query GDS183

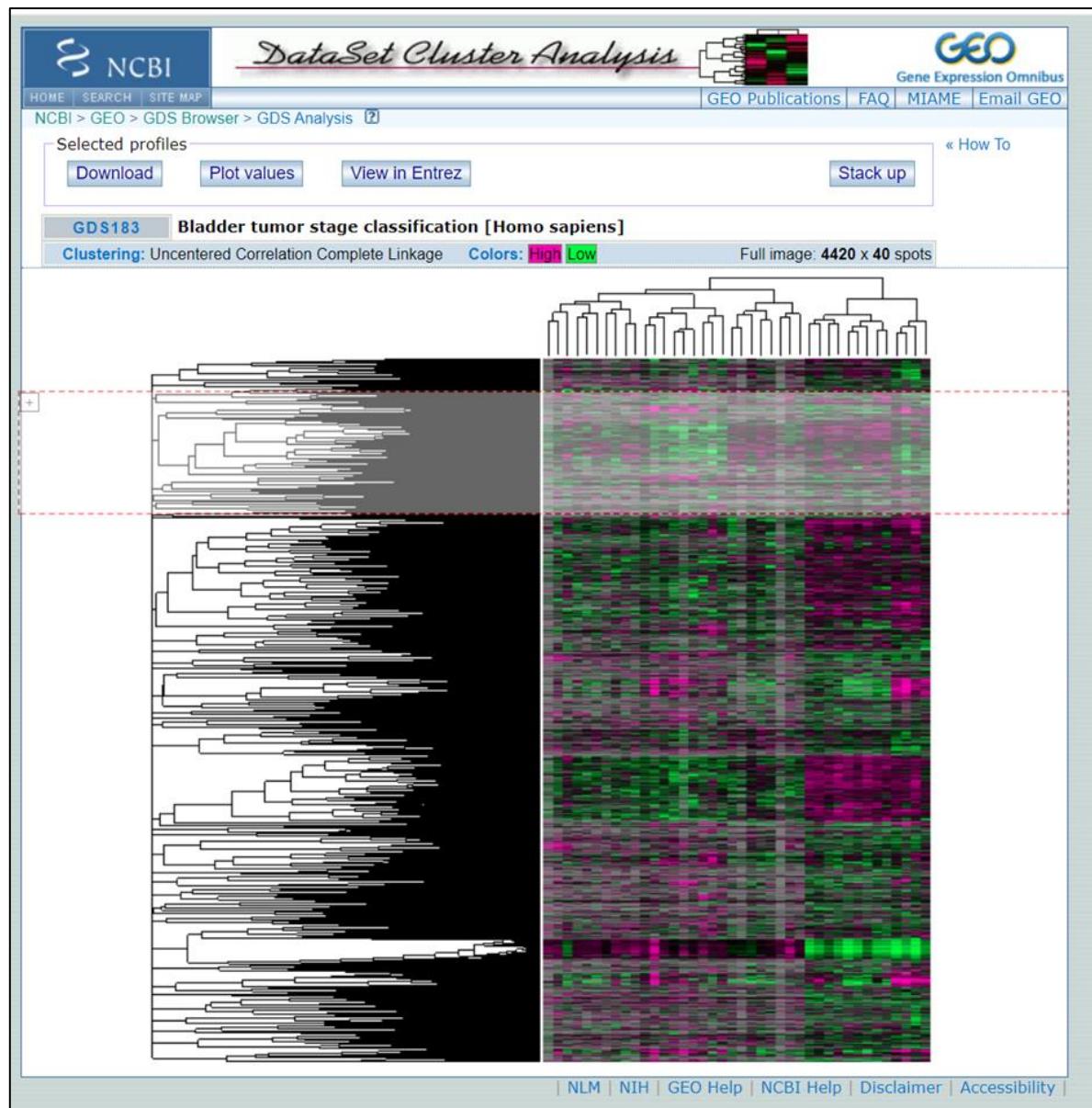


Fig 11: Selecting Region of Interest

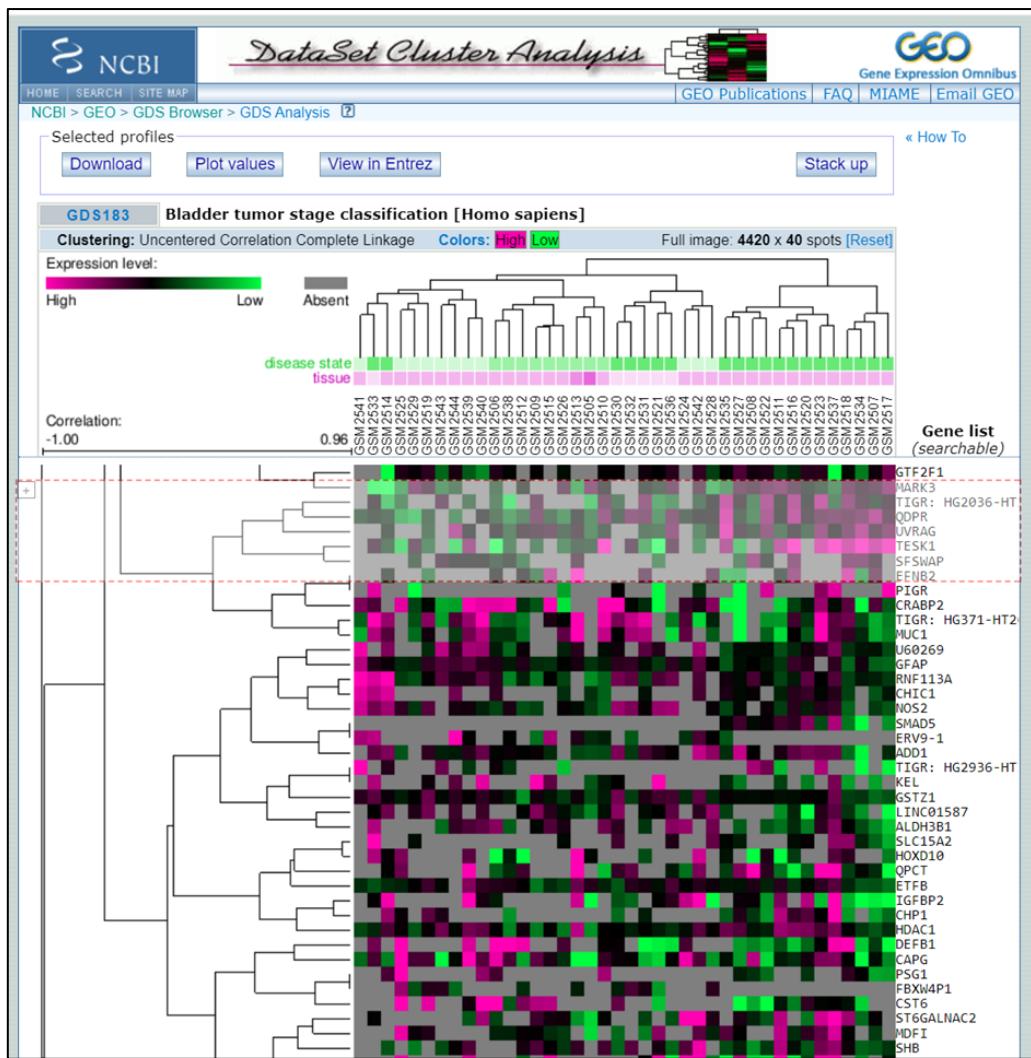


Fig 12: Cluster Heatmaps for selected region

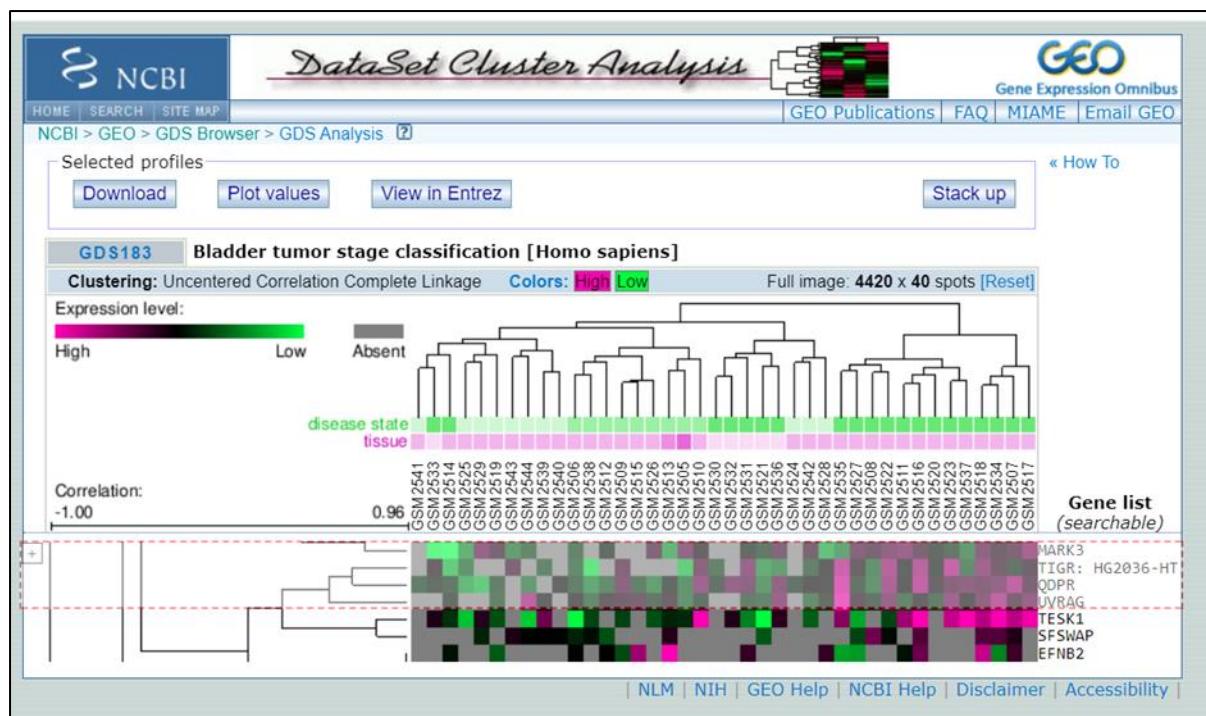


Fig 13: Cluster Heatmaps for Genes selected from Gene Lists in Fig 12

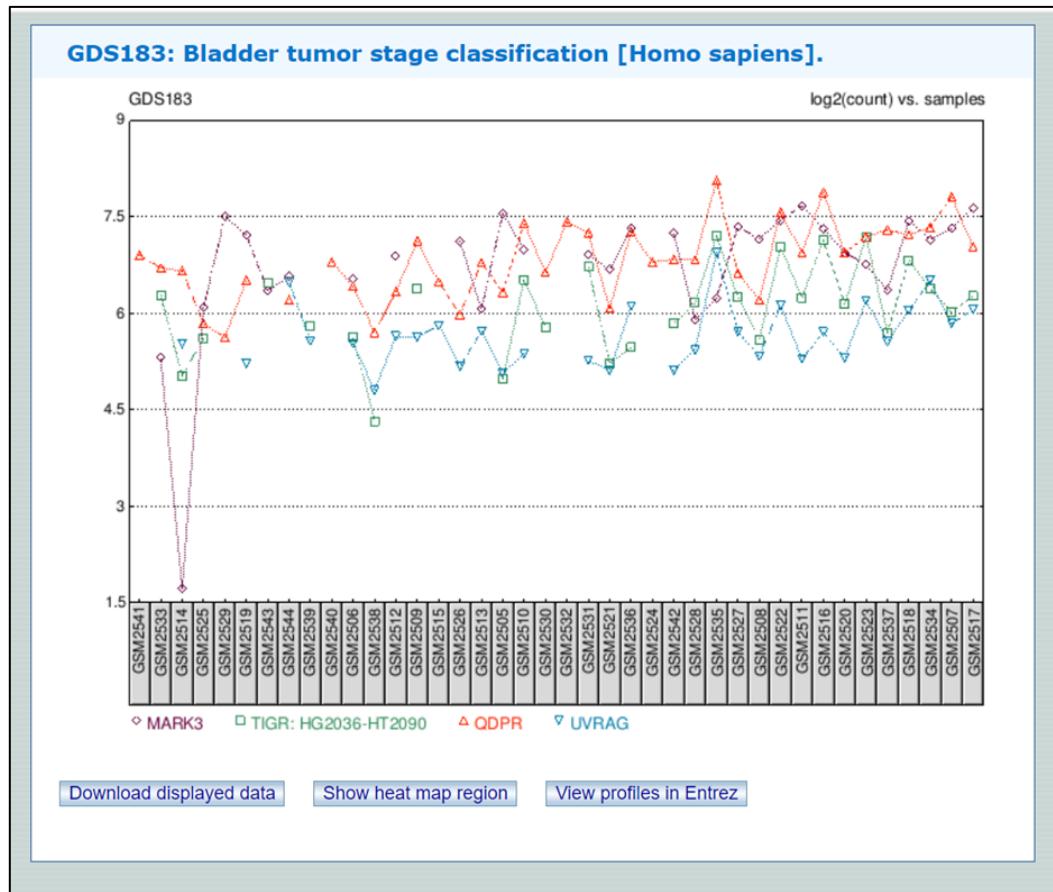


Fig 14: Plot Values of Genes selected in Fig 13

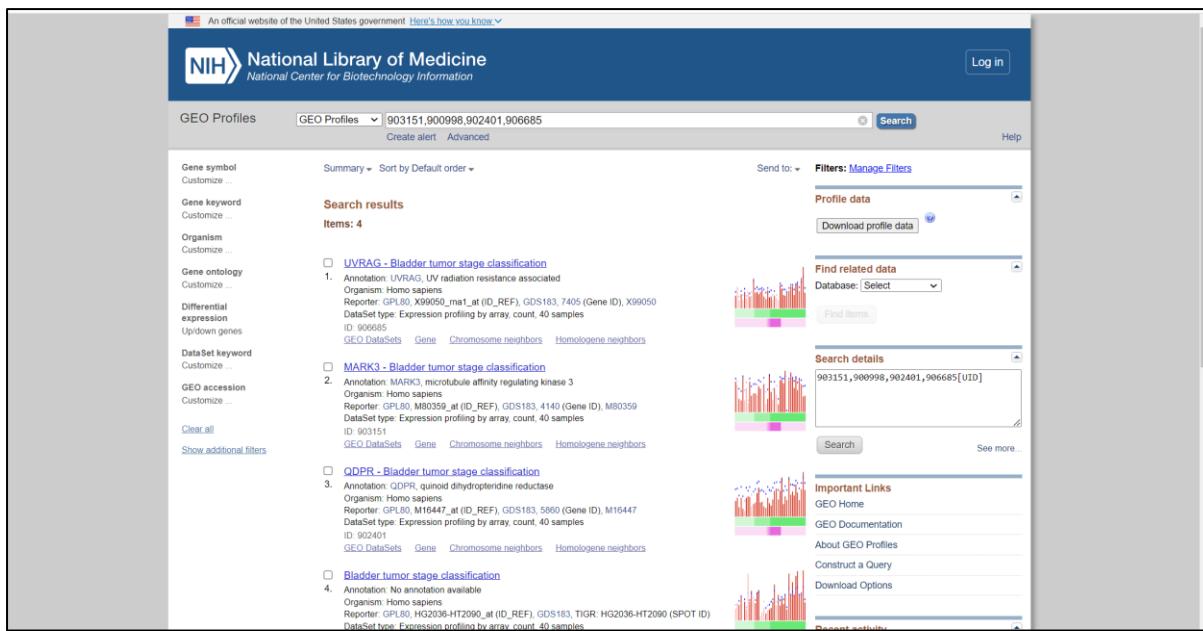


Fig 15: GEO Profiles in Entrez for the Selected Genes

RESULTS:

GEO Database showed differentially expressed genes for the query ‘Bladder tumor’. Structured data files containing gene expression measurements and associated metadata were retrieved for the query. It also provided visualization tools and links to external analysis tools to analyze the data. Results were obtained into two categories which were 4678 in GEO DataSets Database and 29412 in GEO Profiles Database.

CONCLUSION:

GEO Database was explored in order to understand differential gene expression for the query ‘Bladder tumor’.

REFERENCES:

1. What is skin cancer? | CDC. (n.d.). https://www.cdc.gov/cancer/skin/basic_info/what_is-skin-cancer.htm
2. Skin cancer (Including Melanoma)—Patient version. (n.d.). National Cancer Institute. <https://www.cancer.gov/types/skin>
3. Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. In Methods in molecular biology (pp. 93–110). https://doi.org/10.1007/978-1-4939-3578-9_5
4. *Bladder Cancer - Introduction*. (2022, June 7). Cancer.Net. <https://www.cancer.net/cancer-types/bladder-cancer/introduction>
5. *What Is Bladder Cancer?* (2023, February 16). National Cancer Institute. <https://www.cancer.gov/types/bladder>

DATE: 17/03/2024

WEBLEM 3
GPCR Database
(URL: <https://gpcrdb.org/>)

The GPCR database, GPCRdb was started in 1993 by Gert Vriend, Ad IJzerman, Bob Bywater and Friedrich Rippmann. Over two decades, GPCRdb evolved to be a comprehensive information system storing and analyzing data. In 2013, the stewardship of GPCRdb was transferred to the David Gloriam group at the University of Copenhagen, backed up by an international team of contributors and developers from the EU COST Action ‘GLISTEN’.

GPCRdb offers reference data and easy-to-use web tools and diagrams for a multidisciplinary audience investigating GPCR function, drug design or evolution. It stores a manual annotation of all GPCR crystal structures, the largest collections of receptor mutants and reference sequence alignments. The tools run directly in the web browser allowing for swift analysis of structures, sequence similarities, receptor relationships, and ligand target profiles. Diagrams illustrate receptor sequences (snake-plot and helix box diagrams) and relationships (phylogenetic trees). A visual overview can be seen in the GPCRdb poster.

The GPCRdb contains data, diagrams, and web tools for G protein-coupled receptors (GPCRs). Users can browse all GPCR crystal structures and the largest collection of receptor mutants. Diagrams can be produced and downloaded to illustrate receptor residues (snake-plot and helix box diagrams) and relationships (phylogenetic trees). Reference (crystal) structure-based sequence alignments taken into account helix bulges and constrictions, display statistics of amino acid conservation and have been assigned generic residue numbering for equivalent residues in different receptors.

GPCRs or G protein-coupled receptors are proteins located on the cell surface that recognize extracellular substances and transmit signals across the cell membrane. GPCRs do this by activating guanine nucleotide-binding proteins (G protein) that are responsible for signal transduction inside the cell. The signal delivery is very important in various cellular responses including cell growth, gene transcription, post-translational changes, and communication with other cells. This brings on adjustments in the body to adapt to the environmental changes such as increased heart rate when feeling threatened or change in vision in response to dim light. The human genome alone consists of at least 1,000 different GPCRs that detect hormones, lipids, amines, neurotransmitters, and light, among other things. A GPCR consists of three regions. The extracellular part detects and binds the ligand. Then, the seven-transmembrane region undergoes a conformational change.

OBJECTIVES COVERED UNDER (GPCR):

1. Sequence Alignment
2. To Retrieve Structures
3. Endogenous Ligand
4. Generating models
5. Receptor search
6. Mutation studies

1) Sequence Alignment:

1. **Structure-based alignments:** Users select receptors and sequence segments, then view an alignment with sequence and generic numbers. A consensus sequence and conservation statistics are displayed.
2. **Phylogenetic trees:** Users select receptors and sequence segments, then configure settings such as bootstrapping replicas and tree type. Trees are generated using PHYLP and jsPhyloSVG, with an option to view the alignment.
3. **Similarity search - GPCRdb:** Users select a reference receptor, sequence segments, and a comparison set. The tool computes similarities, displaying an alignment ranked by similarity with computed properties.
4. **Similarity search - BLAST:** Users submit sequences for a standard BLAST search against a custom database containing GPCRdb sequences. results show the best BLAST hits.
5. **Similarity matrix:** Users select receptors and sequence segments, then view a table of identities and similarities, color-coded for easy interpretation. Each tool warns about potential performance issues with large datasets. These tools collectively offer a comprehensive toolkit for GPCR sequence analysis, covering alignment, phylogenetics, similarity search, and similarity visualization.

2) To Retrieve Structures:

The GPCRdb database provides comprehensive resources for G protein-coupled receptor (GPCR) structures, including a structure browser, refined structures, statistics, structure models, validation, and descriptors. The structure browser allows sorting and filtering of annotated GPCR structures, while refined structures incorporate missing segments and revert mutations. Structure statistics offer insights into available structures by year, resolution, and phylogenetic trees. Structure models include homology models for human GPCRs in various activation states. Model validation employs root-mean-square deviation (RMSD) calculations. Activation state definitions rely on C α atom distances, while TM6 tilt measures assess receptor conformation. These resources aid in understanding GPCR structure-function relationships and are invaluable for research, with automated updates ensuring data currency and reliability.

3) Endogenous Ligand:

The Ligand Coverage section of GPCRdb provides an extensive overview of ligands associated with G protein-coupled receptors (GPCRs). A table displays ligand counts per receptor class, average ligands per GPCR, and the number of GPCRs with ligands. Phylogenetic trees visualize ligand data, with shaded inner circles indicating ligand counts per receptor. Users can explore ligands through receptor-based or ligand-based queries. Receptor-based queries allow selection of receptors and exploration of ligand bioactivities. Ligand-based queries enable searching by name, ID, or structure, leading to detailed ligand info pages. Additionally, an endogenous ligand browser offers data on endogenous ligands for human and rodent GPCRs. GPCRdb's Ligand Coverage section offers a user-friendly interface for comprehensive ligand exploration, serving both research and clinical needs.

4) Generating Models:

1. Structural data, such as crystallography or cryo-electron microscopy. These models serve as valuable tools for understanding the structural and functional properties of GPCRs, facilitating drug discovery and therapeutic development.
2. The generation of models typically involves several steps, including sequence alignment, homology modelling, and molecular dynamics simulations. Sequence alignment compares the amino acid sequences of the target receptor with known structures of related receptors, identifying conserved regions and structurally significant residues. Homology modelling then utilizes this alignment to predict the 3D structure of the target receptor based on the known structures of homologous proteins.
3. Molecular dynamics simulations refine the homology models by simulating the movement and interactions of atoms within the receptor structure, providing insights into its dynamic behavior and ligand binding properties. These computational models can be further validated and refined using experimental data, such as biochemical assays or mutagenesis studies.

5) Receptor Search:

1. In the GPCR database, users can perform receptor searches to access detailed information on G protein-coupled receptors (GPCRs). By navigating to the receptor search functionality, users can input specific receptor names, aliases, or identifiers to retrieve relevant data. The search results typically include comprehensive information about the selected receptors, such as their names, classifications, structures, functions, and associated ligands.
2. Furthermore, users may encounter additional features within the receptor search interface, allowing for advanced filtering options or customization of search parameters. These features enable researchers and clinicians to tailor their searches to specific criteria, such as receptor families, species, or functional properties.
3. The receptor search functionality in the GPCR database serves as a valuable tool for accessing detailed information on GPCRs, facilitating research, drug discovery, and therapeutic development in the field of pharmacology and molecular biology. It provides a user-friendly interface for efficiently retrieving and exploring data related to GPCR receptors, contributing to advancements in understanding their roles in various physiological processes and diseases.

6) Mutational Studies:

The GPCRdb features a Mutation Browser enabling users to explore mutant data for specific receptors or receptor families. Users first select receptors and sequence segments, then view mutants in a table that can be sorted and filtered. Helix box and snake plots below highlight mutated residues on consensus sequences. Additionally, a table displays all residues with mutations highlighted. The database continuously welcomes contributions of mutational data, encouraging researchers to submit their findings, including data on pharmacological effects. A standardized Excel spreadsheet facilitates data submission, focusing on effects like ligand binding, function, surface expression, basal activity, and Emax. Future includes

expanding data collection to mutations affecting thermo-stabilization, biased signaling, G-protein binding, and dimerization. The Mutation Browser allows users to search mutations by receptor, sub-family, or generic numbering position, offering visualizations like snake and helix box diagrams. These tools aid in understanding mutational effects on ligand binding and function, with future to integrate mutational data into 3D structures.

	GPCRdb(A)	GPCRdb(C)	GPRC5B	GPRC5C	GPRC5D	CaS receptor	GPRC ₆ receptor	GABA _{A1}	GABA _{A2}	mGlu ₁ receptor	mGlu ₂ receptor	mGlu ₃ receptor	mGlu ₄ receptor	mGlu ₅ receptor	mGlu ₆ receptor	mGlu ₇ receptor	mGlu ₈ receptor	TAS1R1	TAS1R2	TAS ⁺
TM1																				
1x35	1.39x39	W55	W49	W22	F612	L593	L591	L481	E592	W567	W576	W587	E579	W585	W590	W583	T566	P565	A568	
1x37	1.41x41	I57	I51	I24	I614	I595	I593	S483	I594	V569	I578	V589	I581	A587	V592	V585	W568	I567	L570	
1x39	1.43x43	V59	L53	L26	L616	L597	V595	L485	A596	P571	P580	P591	A583	P589	P594	P587	L570	V569	L572	
1x42	1.46x46	V62	V56	L29	F619	L600	L598	L488	F599	I574	I583	L594	F586	L592	L597	V590	A573	L572	L575	
TM2																				
2x42	2.38x38	L95	T89	T62	L650	V632	L629	M519	L630	L605	L614	L625	L617	L623	L628	L621	L604	M603	L606	
2x49	2.45x45	G102	G96	S69	S657	C639	G636	G526	G637	G612	G621	G632	G624	G630	G635	G628	S611	L610	C613	
2x53	2.49x49	L106	L100	L73	C661	N643	A640	S530	G641	C616	S625	C636	G628	I634	C639	C632	G615	A614	V617	
2x54	2.50x50	F107	F101	F74	F662	F644	L641	Y531	Y642	Y617	Y626	Y637	Y629	Y635	Y640	Y633	S616	Y615	C618	
2x55	2.51x51	G108	C102	G75	S663	A645	A642	A532	V643	C618	C627	A638	L630	A636	I641	S634	G617	M616	L619	
2x56	2.52x52	L109	L103	L76	S664	S646	A643	S533	C644	M619	M628	T639	C631	I637	I642	I635	S618	V617	S620	
2x57	2.53x53	T110	V104	A77	S665	T647	V644	I534	P645	T620	T629	T640	T632	T638	T643	T636	L619	V618	V621	
2x59	2.55x55	A112	A106	A79	F667	F649	P646	L536	T647	I622	F631	L642	C634	L640	L645	L638	G621	V620	L623	
2x60	2.56x56	F113	C107	F80	F668	F650	L647	F537	L648	F623	F632	M643	L635	M641	M646	M639	F622	Y621	F624	
TM3																				
3x27	3.31x31	V124	S118	V91	L679	T661	A665	V555	L659	L634	L643	L654	L646	A652	F657	F650	L633	C632	A635	
3x28	3.32x32	R125	R119	R92	R680	R662	R666	R556	Q660	R635	R644	R655	Q647	R653	R658	R651	R634	R633	Q636	
3x29	3.33x33	R126	R120	Y93	Q681	Q663	L667	T557	R661	R636	R645	R656	R648	R654	R659	R652	Q635	Q634	Q637	
3x30	3.34x34	F127	F121	F94	P682	T664	W668	W558	L662	L637	L646	I657	I649	L655	V660	V653	A636	A635	P638	
3x31	3.35x35	L128	L122	L95	A683	M665	L669	I559	L663	G638	G647	F658	G650	F656	F661	F654	L637	L636	L639	

Fig 1: Sequence comparison of the 7TM domain binding pocket in the eight mGlu receptor subtypes with all residues that have been mutated

Color-coding: Green indicates increased binding/potency of >5-fold (light green) or >10-fold (dark green), red indicates reduced binding/potency of >5-fold (pink) or >10-fold(red), yellow indicates No/low effect (<5-fold), and grey indicates that no effect is annotated. The first two columns show generic GPCRdb residue numbers for each row of residues

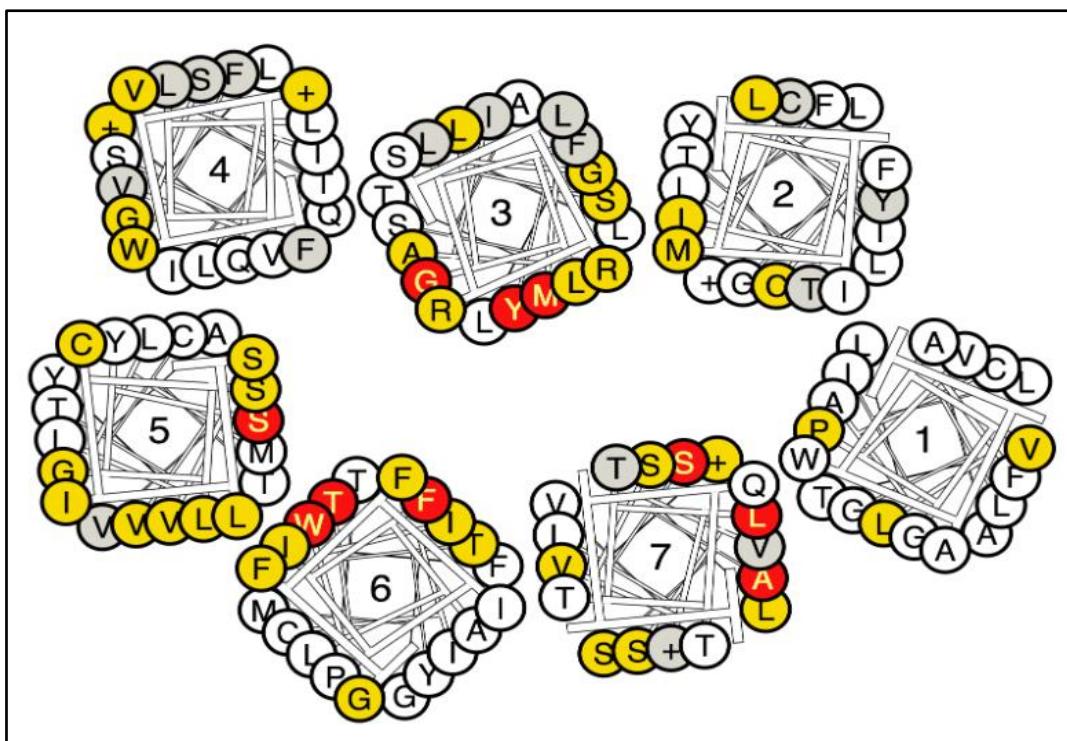


Fig 2: A helix box diagram of the metabotropic glutamate receptors displaying mutated residue positions from the extracellular side with all residues that have been mutated

Color-coding: Green indicates increased binding/potency of >5-fold (light green) or >10-fold (dark green), red indicates reduced binding/potency of >5-fold (pink) or >10-fold (red), yellow indicates No/low effect (<5-fold), and grey indicates that no effect is annotated

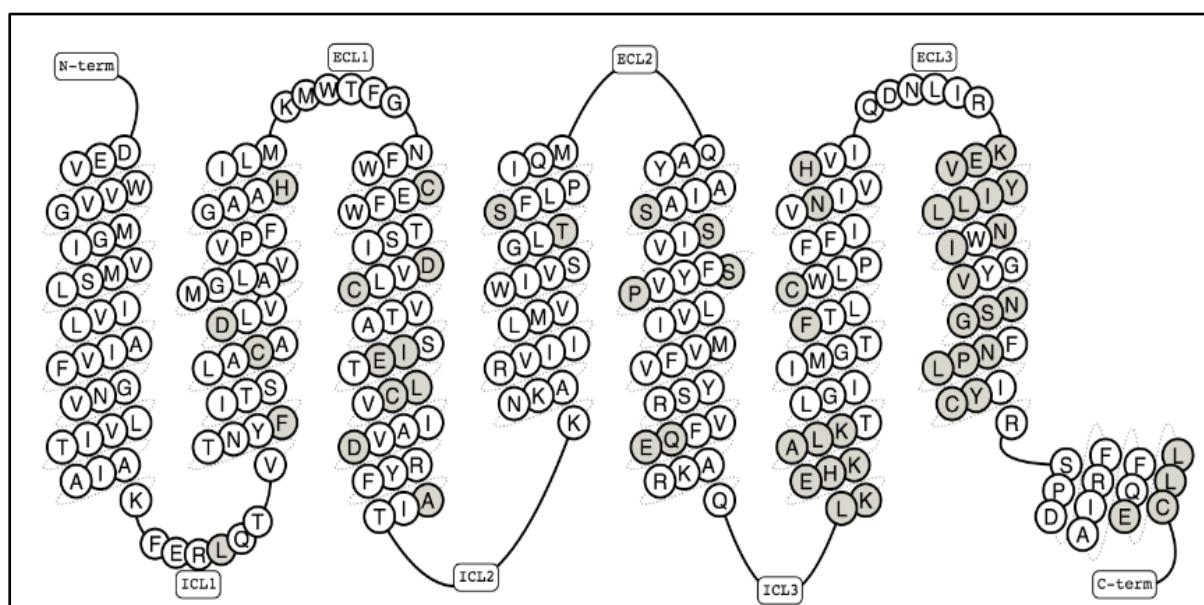


Fig 3: Snake diagram of the human β_2 -adrenoceptor showing all residues (grey) for which mutational experiments have been deposited in the GPCRdb

The screenshot shows the GPCRdb homepage with the following sections:

- Header:** Info, GPCRdb, Drugs & ligands, GproteinDb, ArrestinDb, Biased Signaling Atlas, Join us, Quick search.
- Top Left:** A detailed 3D structure of the 5-HT_{1A} receptor.
- Top Center:** A table of receptor models, including columns for ID, Date, Model, Species, Receptor, and Template.
- Top Right:** A refined structure of the 3RZE receptor.
- Middle Left:** A section titled "New state-specific structure models" featuring a 3D structure of a receptor.
- Middle Center:** A section titled "Models and refined structures using AlphaFold2" with a progress bar.
- Middle Right:** A "Related resources" section with links to GPCRdb, GproteinDb, ArrestinDb, and Biased Signaling Atlas.
- Bottom Left:** A "GPCRdb offers" section with a list of features: Reference data, Analysis tools, Visualization, Experiment design, Data deposition, and Overview figure and references.
- Bottom Center:** Three boxes: "Data" (424 Human proteins, 40,450 Species orthologs), "Latest release" (Feb. 12, 2024, New structures), and "Publications" (GPCRdb in 2023 State-specific structure models using AlphaFold2 and expansion of ligand).
- Bottom Right:** A "Tweets from @gpcrdb" section.

Fig 4: Homepage of GPCR Database

REFERENCES:

1. Rosenbaum DM, Rasmussen SG, Kobilka BK. The structure and function of G-protein-coupled receptors. *Nature*. 2009 May 21;459(7245):356-63. doi: 10.1038/nature08144. PMID: 19458711; PMCID: PMC3967846.
2. Cheng, L., Xia, F., Li, Z. et al. Structure, function and drug discovery of GPCR signaling. *Mol Biomed* 4, 46 (2023). <https://doi.org/10.1186/s43556-023-00156-w>
3. Levoye A, Dam J, Ayoub MA, Guillaume JL, Jockers R. Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Rep*. 2006 Nov;7(11):1094-8. doi: 10.1038/sj.embor.7400838. PMID: 17077864; PMCID: PMC1679777.

DATE: 17/03/2024

WEBLEM 3(A)
GPCR DATABASE
(URL: <https://gpcrdb.org/>)

AIM:

To explore structure and functional characteristics of query ‘PTH1 receptor’ by using GPCR database.

INTRODUCTION:

The GPCR database, GPCRdb was started in 1993 by Gert Vriend, Ad IJzerman, Bob Bywater and Friedrich Rippmann. Over two decades, GPCRdb evolved to be a comprehensive information system storing and analyzing data. In 2013, the stewardship of GPCRdb was transferred to the David Gloriam group at the University of Copenhagen, backed up by an international team of contributors and developers from the EU COST Action ‘GLISTEN’.

GPCRdb offers reference data and easy-to-use web tools and diagrams for a multidisciplinary audience investigating GPCR function, drug design or evolution. It stores a manual annotation of all GPCR crystal structures, the largest collections of receptor mutants and reference sequence alignments. The tools run directly in the web browser allowing for swift analysis of structures, sequence similarities, receptor relationships, and ligand target profiles. Diagrams illustrate receptor sequences (snake-plot and helix box diagrams) and relationships (phylogenetic trees). A visual overview can be seen in the GPCRdb poster.

The GPCRdb contains data, diagrams and web tools for G protein-coupled receptors (GPCRs). Users can browse all GPCR crystal structures and the largest collection of receptor mutants. Diagrams can be produced and downloaded to illustrate receptor residues (snake-plot and helix box diagrams) and relationships (phylogenetic trees). Reference (crystal) structure-based sequence alignments take into account helix bulges and constrictions, display statistics of amino acid conservation and have been assigned generic residue numbering for equivalent residues in different receptors.

PTH1:

PTHR1, or parathyroid hormone receptor-1, encodes a G protein-coupled receptor pivotal in bone biology, mediating the actions of parathyroid hormone (PTH) and PTH-related protein (PThrP). Its roles in regulating bone formation and resorption make it a promising therapeutic target for osteoporosis treatment. PTHR1 exhibits distinct high-affinity conformations, R(0) and RG, with ligand analogs capable of selectively binding to these conformations, influencing signaling responses and calcemic effects. Abaloparatide, a synthetic PThrp analog, demonstrates potential as an osteoporosis therapy by stimulating bone formation with reduced bone resorption and hypercalcemic effects compared to PTH (1-34), attributed to its selective binding to the RG conformation of PTHR1.

Beyond bone biology, PTHR1 involvement extends to cartilage differentiation regulation. Indian hedgehog (Ihh) regulates hypertrophic differentiation rate in cartilage elements by modulating PThrp expression in the perichondrium, thereby blocking hypertrophic

differentiation in prehypertrophic chondrocytes. This mechanism forms a negative feedback loop that regulates chondrocyte differentiation rate, highlighting the intricate interplay between PTHR1, Ihh, and PTHrP in skeletal development and homeostasis. These insights deepen our understanding of PTHR1's multifaceted roles in bone and cartilage biology, offering potential avenues for therapeutic interventions in skeletal disorders.

METHODOLOGY:

1. Open homepage of GPCR database.
2. Search for query 'PTH1'.
3. Under GPCRdb option select sequence alignment, structure models and interpret the results.
4. Under Drugs and ligand option select endogenous ligand (GtP), Mutations and interpret the results.

OBSERVATIONS:

The screenshot shows the GPCRdb homepage with a search query 'bkrb1' entered. The results are displayed under the 'Receptors' heading, specifically for the 'B1 receptor [Human]'. On the left side, there are two highlighted receptor structures: '5-HT_{1A} receptor structure model' and '3RZE refined structure'. The central part of the page contains a summary of what GPCRdb offers, including reference data, analysis tools, visualization, experiment design, data deposition, and an overview figure and references. Below this, there are four main sections: 'Data' (listing 424 Human proteins, 40,450 Species orthologs, 69,580 Genetic variants, and 968 Drugs), 'Latest release' (dated Feb. 12, 2024, with a note about new structures and minor bug fixes), 'Publications' (mentioning GPCRdb's 2023 state-specific structure models using AlphaFold2 and expansion of ligand resources), and 'Related resources' (linking to GPCRdb, GproteinDb, ArrestinDb, and Biased Signaling Atlas). A sidebar on the right shows 'Tweets from @gpcrdb'.

Fig 1: Homepage of GPCR database

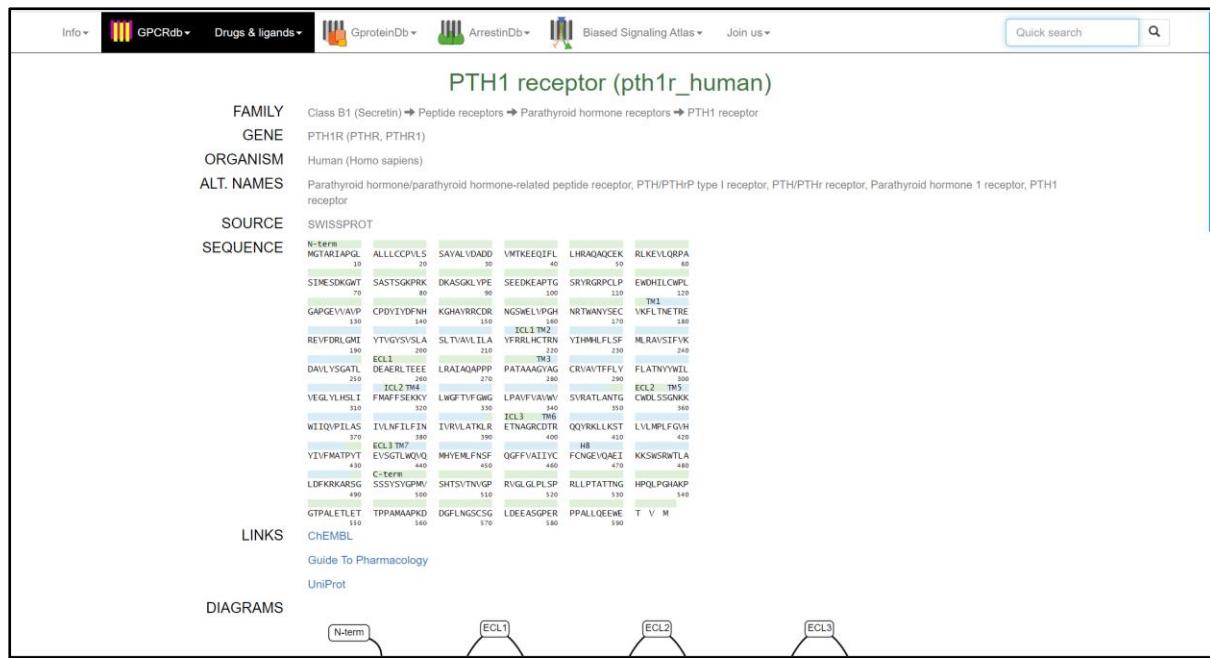


Fig 2: Result page for query 'PTH1 receptor'

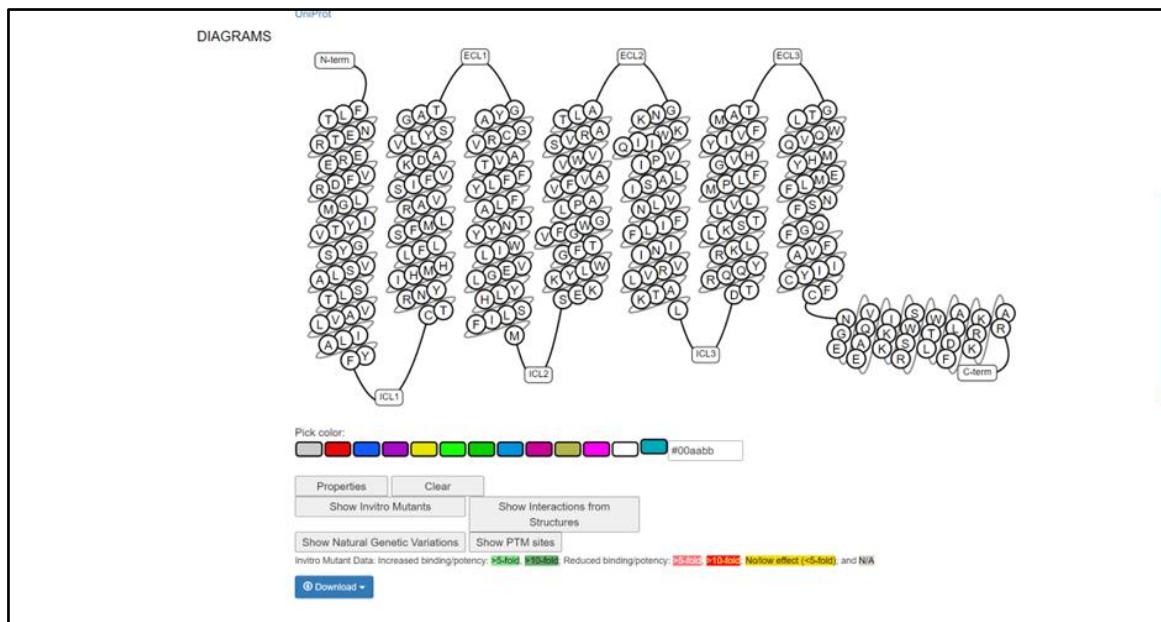


Fig 3: Snake diagram

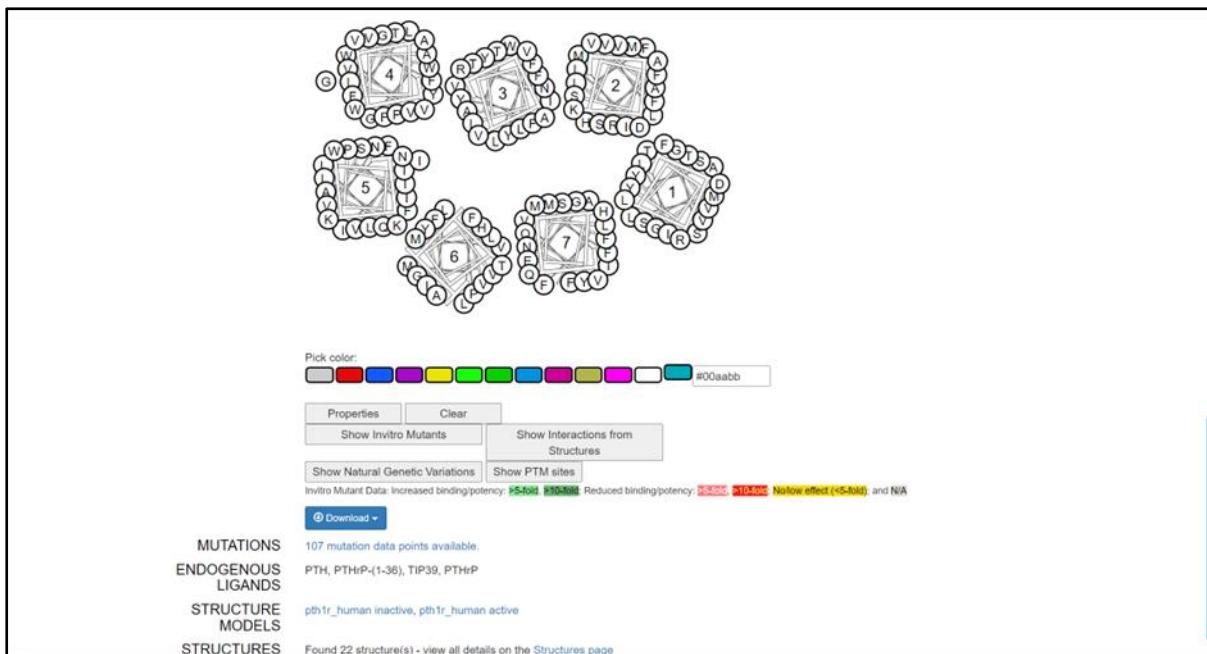


Fig 4: Necklace diagram

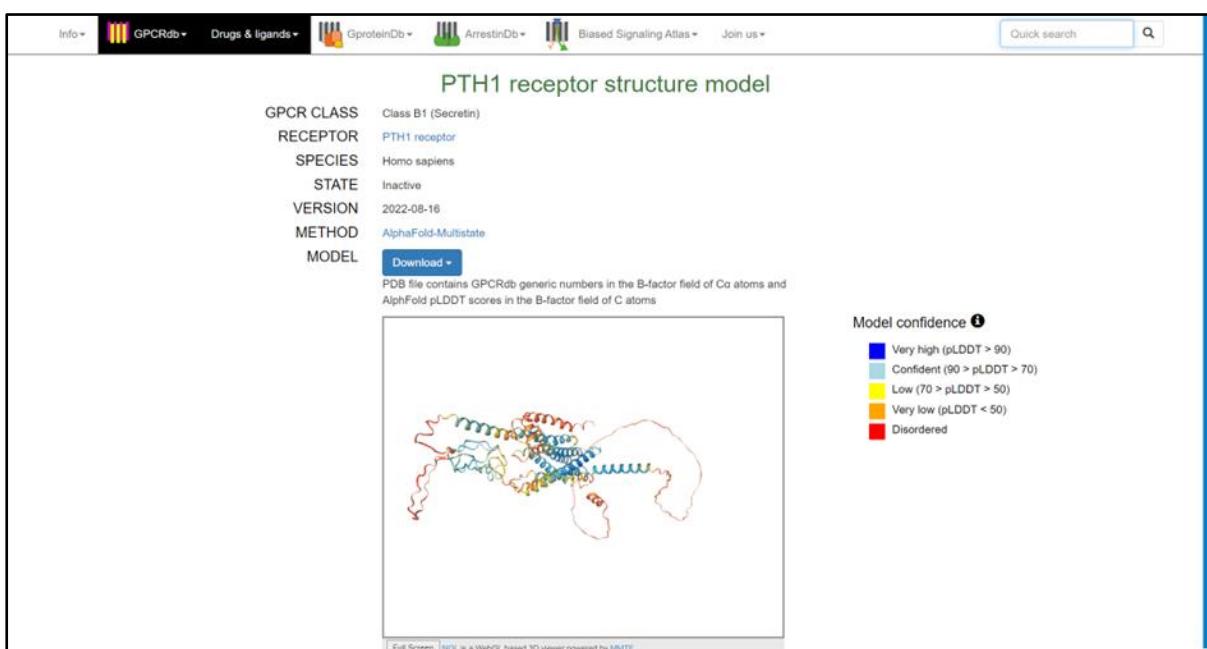


Fig 5: ‘PTH1 receptor’ structure model



Fig 6: Result for sequence alignment

This table lists endogenous ligands for various GPCRs, categorized by receptor class. The columns include RECEPTOR, ENDOGENOUS LIGANDS, pEC50, pKi, and REFERENCES. The table is filtered to show results for B1 (Secretin) receptors.

RECEPTOR	ENDOGENOUS LIGANDS	pEC50	pKi	REFERENCE...
B1 (Secretin) VIP and PA...	PACR PAC ₁	Human Phv	4	Peptide
B1 (Secretin) VIP and PA...	PACR PAC ₁	Human Phm	4	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1	Human Pth	3	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1	Human Pth	3	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1	Human Pth	3	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1 Human Pthr-(1-36)	Human Pthr-(1-36)	2	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1 Human Pth	Human Pth	3	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1 Human Tgfb	Human Tgfb	1	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH1 Human Tgfb	Human Tgfb	2	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Tgfb	Human Tgfb	1	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Pth	Human Pth	5	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Pthr	Human Pthr	4	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Pthr-(1-34)...	None	3	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Pth	Human Pth	5	Peptide
B1 (Secretin) Parathyroid...	PTHR PTH2 Human Pthr	Human Pthr	5	Peptide
B1 (Secretin) Glucagon re...	SCTR secretin	Human Vip	2	Peptide
B1 (Secretin) Glucagon re...	SCTR secretin	Human Secretin	2	Peptide
B1 (Secretin) Glucagon re...	SCTR secretin	Human Secretin	1	Peptide

Fig 7: Result for Endogenous ligand

glr_human	7.38x47	381	TM7	K => L	Abolished	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.39x38	382	TM7	L => A	3.663↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.39x38	382	TM7	L => V	5.348↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.40x39	383	TM7	F => A	3.175↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.41x40	384	TM7	F => T	3.584↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.43x42	386	TM7	L => F	Abolished	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.43x42	386	TM7	L => A	9.091↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.43x42	386	TM7	L => V	2.37↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.44x43	387	TM7	F => S	2.342↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.46x45	389	TM7	S => A	1.499↓	IC50 (Binding - Radioligand competition/displacement)	
glr_human	7.46x45	389	TM7	S => T	1.605↓	IC50 (Binding - Radioligand competition/displacement)	
pth1r_rat	2.60x60	233	TM2	R => H	N/A	N/A	
pth1r_rat	2.60x60	233	TM2	R => Q	N/A	N/A	
pth1r_rat	2.60x60	233	TM2	R => N	N/A	N/A	
pth1r_rat	2.60x60	233	TM2	R => K	N/A	N/A	
pth1r_rat	7.49x49	451	TM7	Q => K	N/A	N/A	
gip1r_rat	8.57x57	416	H8	S => A	N/A	N/A	
pth1r_human	2.50x50	223	TM2	H => R	N/A	N/A	
ghfrh_human	1.39x39	129	TM1	V => L	N/A	N/A	
vipr1_rat	1.39x39	143	TM1	V => L	N/A	N/A	
ghfrh_human	1.39x39	129	TM1	V => L	N/A	N/A	
vipr1_human	3.33x33	219	TM3	M => V	N/A	N/A	
scl_rat	2.50x50	178	TM2	H => R	N/A	N/A	

Fig 8: Result for Mutations

RESULTS:

GPCR database were explored. Structural and functional characteristics of query ‘PTH1 receptor’ were observed. Sequence alignments, endogenous ligands, snake diagram, necklace diagram, structure models and mutations for query calcitonin were observed and studied.

CONCLUSION:

Structural and functional characteristics of query ‘PTH1 receptor’ was studied by exploring GPCR database.

REFERENCES:

1. Kooistra, A. J., Mordalski, S., Pády-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., Keserű, G. M., & Gloriam, D. E. (2021). GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic acids research*, 49(D1), D335–D343. <https://doi.org/10.1093/nar/gkaa1080>
2. Albert J Kooistra, Stefan Mordalski, Gáspár Pády-Szekeres, Mauricio Esguerra, Alibek Mamyrbekov, Christian Munk, György M Keserű, David E Gloriam, GPCRdb in 2021: integrating GPCR sequence, structure and function, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D335–D343, <https://doi.org/10.1093/nar/gkaa1080>
3. Cheloha, R. W., Gellman, S. H., Vilardaga, J. P., & Gardella, T. J. (2015, August 25). PTH receptor-1 signalling—mechanistic insights and therapeutic prospects. *Nature Reviews Endocrinology*, 11(12), 712–724. <https://doi.org/10.1038/nrendo.2015.139>
4. Hattersley, G., Dean, T., Corbin, B. A., Bahar, H., & Gardella, T. J. (2016, January 1). Binding Selectivity of Abaloparatide for PTH-Type-1-Receptor Conformations and Effects on Downstream Signaling. *Endocrinology*, 157(1), 141–149. <https://doi.org/10.1210/en.2015-1726>
5. Vortkamp, A., Lee, K., Lanske, B., Segre, G. V., Kronenberg, H. M., & Tabin, C. J. (1996, August 2). Regulation of Rate of Cartilage Differentiation by Indian Hedgehog and PTH-Related Protein. *Science*, 273(5275), 613–622. <https://doi.org/10.1126/science.273.5275.613>

WEBLEM 4

INTRODUCTION TO EXPRESSED SEQUENCE TAGS DATABASE

Expressed Sequence Tags are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene. The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "tags" to fish a gene out of a portion of chromosomal DNA by matching base pairs. The challenge associated with identifying genes from genomic sequences varies among organisms and is dependent upon genome size as well as the presence or absence of introns--the intervening DNA sequences interrupting the protein coding sequence of a gene.

ESTs and NCBI:

Expressed Sequence Tags, or ESTs, changed genomics because they were quick, cheap, and efficient. Scientists have produced hundreds of thousands of ESTs for public usage very quickly, including small-scale genome sequencing institutes and individual researchers. Nevertheless, the enormous number of ESTs being produced made it difficult to efficiently organize and retrieve this important data.

The National Institutes of Health (NIH) division, the National Center for Biotechnology Information (NCBI), realized the necessity of an EST-specific database in order to address this issue. In 1992, researchers at the NCBI created dbEST in response. This database was created especially to act as an EST central repository, making it simple for researchers to access and make use of this abundance of genetic data.

Before being deposited into dbEST, ESTs were subjected to screening and annotation procedures following submission to GenBank, the NIH's main sequence database. As a result, the database was guaranteed to have excellent, fully annotated EST sequences that were prepared for investigation and examination. The NCBI gave researchers an effective tool for genomic and gene discovery by centralizing EST data in dbEST.

Moreover, dbEST was more than just an EST repository; it was included into the larger ecosystem of genome data that the NCBI maintained. This integration made it easier to conduct thorough searches across a variety of genomic information types, which improved researchers' capacity to examine and interpret genetic data from a variety of perspectives.

dbEST: a descriptive catalog of ESTs

To arrange, preserve, and make accessible the vast amount of publicly available EST data that has previously been collected and is still growing every day, scientists at NCBI developed dbEST. A scientist can obtain information on ESTs from more than 300 different organisms, in addition to human ESTs, by using dbEST. Scientists at NCBI annotate the EST record with any information that they know, wherever possible. For instance, the name and function of a known gene are recorded on the EST record if an EST matches a DNA sequence that codes for that gene. Public scientists can utilize dbEST as a gene discovery tool by annotating EST

records. Using a database search engine like NCBI's BLAST, anybody with an interest can do sequence similarity searches against dbEST.

UniGene: an indistinguishable collection of gene-focused clusters

ESTs that are ultimately produced from mRNA may be redundant since a gene can express itself as mRNA several times. That is to say, multiple copies of the same EST that are identical or comparable may exist. Because of this redundancy and overlap, a user searching dbEST for a specific EST may get a lengthy list of tags, many of which could be for the same gene. Going through each of these similar ESTs can take a lot of time. UniGene was created by NCBI researchers as a solution to the redundancy and overlap issue.

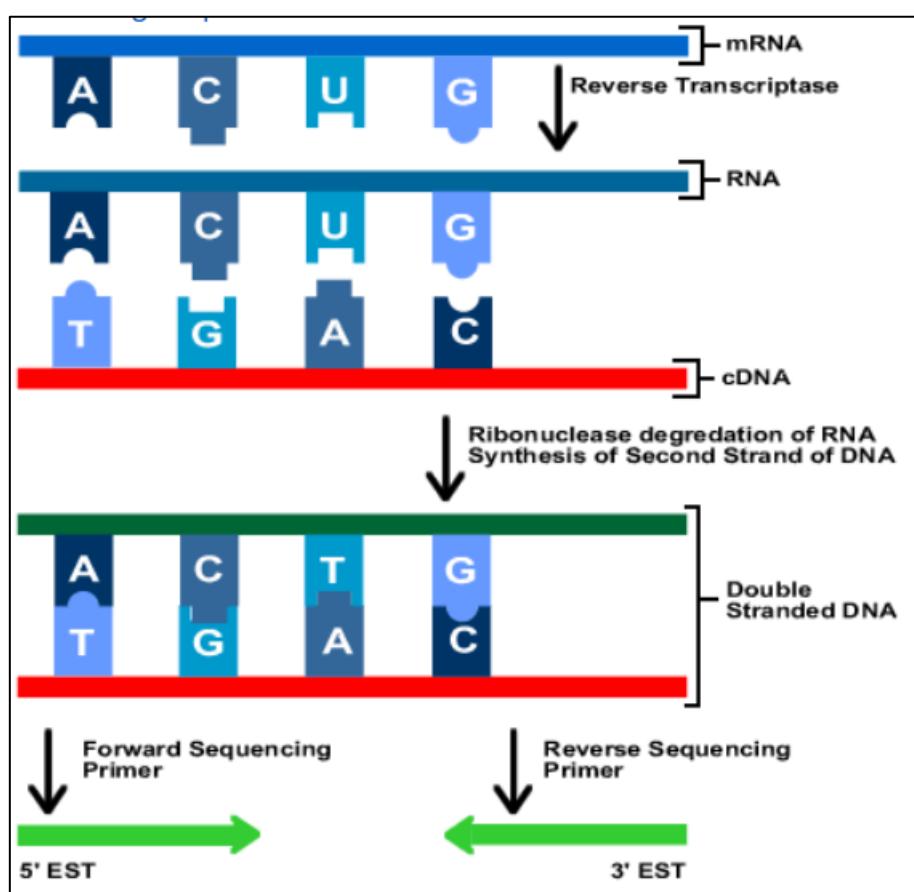


Fig 1: An overview of how Expressed Sequence Tags is generated

GenBank's "Expressed Sequence Tags" (or "single-pass" cDNA sequences) from various organisms are the subject of sequence data and other material in the EST Database (Nature Genetics 4:332-3;1993) subdivision. A synopsis of the development of human ESTs in GenBank can be found in Trends Biochemistry. Expressed Sequence Tags (ESTs) are single-pass sequence readings from mRNA (cDNA) that are brief (often less than 1000 bp). They are usually made in big quantities. They provide an image of the genes expressed in a particular tissue and/or at a particular stage of development. These are expression tags for the relevant cDNA library, some of which are coding and some of which are not.

The majority of EST projects produce a lot of sequences. These are typically uploaded in batches of dozens to thousands of entries, with shared submitter, source, and citation data, to GenBank and dbEST.

1. **Gene Discovery:** EST databases have proven invaluable in the search for new genes, particularly in organisms for which a complete genome sequence may not yet be accessible or adequately annotated. Researchers are able to discover new genes, examine their structures, and deduce their activities by sequencing segments of expressed genes.
2. **Gene Expression Analysis:** When examining patterns of gene expression in various tissues, developmental stages, or experimental settings, EST databases are invaluable resources. Through comparing the quantity of ESTs obtained from various libraries, scientists can learn which genes are expressed in particular biological circumstances.
3. **Comparative Genomics:** By offering sequences from several species, EST databases can also help in comparative genomics research. It is possible to determine conserved genes and regulatory elements as well as comprehend evolutionary changes in gene expression patterns by comparing ESTs from related animals.
4. **Functional Genomics:** By offering sequence data for relevant genes, EST databases can support functional genomics research. Utilizing ESTs, scientists can validate gene predictions from genome sequencing projects, create probes for gene expression microarrays, and investigate how genetic diversity affects gene expression.
5. **Database Resources:** There are numerous EST databases that are kept up to date by different organizations and academic institutions. ESTScan, TIGR Gene Indices, and dbEST (a component of the NCBI's Entrez system) are a few examples. Researchers from anywhere around the world can access EST data through these databases, which usually offer tools for browsing, searching, and analysis.

REFERENCES:

1. ESTs. (n.d.). <http://www.cyto.purdue.edu/cdroms/cyto6/content/primer/est.htm>
 2. What is dbEST? (n.d.). <https://www.ncbi.nlm.nih.gov/genbank/dbest/>
-

DATE: 16-03-2024

WEBLEM 4(A)
EST DATABASE

(URL: <https://www.ncbi.nlm.nih.gov/genbank/dbest/>)

AIM:

To identify and characterize gene expression patterns for query 'DDIT3' (Accession ID: DN990078.1) using an EST database.

INTRODUCTION:

GenBank's "Expressed Sequence Tags" (or "single-pass" cDNA sequences) from various organisms are the subject of sequence data and other material in the EST Database (Nature Genetics 4:332-3;1993) subdivision. A synopsis of the development of human ESTs in GenBank can be found in Trends Biochemistry. Expressed Sequence Tags (ESTs) are single-pass sequence readings from mRNA (cDNA) that are brief (often less than 1000 bp). They are usually made in big quantities. They provide an image of the genes expressed in a particular tissue and/or at a particular stage of development. These are expression tags for the relevant cDNA library, some of which are coding and some of which are not.

The majority of EST projects produce a lot of sequences. These are typically uploaded in batches of dozens to thousands of entries, with shared submitter, source, and citation data, to GenBank and dbEST.

1. Gene Discovery: EST databases have proven invaluable in the search for new genes, particularly in organisms for which a complete genome sequence may not yet be accessible or adequately annotated. Researchers are able to discover new genes, examine their structures, and deduce their activities by sequencing segments of expressed genes.
2. Gene Expression Analysis: When examining patterns of gene expression in various tissues, developmental stages, or experimental settings, EST databases are invaluable resources. Through comparing the quantity of ESTs obtained from various libraries, scientists can learn which genes are expressed in particular biological circumstances.
3. Comparative Genomics: By offering sequences from several species, EST databases can also help in comparative genomics research. It is possible to determine conserved genes and regulatory elements as well as comprehend evolutionary changes in gene expression patterns by comparing ESTs from related animals.
4. Functional Genomics: By offering sequence data for relevant genes, EST databases can support functional genomics research. Utilizing ESTs, scientists can validate gene predictions from genome sequencing projects, create probes for gene expression microarrays, and investigate how genetic diversity affects gene expression.
5. Database Resources: There are numerous EST databases that are kept up to date by different organizations and academic institutions. ESTScan, TIGR Gene Indices, and dbEST (a component of the NCBI's Entrez system) are a few examples. Researchers from anywhere around the world can access EST data through these databases, which usually offer tools for browsing, searching, and analysis.

DDIT3 gene:

The DDIT3 gene, also known as CHOP or GADD153, encodes a member of the CCAAT/enhancer-binding protein (C/EBP) family of transcription factors. Located on chromosome 12, it regulates diverse biological processes such as transcriptional regulation, blood vessel maturation, and cellular responses to various stressors like DNA damage and endoplasmic reticulum stress. Its involvement extends to cell cycle regulation, Wnt signaling, and modulation of interleukin-8 production, among others.

Functionally, DDIT3 participates in transcriptional regulation by binding to specific DNA sequences near RNA polymerase II promoters and modulating gene expression. It acts as a transcriptional activator or repressor depending on the context, exerting its effects through protein-protein interactions and DNA binding. Moreover, DDIT3 plays roles in cellular processes such as protein binding, homodimerization, and interaction with other transcription factors like cAMP response element-binding protein (CREB). Its leucine zipper domain facilitates dimerization, essential for its transcriptional activity. Overall, DDIT3's multifaceted functions highlight its significance in cellular homeostasis and stress response pathways.

METHODOLOGY:

1. Launch a browser and go to the NCBI website.
2. By selecting the "All Databases" dropdown menu at the top of the NCBI webpage, you can gain access to the EST database. To search nucleotide sequences, including ESTs, choose "Nucleotide" from the dropdown menu.
3. Search query 'DDIT3'.
4. On the left-hand side of the screen, use filter EST to screen for ESTs for your query.
5. Interpret the results displayed for query 'DDIT3'.

OBSERVATIONS:

The screenshot shows the NCBI homepage with a blue header containing the NIH logo and "National Library of Medicine" text. A search bar with dropdown options for "All Databases" and "Search" is at the top right. On the left, a sidebar lists categories like "NCBI Home", "Resource List (A-Z)", and "All Resources". The main content area features sections for "Welcome to NCBI", "Submit", "Download", "Learn", "Develop", "Analyze", and "Research". Each section has a brief description and a corresponding icon. To the right, there's a "Popular Resources" sidebar with links to PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below that is a "NCBI News & Blog" section with recent posts. At the bottom right, there's a "Now Available: RefSeq Release 223" notice.

Fig 1: NCBI Homepage

This screenshot is identical to Fig 1, but the "Nucleotide" link in the "All Databases" dropdown menu is highlighted with a red box. The rest of the interface, including the sidebar, main content, and news section, remains the same.

Fig 2: Select 'Nucleotide' option from all Databases

The screenshot shows the Nucleotide search interface. The search term 'DDIT3' is highlighted in a red box in the search bar. The results page displays information about the gene 'DDIT3 - DNA damage inducible transcript 3'. It includes links to orthologs, genome data viewer, and BLAST. A sidebar on the left provides filtering options for species, molecule types, source databases, sequence type, genetic compartments, sequence length, release date, revision date, and clear all filters. The right sidebar includes sections for filters, related data, search details, and recent activity.

Fig 3: Results obtained after searching for the query ‘DDIT3’

The screenshot shows the Nucleotide search interface for the EST entry 'TC117581'. The accession ID 'DN990078.1' is highlighted in the search bar. The results page displays detailed information about the EST entry, including its definition (TC117581 Human adult whole brain, large insert, pCMV expression library Homo sapiens cDNA clone TC117581 5' similar to Homo sapiens DNA-damage-inducible transcript 3 (DDIT3), mRNA sequence), sequence details, and related information. The right sidebar includes sections for change region shown, customize view, analyze this sequence, run BLAST, pick primers, find in this sequence, related information, and recent activity.

Fig 4: EST entry selected for further study: ‘DDIT3’ (Accession ID: DN990078.1)

COMMENT Contact: Kovacs, KF
 High Throughput cDNA Cloning
 Origene Technologies, Inc. (www.origene.com)
 6 Taft Court, Suite 100, Rockville, MD 20850, USA
 Tel: 301 348 3188
 Fax: 301 348 8689
 Email: cDNA@origene.com
 This EST submission is part of an on-going human full-length cloning project at Origene Technologies, Inc.
 Please contact Origene for access.
 Origene Technologies, Inc.
 6 Taft Ct. Suite 100
 Rockville, MD 20850
 Tel: (301) 348-3188
<http://www.origene.com>
 Seq primer: pCMV6 Sprime forward vector primer, Origene
 Technologies Inc.

FEATURES

source	Location/Qualifiers
	1..796
	/organism="Homo sapiens"
	/mol_type="mRNA"
	/db_xref="taxon:9606"
	/clone="TC117581"
	/tissue_type="whole brain"
	/clone_lib="SAWNB0176259 Human adult whole brain, large insert, pCMV expression library"
	/note="Organ: Brain; Vector: pCMV-XL5; Site_1: EcoRI; Site_2: NotI/SalI compatible; Ligation: Oligo-dT primed reverse transcriptase optimized for large size GC rich mRNA transcripts, cDNA size selection optimized ligation for large inserts into mammalian expression vector, random clones selected for end sequence verification of full-length genes"

ORIGIN

```

1  gacgagggtt cccatgtact cggggggcg aggccagaaga acatctggac cgaggaggca
61  gaggttgtcg agtgccggata tgcaccact gaaccttcgg ctggccacgg acaagactt
121  ggctccaaaa aaaaaaaaaaa aaaaaaaaaa agaaaaaaa aagaatgttgc cttctcccc
181  ttccaaaaat gggtgcattt tttttttttt gccccacatgt ttcaaaaggaa aatgttatctt
241  tcatacatca ccacacccca aagccatgtt gcctttccat acgtatccaa ctgcagatgt
301  ggcacgtgg tcattgcattt ttccttcgg gacacttgc acgtggggac ttgaaggctg
361  gtatggggac ctgcacagggtt tcctgttttc agataaaaat gggggtaactt atgttttacc
421  tcctggaaat ctggggggaa aatccaaaaat ttccacactt ttggacccctt ctttttttggc
481  tttggctcatt gggggggggc cagggccacgg aagggttaca agacccccc agggccctca
541  ctctccatgt ttccatgttca gttcccttcgg tcaggagaa gggggggaa accaaggggaa
601  aaccagggtttt cggaaacaca gttgtttatcc cccacggccgg gctggggaaacg acggccatgtaa
661  agaaaaaaatc agggaaatgtt aaggaaatgtt gcacacgtt ctggaa

```

Fig 4.1: Comment and Features for ‘DDIT3’ (Accession ID: DN990078.1)

RESULTS:

Following results were obtained for the query 'DDIT3' (Accession ID: DN990078.1) gene which is a 706bp mRNA linear sequence having only 1 EST tag.

CONCLUSION:

The EST (Expressed Sequence Tag) database makes it easier to identify and describe the genes that are expressed in a variety of organisms. EST databases provide information on gene expression patterns, alternative splicing variations, and tissue-specific expression profiles by supplying brief sequences obtained from mRNA transcripts.

REFERENCES:

1. National Center for Biotechnology Information. (n.d.). <https://www.ncbi.nlm.nih.gov/>
 2. What is dbEST? (n.d.). <https://www.ncbi.nlm.nih.gov/genbank/dbest/>
 3. Diaz-Perez, J. A., & Kerr, D. A. (2023, November 22). Gene of the month: DDIT3. *Journal of Clinical Pathology*, 77(4), 211–216. <https://doi.org/10.1136/jcp-2023-208963>
 4. Ahluwalia, T. S., Troelsen, J. T., Balslev-Harder, M., Bork-Jensen, J., Thuesen, B. H., Cerqueira, C., Linneberg, A., Grarup, N., Pedersen, O., Hansen, T., & Dalgaard, L. T. (2016, September 14). Carriers of aVEGFAenhancer polymorphism selectively binding CHOP/DDIT3 are predisposed to increased circulating levels of thyroid-stimulating hormone. *Journal of Medical Genetics*, 54(3), 166–175. <https://doi.org/10.1136/jmedgenet-2016-104084>
 5. Merk, M., Zierow, S., Leng, L., Das, R., Du, X., Schulte, W., Fan, J., Lue, H., Chen, Y., Xiong, H., Chagnon, F., Bernhagen, J., Lolis, E., Mor, G., Lesur, O., & Bucala, R. (2011, August 4). The D -dopachrome tautomerase (DDT) gene product is a cytokine and functional homolog of macrophage migration inhibitory factor (MIF). *Proceedings of the National Academy of Sciences*, 108(34). <https://doi.org/10.1073/pnas.1102941108>

DATE:15/03/2024

WEBLEM 5
INTRODUCTION TO BLAST2GO Tool
(URL: <https://www.blast2go.com/>)

BioBam's offerings lies BLAST2GO, a comprehensive bioinformatics platform tailored for functional annotation of genomic data. Leveraging cutting-edge algorithms and user-friendly interfaces, BLAST2GO enables researchers to unravel the functional significance of their sequences with ease. From sequence similarity searches using BLAST to mapping gene ontology terms and conducting annotation enrichment analysis, BLAST2GO offers a holistic approach to functional annotation, empowering researchers to glean insights into the biological roles of their sequences.

BLAST2GO tool research has revolutionized our understanding of biological systems, providing vast amounts of raw sequence data waiting to be deciphered. However, extracting meaningful biological insights from these sequences requires sophisticated tools for functional annotation. Enter BLAST2GO, a comprehensive bioinformatics solution designed to streamline the process of functional annotation for genomics data. BLAST2GO combines powerful algorithms with user-friendly interfaces, making it accessible to both seasoned bioinformaticians and bench scientists. At its core lies the integration of three essential steps: sequence similarity search (using BLAST), mapping of gene ontology (GO) terms, and annotation enrichment analysis. The first step involves comparing input sequences against vast databases of known sequences using the widely adopted BLAST algorithm. This process identifies homologous sequences, providing crucial insights into the functional conservation and evolutionary relationships of the query sequences. Once homologous sequences are identified, BLAST2GO maps them to Gene Ontology terms, a structured vocabulary that describes gene products in terms of their associated biological processes, molecular functions, and cellular components. This step assigns functional annotations to the query sequences based on the annotations of their homologs, greatly enhancing our understanding of their potential roles in biological systems. Finally, BLAST2GO offers annotation enrichment analysis, allowing researchers to identify overrepresented GO terms within their dataset compared to a reference set. This analysis unveils biological processes, molecular functions, and cellular components that are significantly enriched within the input data, shedding light on the underlying biological mechanisms at play. Beyond these fundamental features, BLAST2GO boasts a range of additional functionalities, including sequence filtering, statistical analysis tools, and customizable workflows, empowering users to tailor the annotation process to their specific research needs.

Blast2GO is a bioinformatics platform for high-quality functional annotation and analysis of genomic datasets. Main Application Features are:

1. **Easy start-up and low maintenance:** Simply download the BLAST2GO from the portal and install the application and the updates are automatics.
2. **User-friendly:** Blast2GO is designed for experimentalists. An intuitive interface, the many graphical parameters and the detailed user's manual makes the use of the tool possible from the first try.
3. **High-throughput and interactive:** Blast2GO can annotate thousands of sequences, in multiple projects. Users can follow and modify the annotation process at any stage.
4. **Highly configurable:** Blast2GO is a functional annotation workstation. You can design your custom annotation style through the many configurable parameters. Statistical charts are available to guide users in the annotation process.
5. **Datamining:** Blast2GO does not only generate functional annotations. You can interrogate the biological meaning of your data with different graphical and statistical functions.
6. **Vocabularies:** Gene Ontology Terms, InterPro Domains, RFAM IDs and Enzyme Codes are supported by Blast2GO.

INSTALLATION OF THE TOOL:

Installing BLAST2GO involves several steps to ensure proper setup and functionality. Here's a detailed guide on how to install the BLAST2GO application:

Step 1: Obtain the BLAST2GO Software: Visit the official BioBam website or the BLAST2GO website to obtain the software. BLAST2GO is available for download on Windows, macOS, and Linux platforms.

Step 2: Download the Installation Package: Once on the website, navigate to the download section and select the appropriate version of BLAST2GO for your operating system (Windows, macOS, or Linux). Ensure that you download the latest stable version of the software.

Step 3: Check System Requirements: Before proceeding with the installation, ensure that your system meets the minimum requirements specified by BLAST2GO. This includes requirements for operating system version, processor, RAM, and disk space.

Step 4: Installation Process: Locate the downloaded installation package (usually a .exe file) and double-click on it to initiate the installation process. Follow the on-screen instructions provided by the installation wizard. You may be prompted to specify the installation directory and agree to the terms of the license agreement. Once the installation is complete, BLAST2GO should be accessible from the Start menu or desktop shortcut.

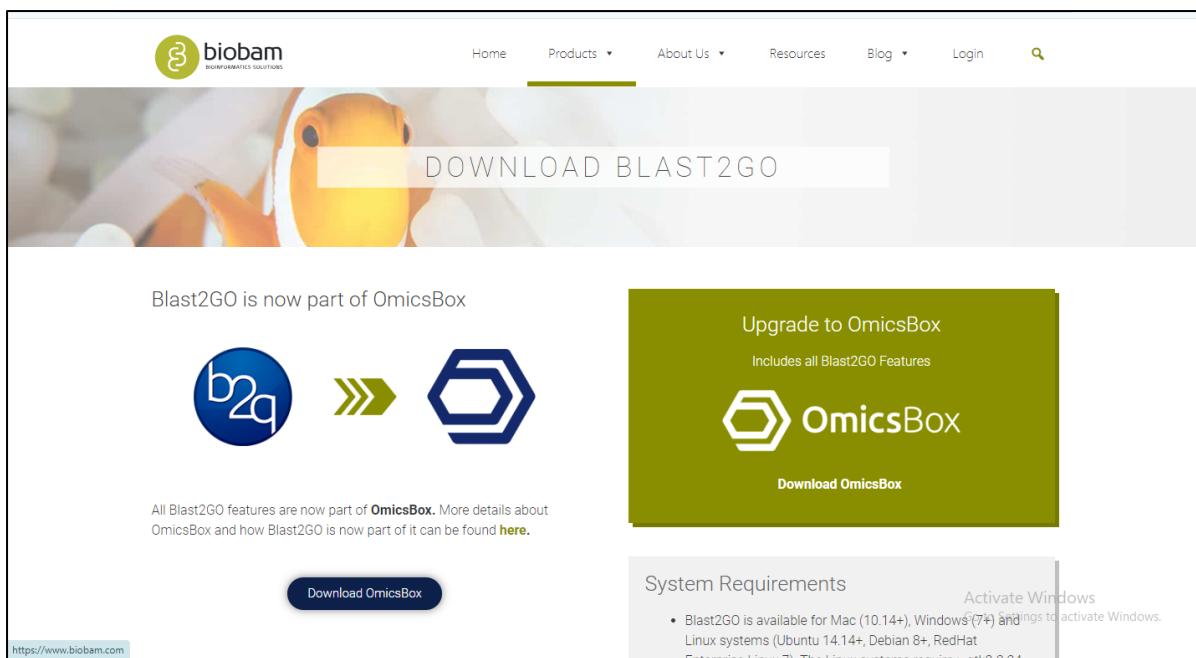


Fig 1: Homepage of BIOBAM

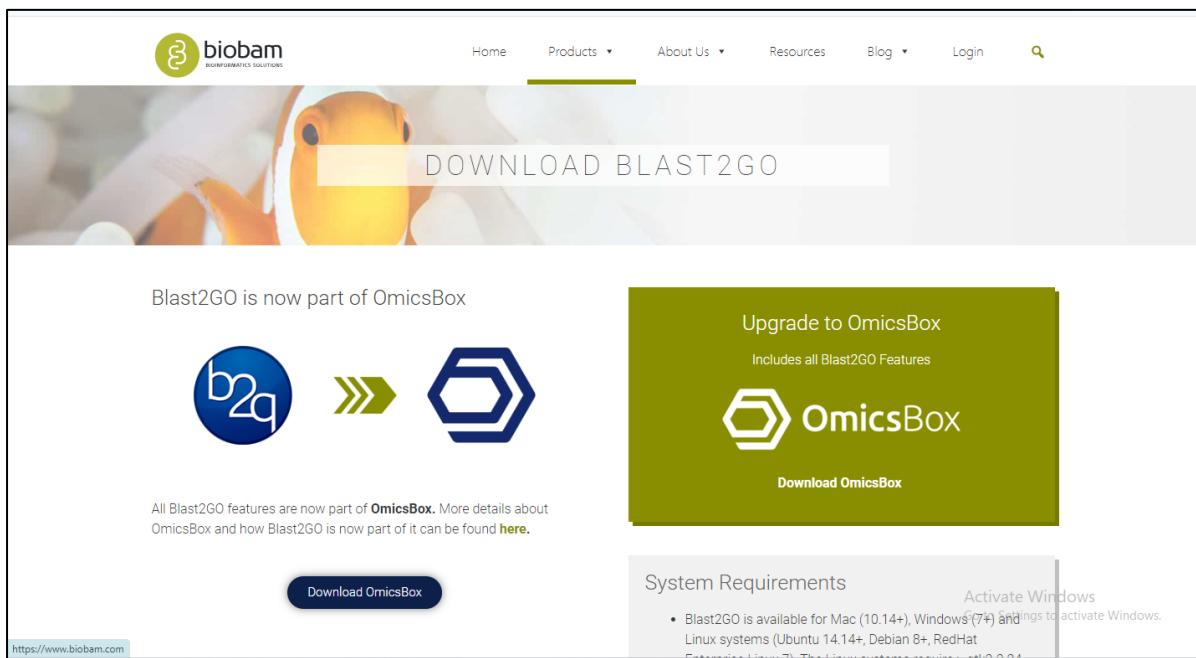


Fig 2: Download the latest version BLAST2GO tool for windows

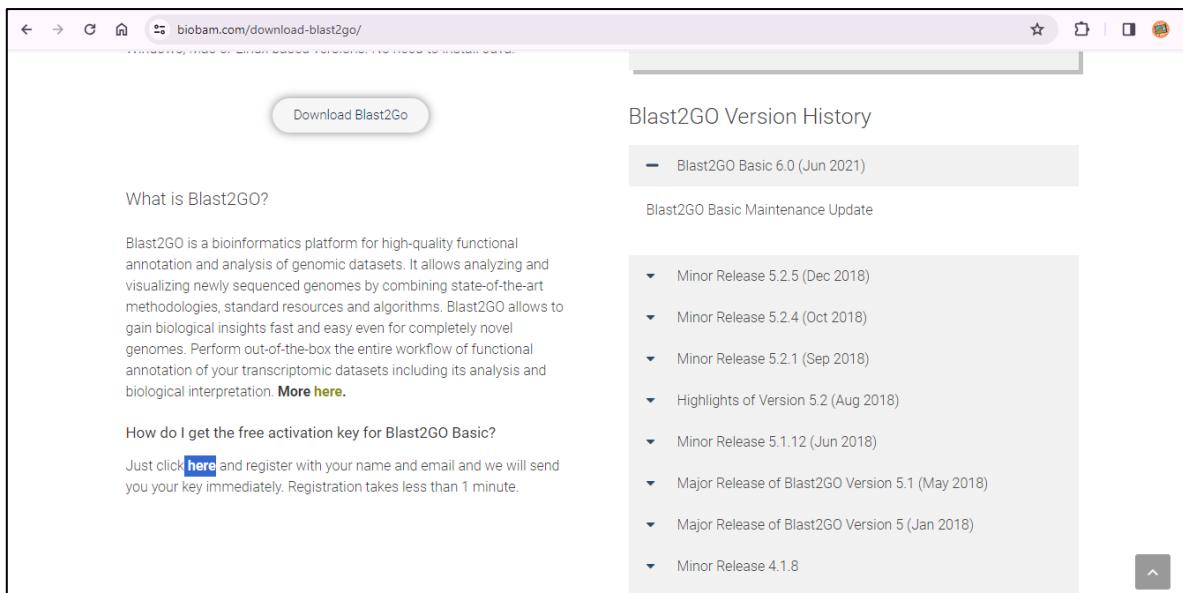


Fig 3: Click on the highlighted part for subscription key

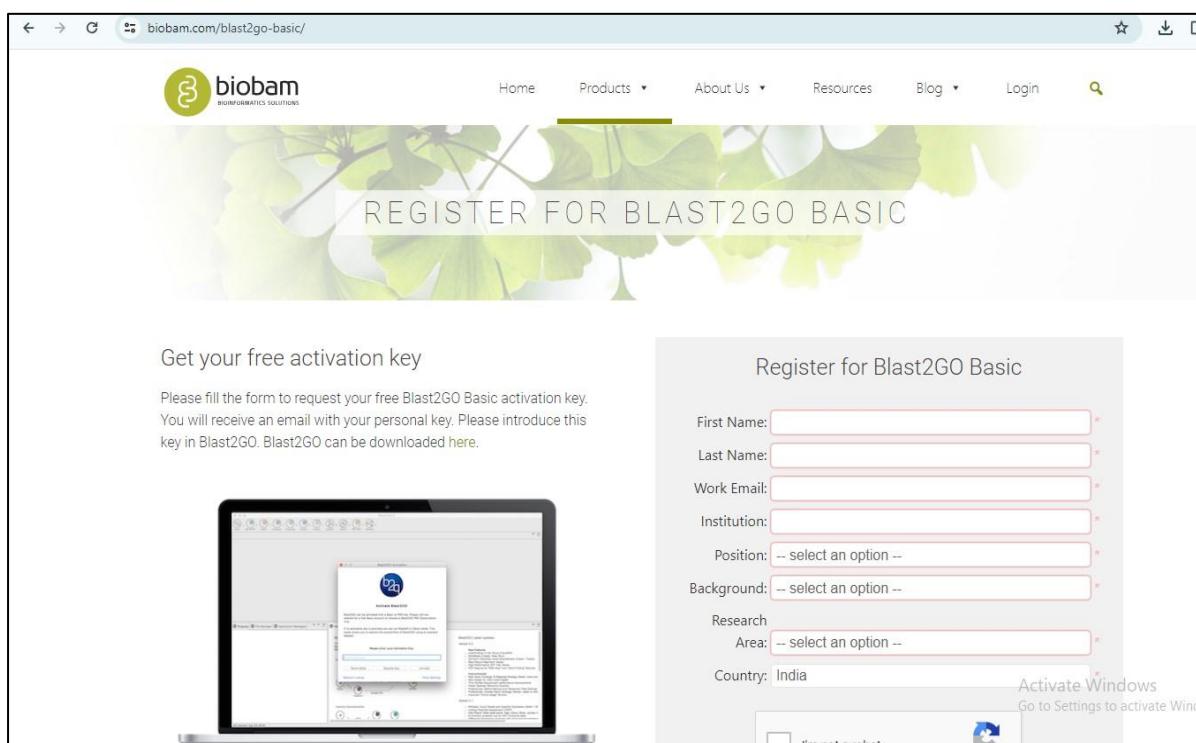


Fig 4: Registration page for subscription key

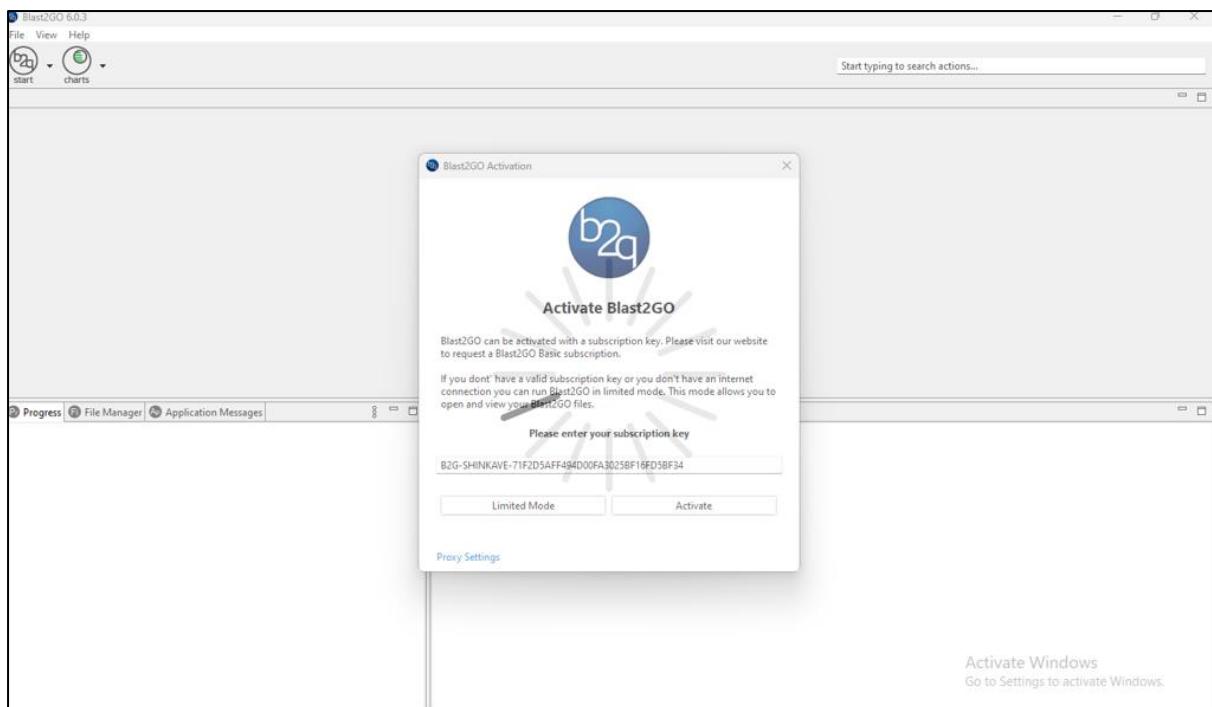


Fig 5: Enter the subscription received through the mail-ID

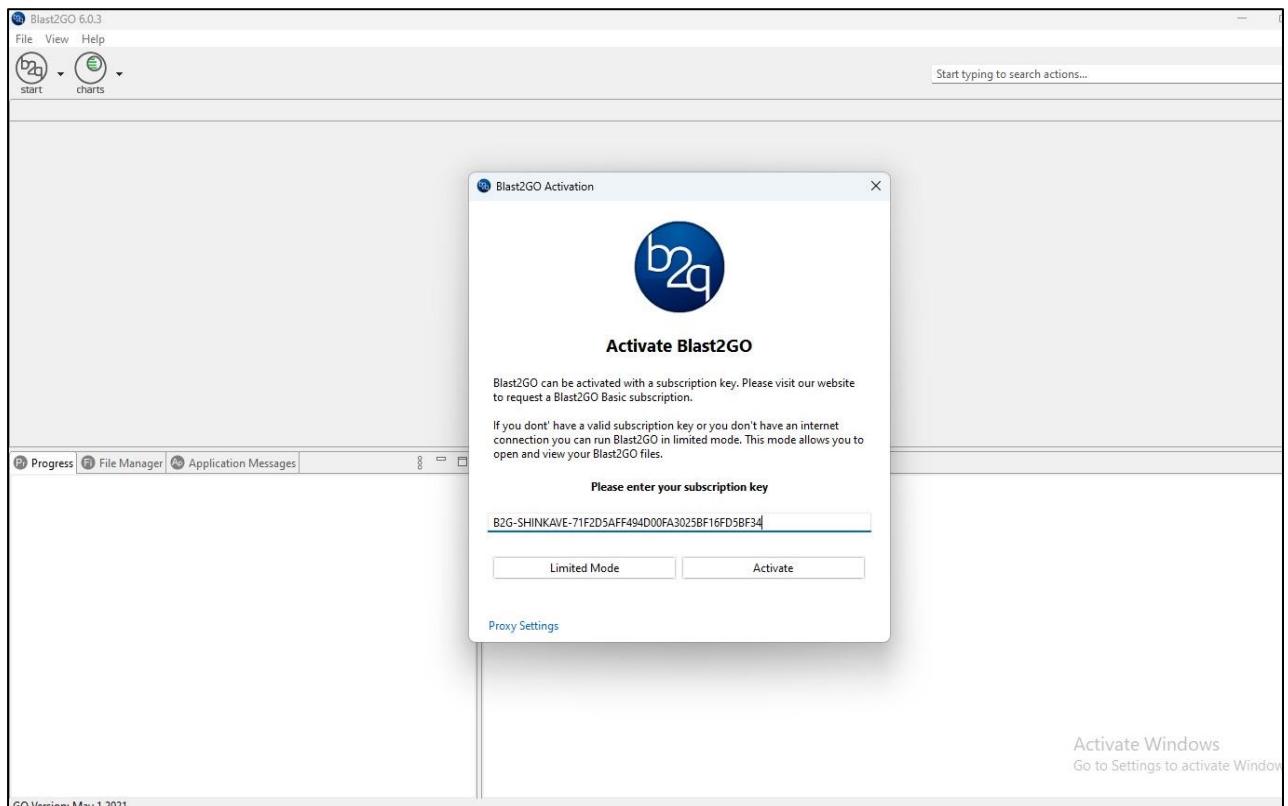


Fig 6: Activation after entering the subscription key

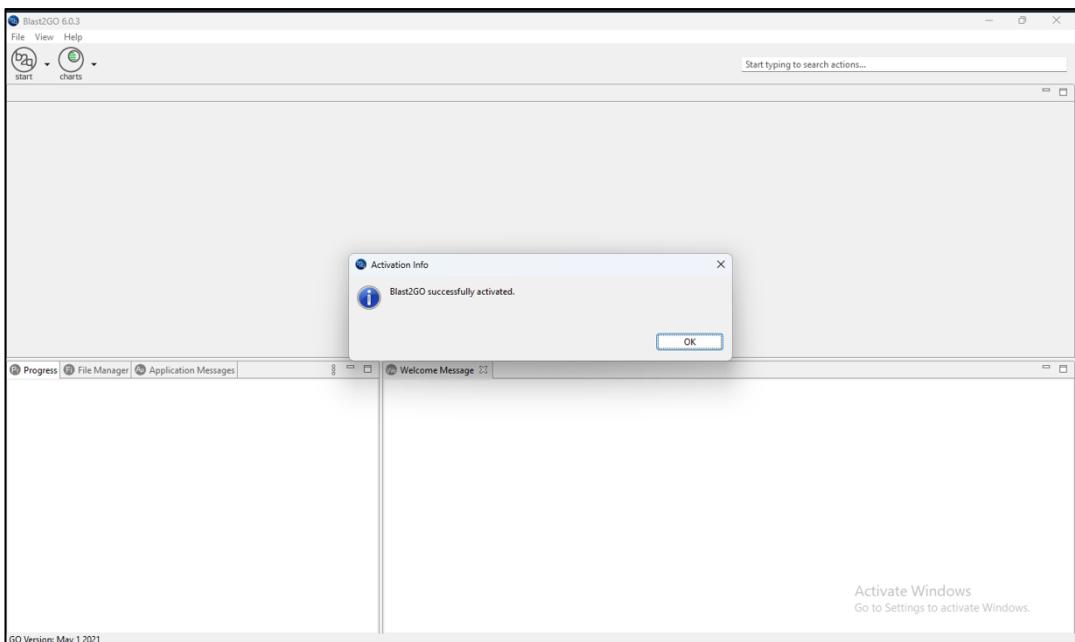


Fig 7: Your BLAST2GO is successfully activated

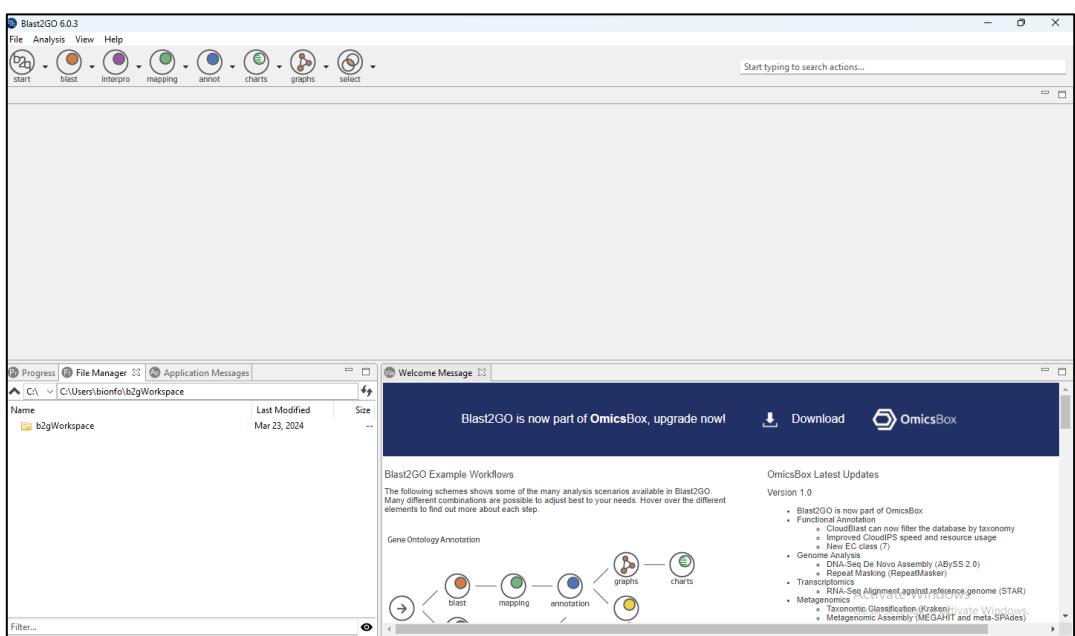


Fig 8: BLAST2GO

REFERENCES:

1. BioBam. (2024, February 28). BioBam - Bioinformatics Made Easy - Bioinformatics Made Easy. <https://www.biobam.com>.
2. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), 3674–3676. <https://doi.org/10.1093/bioinformatics/bti610>

DATE: 15/03/2024

WEBLEM 5(A)
Blast2GO
(URL: <https://www.blast2go.com/>)

AIM:

To annotate and analyze genomic or transcriptomic sequences for query ‘LEPR’ to understand their functional significance by using Blast2GO tool.

INTRODUCTION:

Blast2GO is a comprehensive bioinformatics tool for the functional annotation and analysis of genome-scale sequence datasets. The software was originally developed to provide a user-friendly interface for Gene Ontology annotation. Over the last years, many improvements have considerably increased the functionality of Blast2GO and many different functional genomics tools are now available. Additionally, the application offers a wide array of graphical and analytical tools for data manipulation and mining. The main concept behind the developments is the easiness for biological researchers: minimal set up requirements, automatic updates, simplicity in the usage and visual-oriented information display. Advanced functionality is instantly available requiring a minimal computational background. Basically, Blast2GO uses local or remote BLAST searches to find similar sequences to one or several input sequences. The program extracts the GO terms associated to each of the obtained hits and returns an evaluated GO annotation for the query sequence(s). Enzyme codes are obtained by mapping from equivalent GOs while InterPro motifs are directly queried at the InterProScan web service. GO annotation can be visualized reconstructing the structure of the Gene Ontology relationships.

A typical basic use case of Blast2GO consists of 5 steps: BLASTing, mapping, annotation, statistical analysis and visualization. These steps will be described in this document including installation instructions, further explanations and information on additional functions.

The program extracts GO terms to each obtained hit by mapping to existent annotation associations. An annotation rule finally assigns GO terms to the query sequence. Annotation and functional analysis can be visualized in a graph form reconstructing the GO relationships and color-highlighting the most relevant areas. B2G was conceived to be an attractive tool for research environments where genetic and/or computational resources are limited and where much work is still done in an explorative fashion. B2G is a user-friendly, easy to distribute and low maintenance tool. It allows monitoring and interaction at different steps of the analysis, and emphasizes visualization as an important component of knowledge acquisition. B2G is a Java application made available by Java Web Start. It is platform independent and has no further requirements than an Internet connection.

LEPR:

The LEPR gene, also known as the leptin receptor gene, is located on chromosome 1 and provides instructions for making a protein called the leptin receptor. This protein is found on the surface of cells in many organs and tissues of the body, including the hypothalamus, a part of the brain that controls hunger and energy balance. The leptin receptor plays a crucial role in the regulation of body weight by binding to the hormone leptin, which is produced by adipose tissue and signals available energy reserves to the brain.

In mammals, six isoforms of the leptin receptor (LepRa-f) have been identified, all of which have a common extracellular ligand binding domain. The long form of the leptin receptor (LepRb) has an approximately 300 amino acid cytoplasmic tail that mediates intracellular signaling, while the LepRa, -c, -d, and -f have short (~30-40) amino acid cytoplasmic extensions. The LepRe lacks transmembrane and cytoplasmic domains and may function as a secreted leptin binding protein. The LepRb is the only isoform that contains intracellular tyrosine residues necessary for signaling; the physiological functions, if any, for other LepR isoforms are unknown.

METHODOLOGY:

1. Open homepage of UNIPROT.
2. Extract 5 FASTA sequences of the same protein.
3. Save it in a notepad.
4. Open Blast2Go and load the file in sequences in FASTA file format.
5. After loading click on Blast and use NCBI Blast.
6. Set Blast program as BlastP and Blast DB as Swissprot.
7. Let the advanced configurations be as it is.
8. Run the file and save the results in XML file.
9. For interpro, let the configurations for the first page be default.
10. Only select Superfamily in other sequence features and let the others remain unchecked.
11. Run the file and save the results in XML file format.
12. After results are obtained and loaded onto the homepage, run the Blast2go statistics.
13. Select all four statistic options and hit run.\
14. For getting InterPro results, select InterProScan results and hit run.
15. Obtain results in graph formats.

OBSERVATIONS:

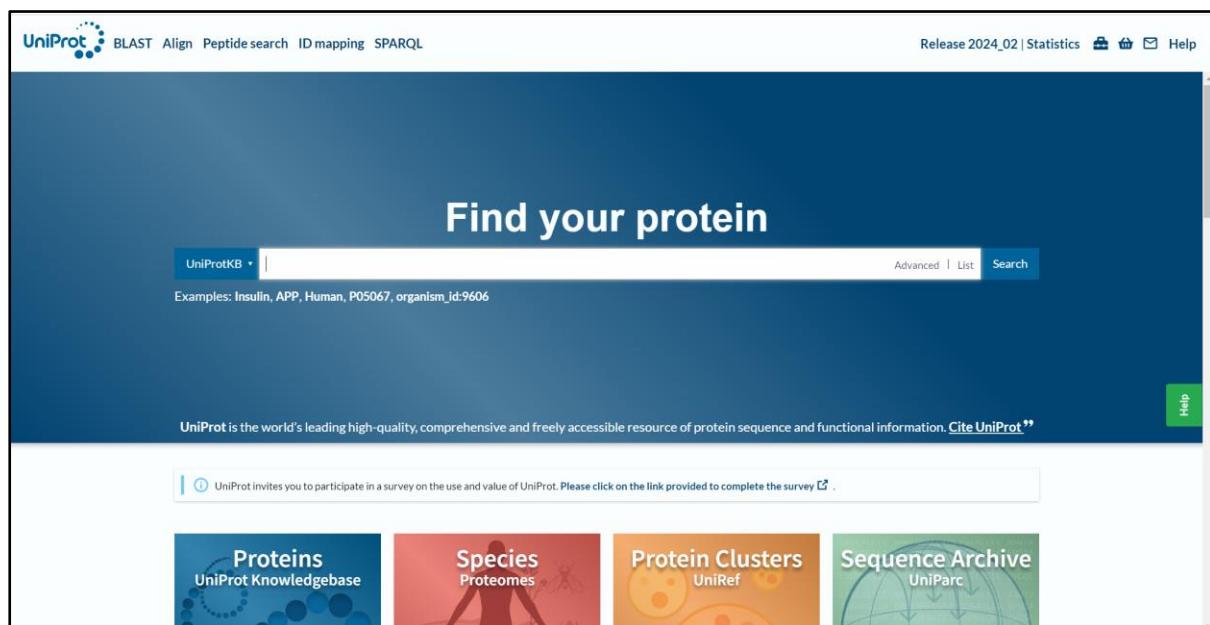


Fig 1: Homepage of UniProt database

The screenshot shows the UniProtKB search results for the query "LEPR". The header is identical to Fig 1. The main content area displays a table of 3,085 results. The columns are: Entry, Entry Name, Protein Names, Gene Names, Organism, and Length. The table lists several entries, including Q62959 (LEPR_RAT), O02671 (LEPR_PIG), Q9MYL0 (LEPR_MACMU), P48357 (LEPR_HUMAN), P48356 (LEPR_MOUSE), O15243 (OBRG_HUMAN), O89013 (OBRG_MOUSE), Q62120 (JAK2_MOUSE), Q9JLS8 (OBRG_RAT), G1NH43 (G1NH43_MELGA), and O9I8V6 (O9I8V6_CHICK). The table has a "Feedback" button on the right and a "Help" button at the bottom right.

Fig 2: 5 selected sequences of query 'LEPR'

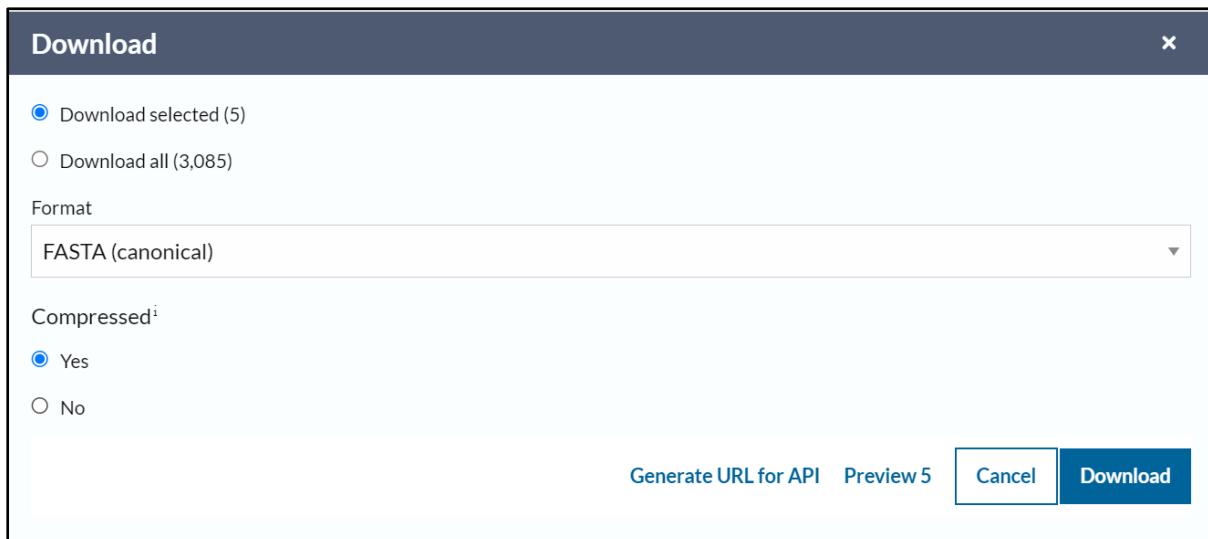


Fig 3: Downloading FASTA sequences of query 'LEPR'

```
>sp|P48357|LEPR_HUMAN Leptin receptor OS=Homo sapiens OX=9606 GN=LEPR PE=1 SV=2
MICQKFCVVLHWEFIYVITAFNLSPYTPWRFKLSCMPNNSTYDFLLPAGLSKNTNS
NGHYETAVEPKFNSSGTHFNSNLKTTFHCCFRSEQRNCSCADNIEGKTFVSTVNSLVF
QQIDANWNIOCWLKGDLKLFIYCYESLFKNLFRNYNKVHLILYVLPEVLEDSPLPVKGS
FQMWHCNCVSHCCECLVPVPTAKLNLDLMLKITSQGVIFQSPPLMSVQPINMVKPDP
LGLHMEITDDGNLKTSWSSPLVPPFLQQYQVKYSENSTTVIREADKIVSATSLVDSILP
GSSYEVQVRGKRLDGPWIWSDWSTPRVFTTQDVYFPPKILTSVGSNSFHCIFYKKENKI
VPSKEIVWWMLAEKIPQSQYDVVSDHVSKVFFNLNETKPRGKFTYDAVYCCNEHECHH
RQAEELYVIDVNINISCTEDGYLTKMTCRWSTSTIQSLAESTQLRHYHRSSLYCSIDPSIH
PISEPKDCYLSQDFYEC1FQPFIFLLSGTYMWIRINHSGSLDSPPTCVLPSVVKPLPP
SSVKAETINIGLLKISWEKPVFPENNLFQFQIRYGLSGKEVQWKMYEVYDAKSKSVLV
PDLCAVYAVQVRCKRLDGLGYWSNWSPAYTVMDIKVPMRGPEFWRIINGDTMKKEKNV
TLLWKPLMKNDLCSVQRYVNHHTCNGTWSEDVGHNHTKFTFLWTEQAHTVTVLAINSI
GASVANFNLTFSWPMSCVKNVIQSL SAYPLNNSCVIVSWSLSPSDYKLMYFIIEWKLNED
GEIKWLRISSVVKKYIHDHFPIIEKYQFSLYPIFMEGVGKPKIINSFTQDDIEKHQSDA
GLYVIVPVISSSILLGTLISHQRMKKLFWEDVPNPNCWSAQGLNFQKPETFEHLFI
KHTASVTCGPLLEPETIDESIVDTSWKNKDEMMPPTVVSLLSTTDLEGKGSVCISDQFN
SVNFSEAEGETVTEDESQRQFVVKYATLINSNSKPSYEEQGLINSSVTCKFSSKNSPL
KDSFSNSSWIEAQAFFILSDQHPNIISPHLTFSEGLDELLKLEGNFPEEENDKKSIYV
GVTSIKKRESGVLLDKSRVSCPAPCLFTDIRVLQDSCSHFVENNINLGTSKKTFAS
YMPQFQTCSTQTHKIMENKMCDLTV
>sp|Q62959|LEPR_RAT Leptin receptor OS=Rattus norvegicus OX=10116 GN=Lepr PE=1 SV=1
MTQKFYVVLHWEFLYVITALNLAYPTSPWRFKLCAPSTTDDSLPAGVPNTSSL
KGASEALVEAKFNSTGIYSELSTKTIHCCFGNEQQNCNSALTGTNTEGKTLASVVKPLVF
RQLGVNWIDECWMKGDLTLFICHMEPLLKNPKNYDSKVHLLYDLPEVIDDLPLPLKDS
FQTWQCNCSVRECECHVPRAKVNAYALLMYLEITSAGVSFQSPPLSLOQPLMLVVKPDPL
GLRMEVTDDGNLKISWDSQTQKAPFPLQYQVKYLENSTIVREAEEIVSDTSLLVDSVLPGS
SYEVQRSKRLDGSQWSDWSLPQLFTTQDVYFPPKILTSVGSNSFCCIFYKNEQNQTS
SKQIVWWMLAEKIPETQYNTVDHISKVTSNLKATRPRGKFTYDAVYCCNEQACHRY
AELYVIDVNINISCTEDGYLTKMTCRWSPSTIQSLVGVSTVQLRHYRSSLYCPDNPNSIRPT
SELKNCVLTQDGFYECVFLSGTYMWIRINHSGSLDSPPTCVLPSVVKPLPPSN
VKAETINTGLLKVSGWEKPVFPENNLFQIRYGLNGKEIOWKTHEVFDAKSKASLPVSD
LCAVYVVQVRCRLDGLGYWSNWSSPAYTLVMDVKVPMRGPEFWRIMDGITKKERNVTL
LWKPLMKNDLCSVRRYVVKRTAHNGTWSQDVGQNTNLTFLWAESAHTVTVLAINSIGA
SLVNFNLTFSWPMSCVKNVIAQSL SAYPLSSSCVILSWTSPNDYSLLYLVIEWKLNNDGG
MKWLRIPSNVNKYIHDNPIIEKYQFSLYPVFMEGVGPKIIINGFTKDDIAKQQNDAGL
YIVVPIIISCVLLGTLISHQRMKKLFWDVNPNCWSAQGLNFQKPETFEHLFTKH
AESVIFGPLLEPEVSEEISVDTAWKNKDEMVPAAVMSSLTTPDTRGSICISDQNS
ANFGAQSTQGTCEDECQSQPSVKYATLVSIVKTVETDEEQGAIHSSVSQCIAKHSPLR
QSFSNSWEIEAQAFFILSDHPPNISPQLSGLDELLELEGNFPEENHGEKSVYYLGV
SSGNKRENDMLLTDEAGVLCPPPAHCLFSDIRILQESCASHFVENNINLGTSGNFVPYMP
QFQSCSTHSHKIIENKMCDLTV
>sp|P48356|LEPR_MOUSE Leptin receptor OS=Mus musculus OX=10090 GN=Lepr PE=1 SV=1
MMCQKFYVVLHWEFLYVIALNAYPISPMKFKLFCGPNTTDDSFSLPAGPNASAL
KGASEAIVEAKFNSSGIYVPELSKTVHCCFGNEQQNCNSALTDTNEGKTLASVVKASVF
RQLGVNWIDECWMKGDLTLFICHMEPLPKNPKNYDSKVHLLYDLPEVIDDPLPPLKDS
```

Fig 4: Opening it in Notepad

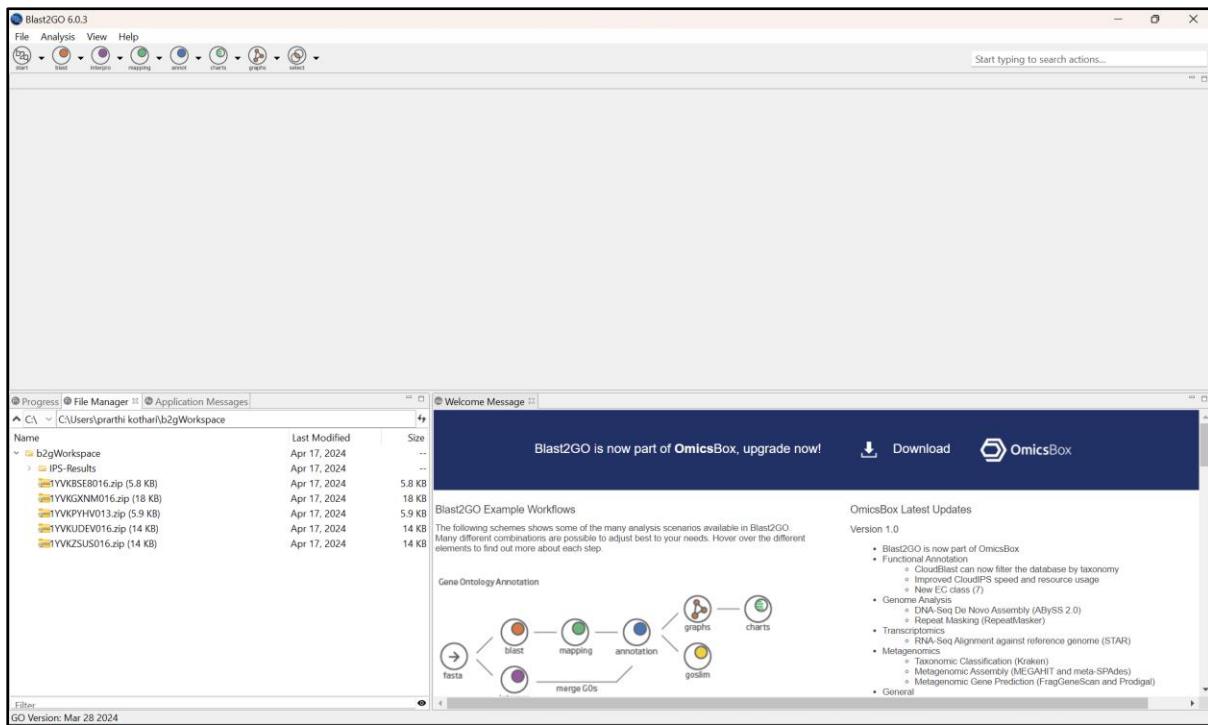


Fig 5: Opening page of Blast2GO

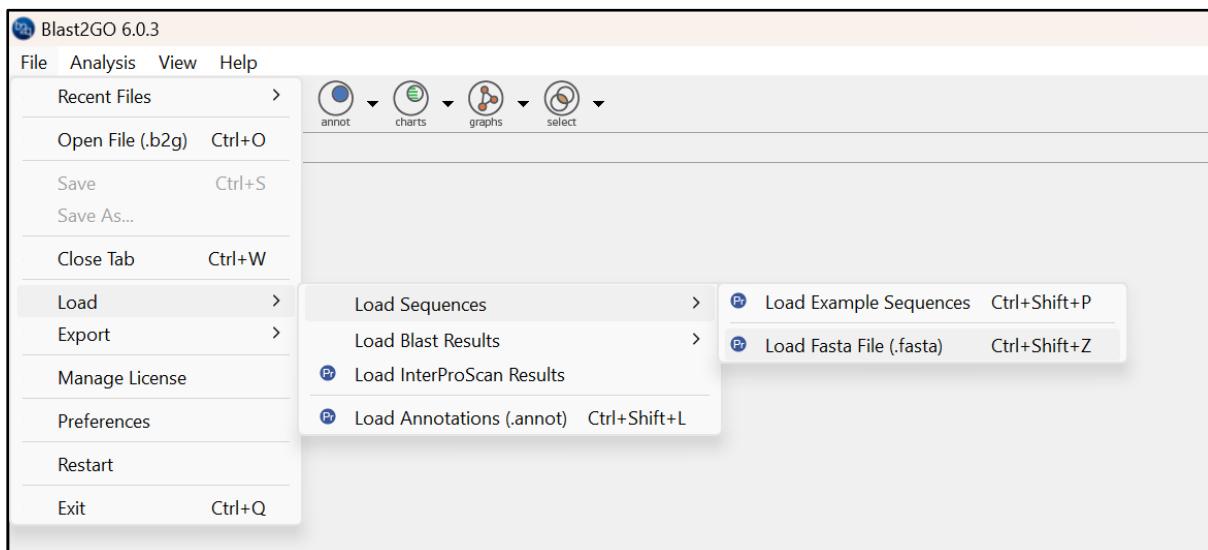


Fig 6: Pathway to load file

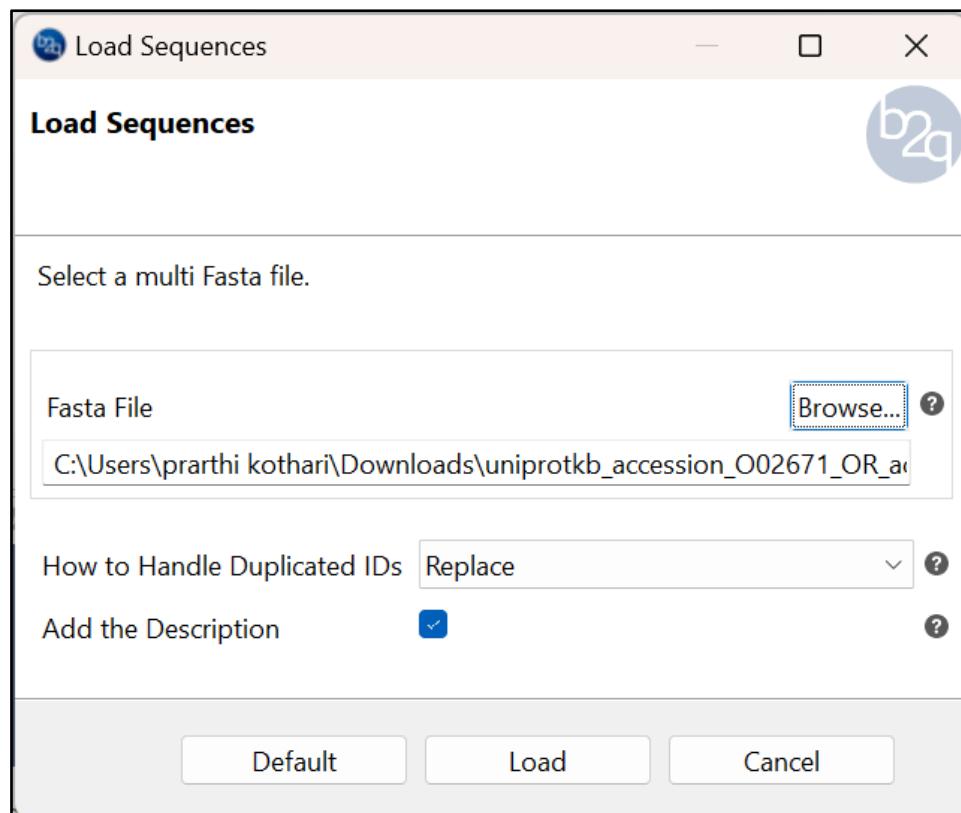


Fig 7: Loaded file

The screenshot shows the Blast2GO 6.0.3 interface. The top navigation bar includes 'File', 'Analysis', 'View', and 'Help'. Below the bar are eight circular icons with labels: 'start', 'blast', 'interpro', 'mapping', 'annot', 'charts', 'graphs', and 'select'. The main area displays a table titled '*Table: uniprotkb_accession_O02671_OR_accession_2024_04_17...'. The table has columns: Nr, Tags, SeqName, Description, Length, and #Hits. The data is as follows:

Nr	Tags	SeqName	Description	Length	#Hits
1		sp O02671 LEPR_PIG	Leptin receptor OS=Sus...	1165	
2		sp P48356 LEPR_MOUSE	Leptin receptor OS=Mus...	1162	
3		sp P48357 LEPR_HUMAN	Leptin receptor OS=Ho...	1165	
4		sp Q62959 LEPR_RAT	Leptin receptor OS=Ratt...	1162	
5		sp Q9MYL0 LEPR_MAC...	Leptin receptor OS=Mac...	1163	

Fig 8: FASTA files displayed in Blast2GO

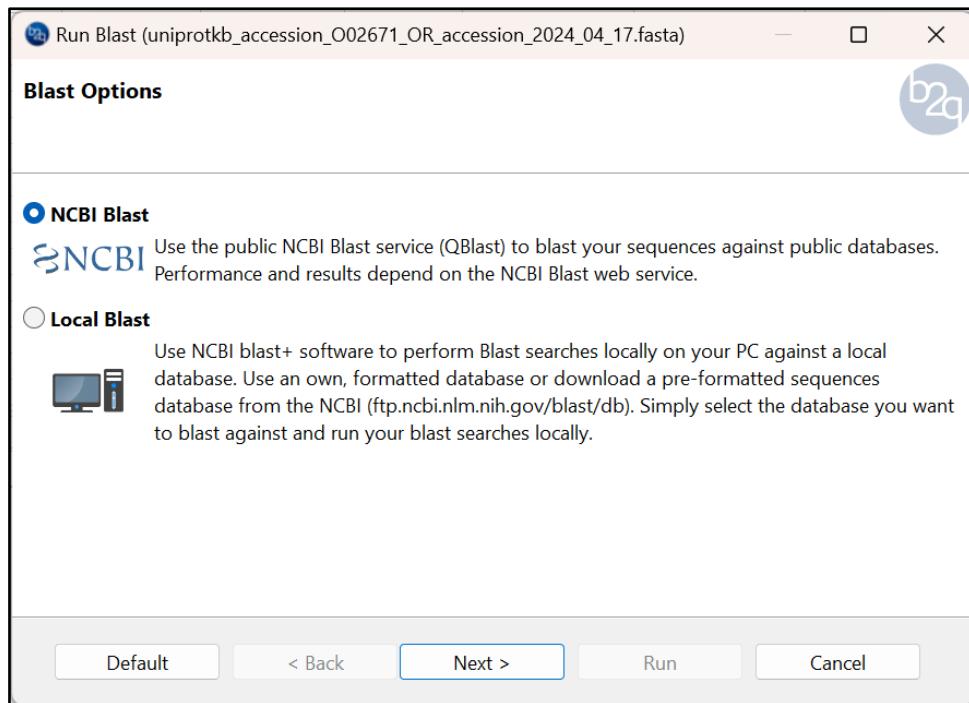


Fig 9: Running Blast

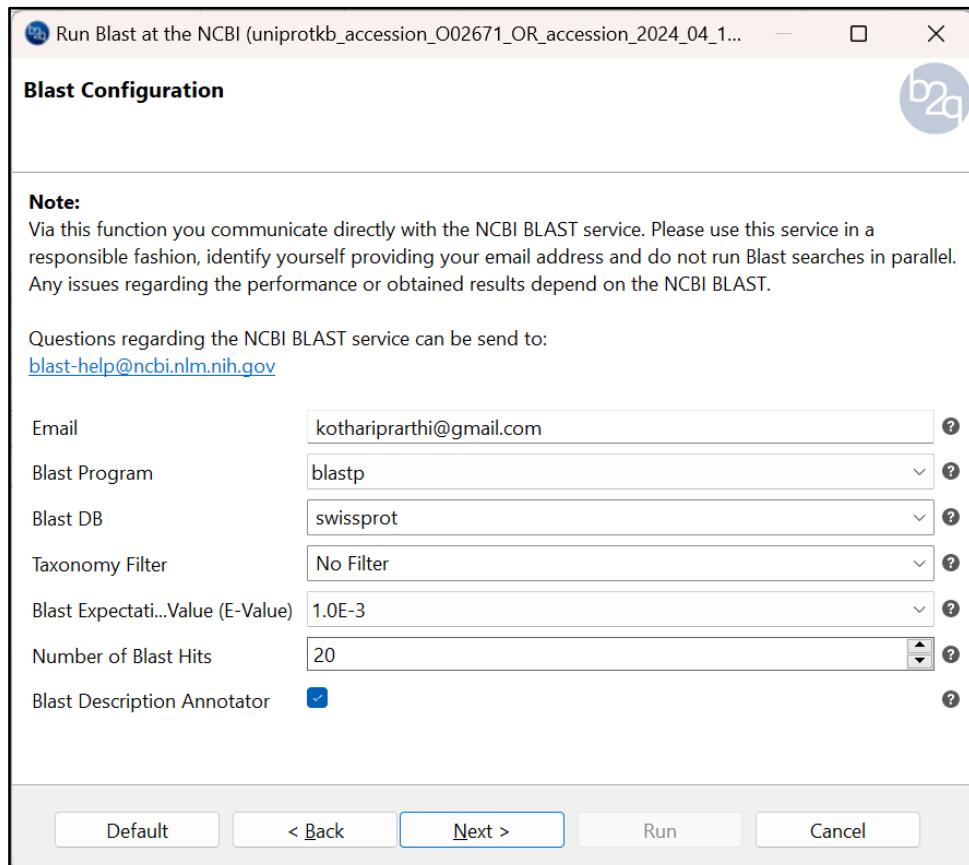


Fig 10: Blast input parameters

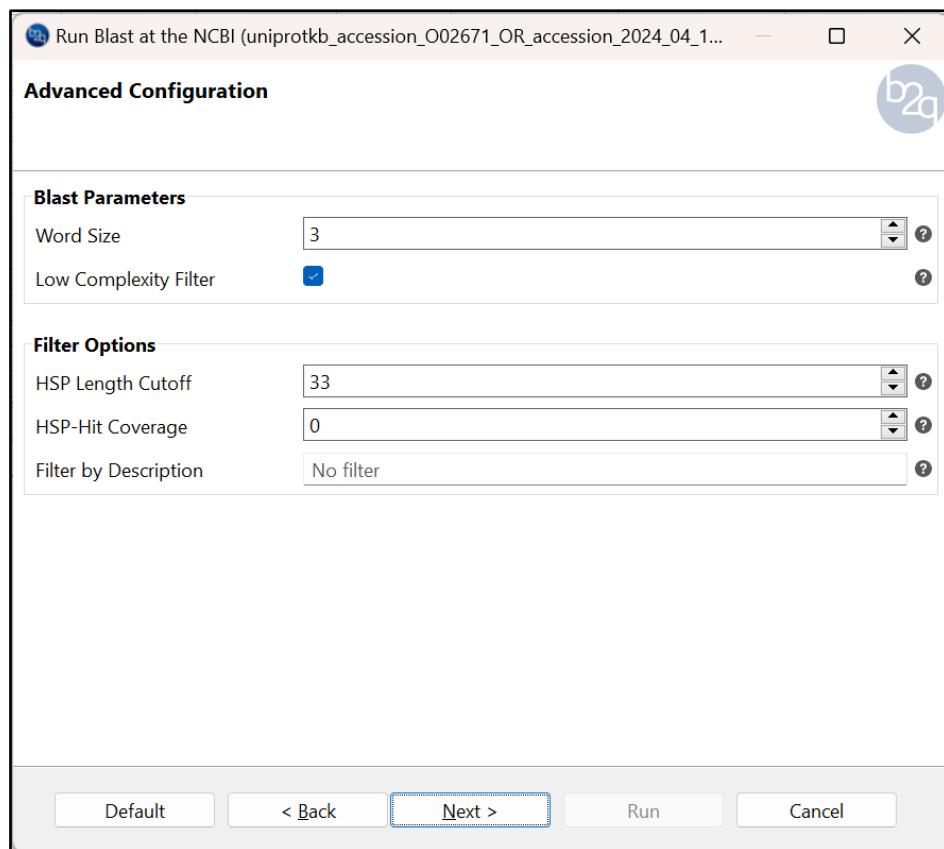


Fig 11: Advanced configurations for the tool

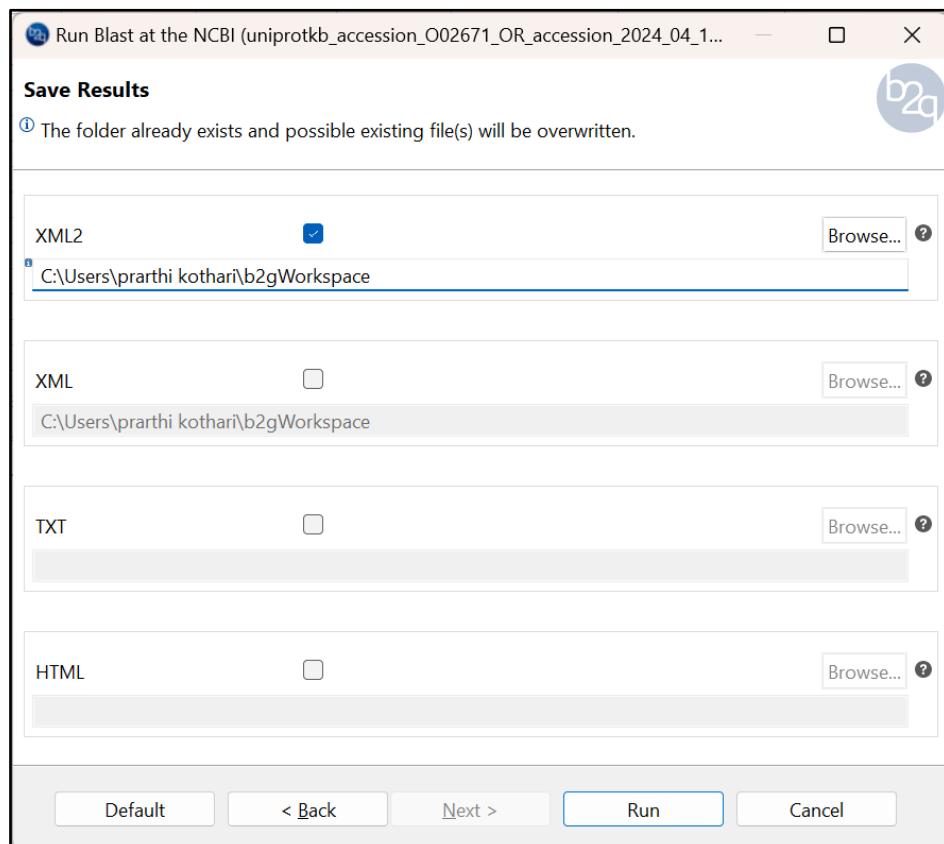


Fig 12: Running the file

Blast2GO 6.0.3

File Analysis View Help

start blast interpro mapping annot charts graphs select

*Table: uniprotkb_accession_O02671_OR_accession_2024_04_17...

Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim mean	GO
<input checked="" type="checkbox"/> 1	BLASTED	sp Q02671 LEPR_PIG	RecName: Full=Leptin re...	1165	20	0E0	53.67%	
<input checked="" type="checkbox"/> 2	BLASTED	sp P48356 LEPR_MOUSE	RecName: Full=Leptin re...	1162	20	0E0	54.34%	
<input checked="" type="checkbox"/> 3	BLASTED	sp P48357 LEPR_HUMAN	RecName: Full=Leptin re...	1165	20	0E0	54.26%	
<input checked="" type="checkbox"/> 4	BLASTED	sp Q62959 LEPR_RAT	RecName: Full=Leptin re...	1162	20	0E0	53.48%	
<input checked="" type="checkbox"/> 5	BLASTED	sp Q9MYL0 LEPR_MAC...	RecName: Full=Leptin re...	1163	20	0E0	53.78%	

Qblast (uniprotkb_accession_O02671_OR_accession_2024_04_17.fasta) Done X

i Qblast finished!
Blasted Sequences: 5
Sequences without results: 0

OK

Fig 13: Results obtained for Blast

Run InterProScan (uniprotkb_accession_O02671_OR_accession_2024_04_17.fa...)

InterProScan Configuration 1

Important:
To run InterProScan, the sequence information (FASTA) is needed.
Nucleotides will be translated to AminoAcids automatically.

Via this function you communicate directly with the public InterProScan web service offered by the EBI.
Please use this service in a responsible fashion, identify yourself providing your email address and do not run multiple searches in parallel.

Email

Families, Domains, Sites and Repeats

CDD	<input checked="" type="checkbox"/>	?
HAMAP	<input checked="" type="checkbox"/>	?
HMMPanther	<input checked="" type="checkbox"/>	?
HMMPfam	<input checked="" type="checkbox"/>	?
HMMPiR	<input checked="" type="checkbox"/>	?
FPrintScan	<input checked="" type="checkbox"/>	?

Default < Back **Next >** **Run** **Cancel**

Fig 14: Input Parameters for Interpro

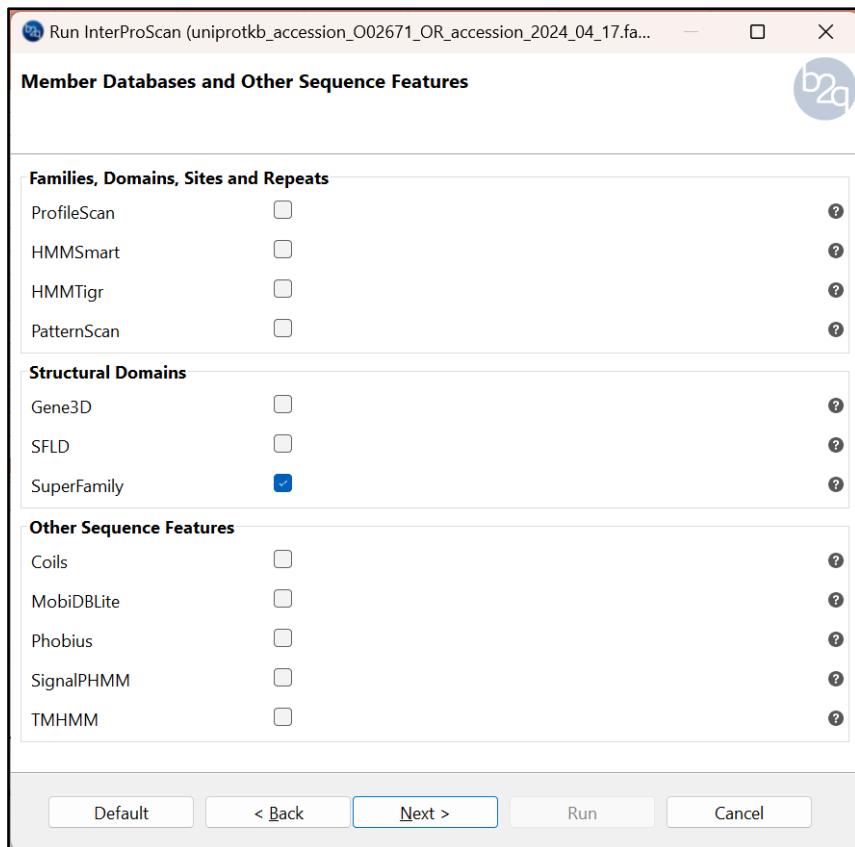


Fig 15: Ticking on Superfamily under structural domains

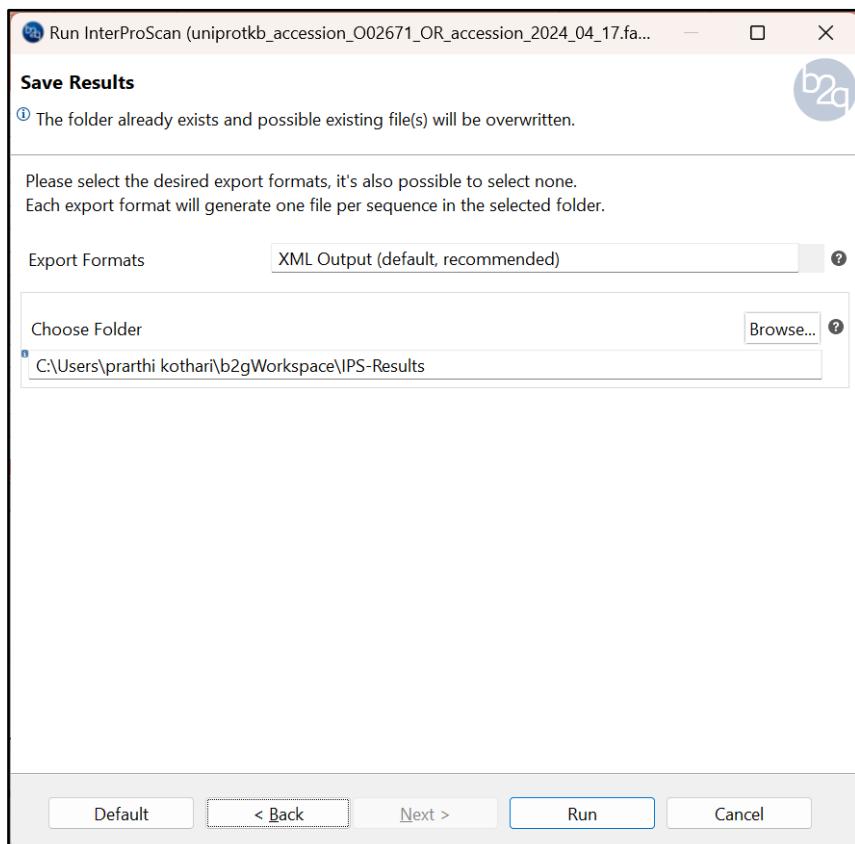


Fig 16: Running Interpro scan

InterPro IDs	InterPro GO IDs	InterPro GO Names
IPR041182 (PFAM); IPR010457 (PFAM); IPR050379 (PANTHER); IPR003961 (CDD); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY)	P:GO:0019221; F:GO:0004896; F:GO:0005515; F:GO:0019955; C:GO:0009897; C:GO:0043235	P:cytokine-mediated signaling pathway; F:cytokine receptor activity; F:protein binding; F:cytokine binding; C:external side of plasma membrane; C:receptor complex
IPR041182 (PFAM); IPR010457 (PFAM); IPR050379 (PANTHER); IPR003961 (CDD); IPR003961 (CDD); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY)	P:GO:0019221; F:GO:0004896; F:GO:0005515; F:GO:0019955; C:GO:0009897; C:GO:0043235	P:cytokine-mediated signaling pathway; F:cytokine receptor activity; F:protein binding; F:cytokine binding; C:external side of plasma membrane; C:receptor complex
IPR010457 (PFAM); IPR041182 (PFAM); IPR050379 (PANTHER); IPR003961 (CDD); IPR003961 (CDD); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY)	P:GO:0019221; F:GO:0004896; F:GO:0005515; F:GO:0019955; C:GO:0009897; C:GO:0043235	P:cytokine-mediated signaling pathway; F:cytokine receptor activity; F:protein binding; F:cytokine binding; C:external side of plasma membrane; C:receptor complex
IPR041182 (PFAM); IPR010457 (PFAM); IPR050379 (PANTHER); IPR003961 (CDD); IPR003961 (CDD); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY)	P:GO:0019221; F:GO:0004896; F:GO:0005515; F:GO:0019955; C:GO:0009897; C:GO:0043235	P:cytokine-mediated signaling pathway; F:cytokine receptor activity; F:protein binding; F:cytokine binding; C:external side of plasma membrane; C:receptor complex
IPR041182 (PFAM); IPR010457 (PFAM); IPR050379 (PANTHER); IPR003961 (CDD); IPR003961 (CDD); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY); IPR036116 (SUPERFAMILY)	P:GO:0019221; F:GO:0004896; F:GO:0005515; F:GO:0019955; C:GO:0009897; C:GO:0043235	P:cytokine-mediated signaling pathway; F:cytokine receptor activity; F:protein binding; F:cytokine binding; C:external side of plasma membrane; C:receptor complex

Fig 17: Results obtained after InterPro Scan

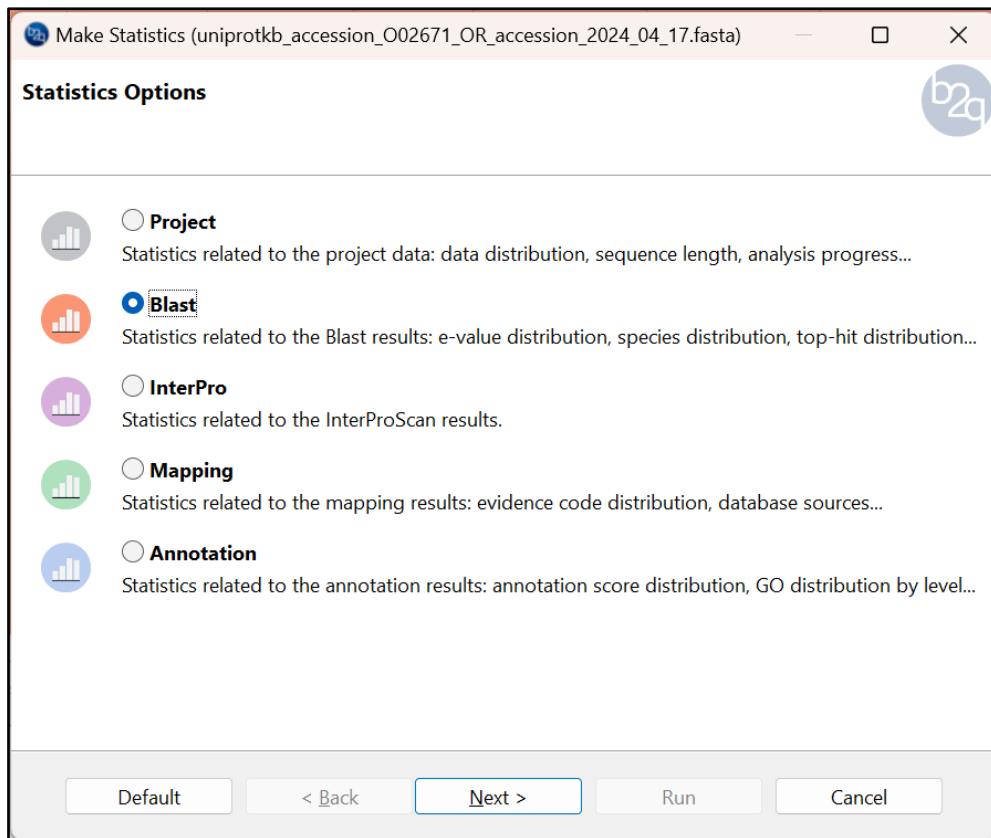


Fig 18: Statistics option after running Blast

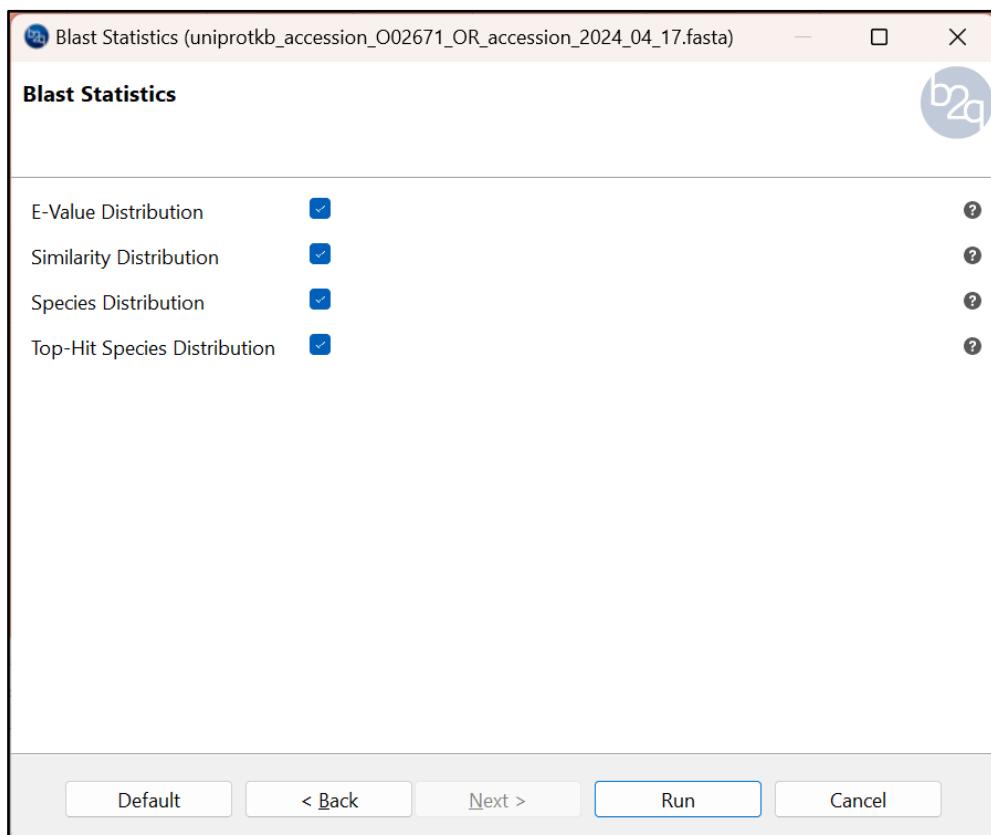


Fig 19: Adding input parameters and hitting run

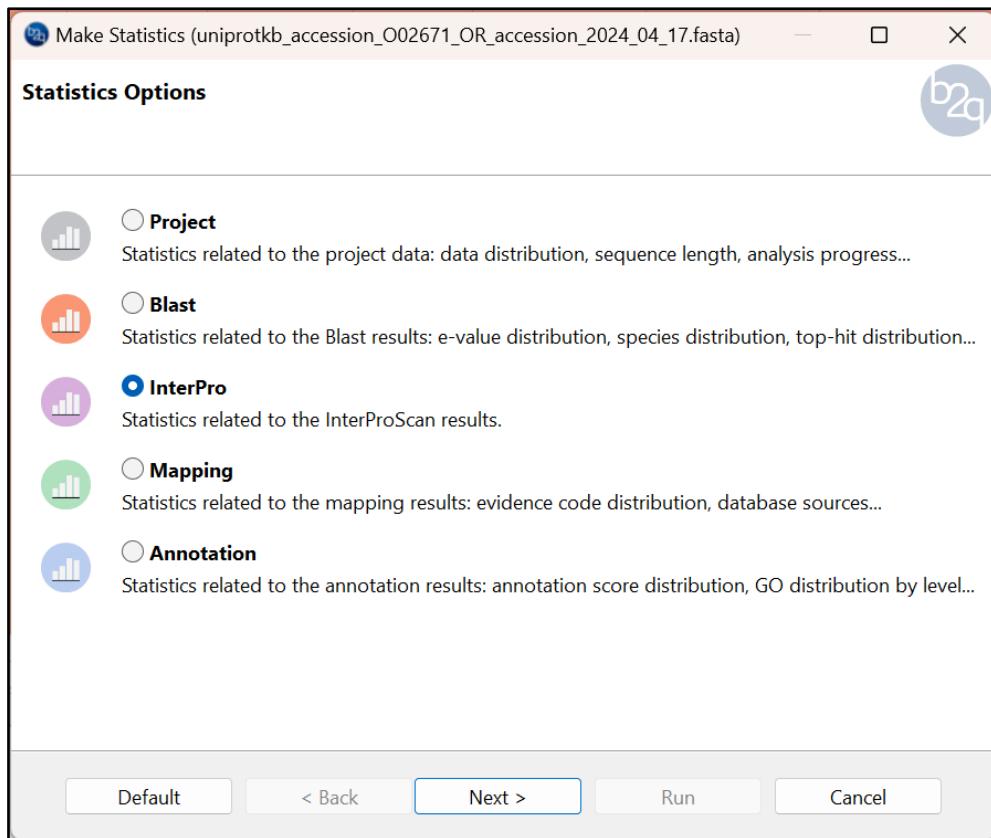


Fig 20: InterPro statistics option

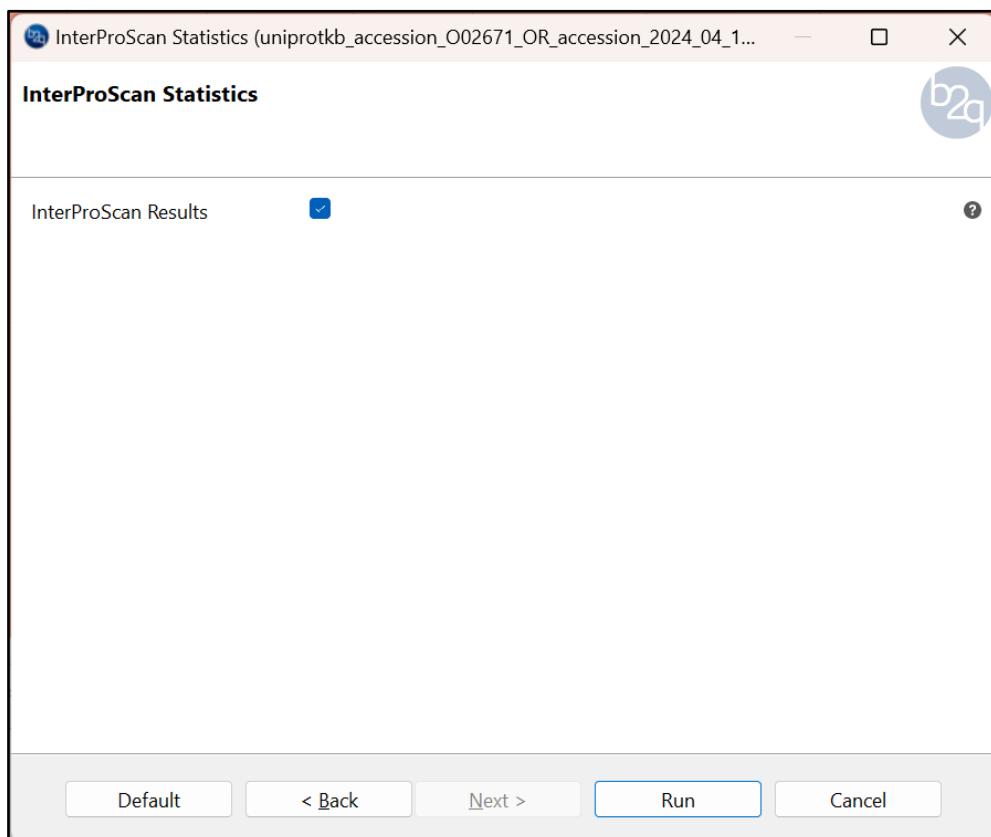


Fig 21: InterProScan Statistics and hit run

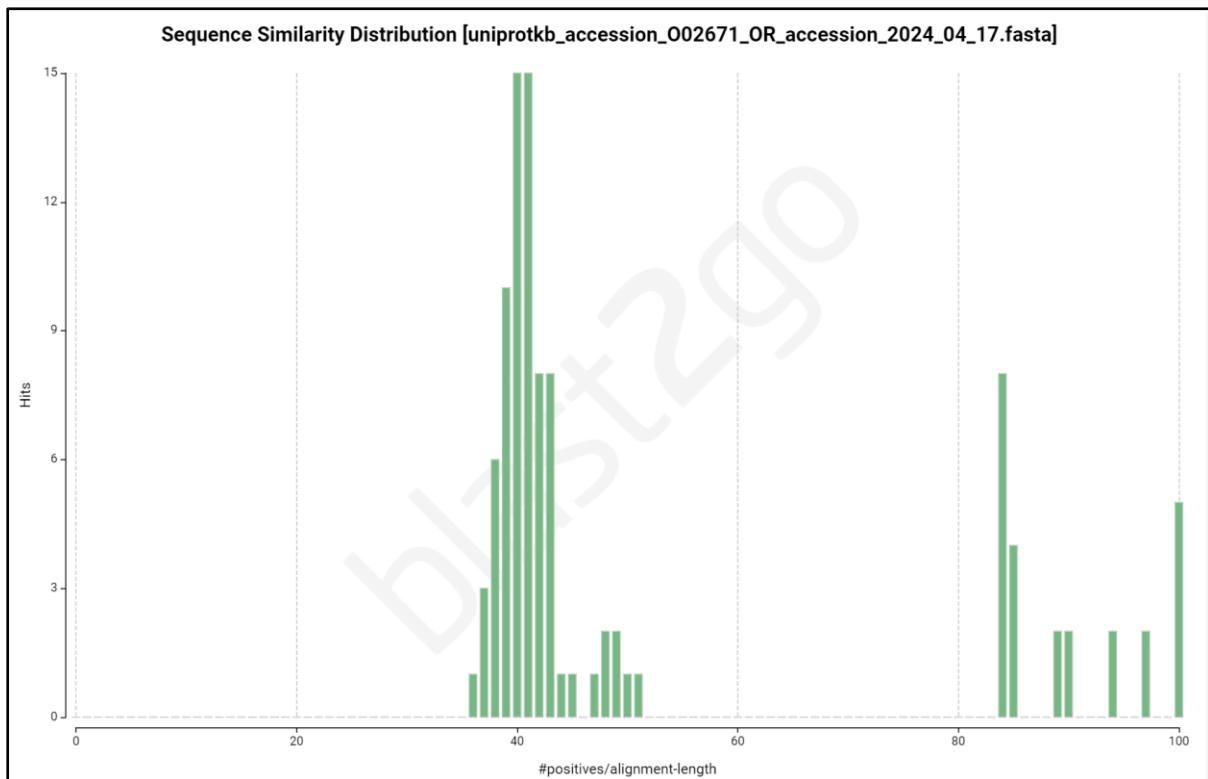


Fig 22: Sequence similarity distribution chart

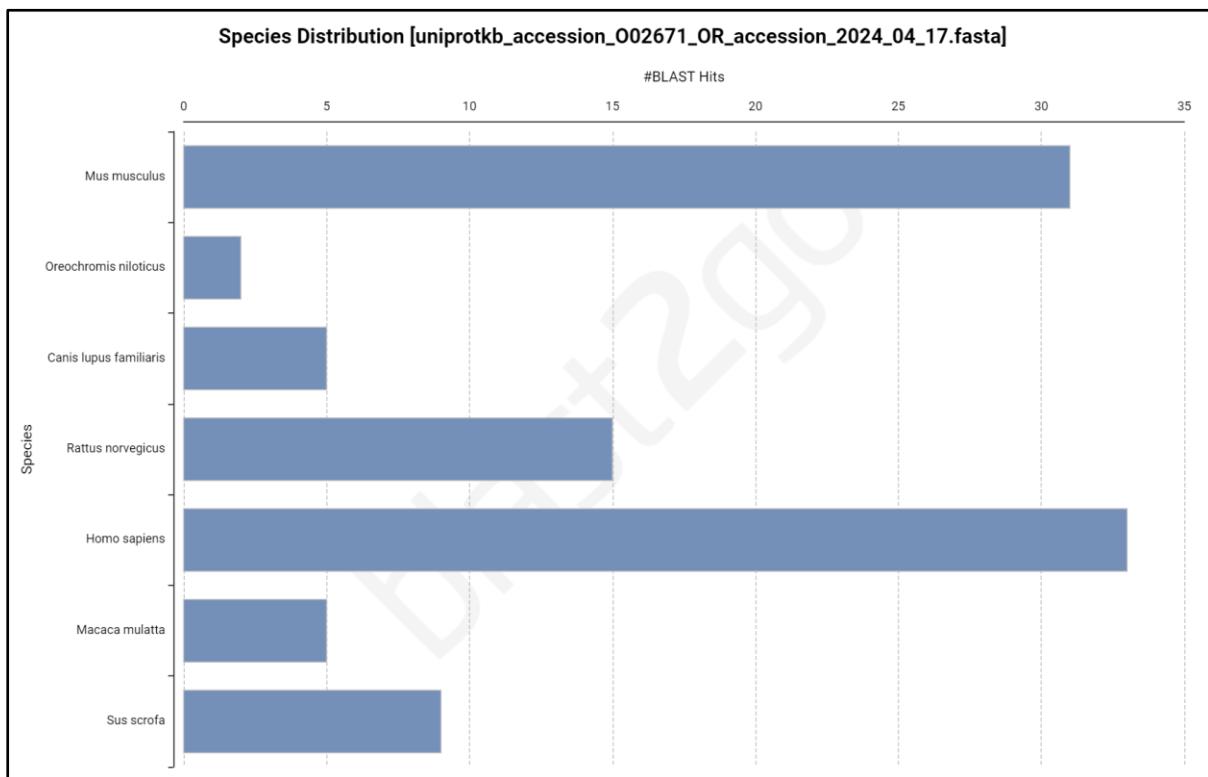


Fig 23: Species distribution chart

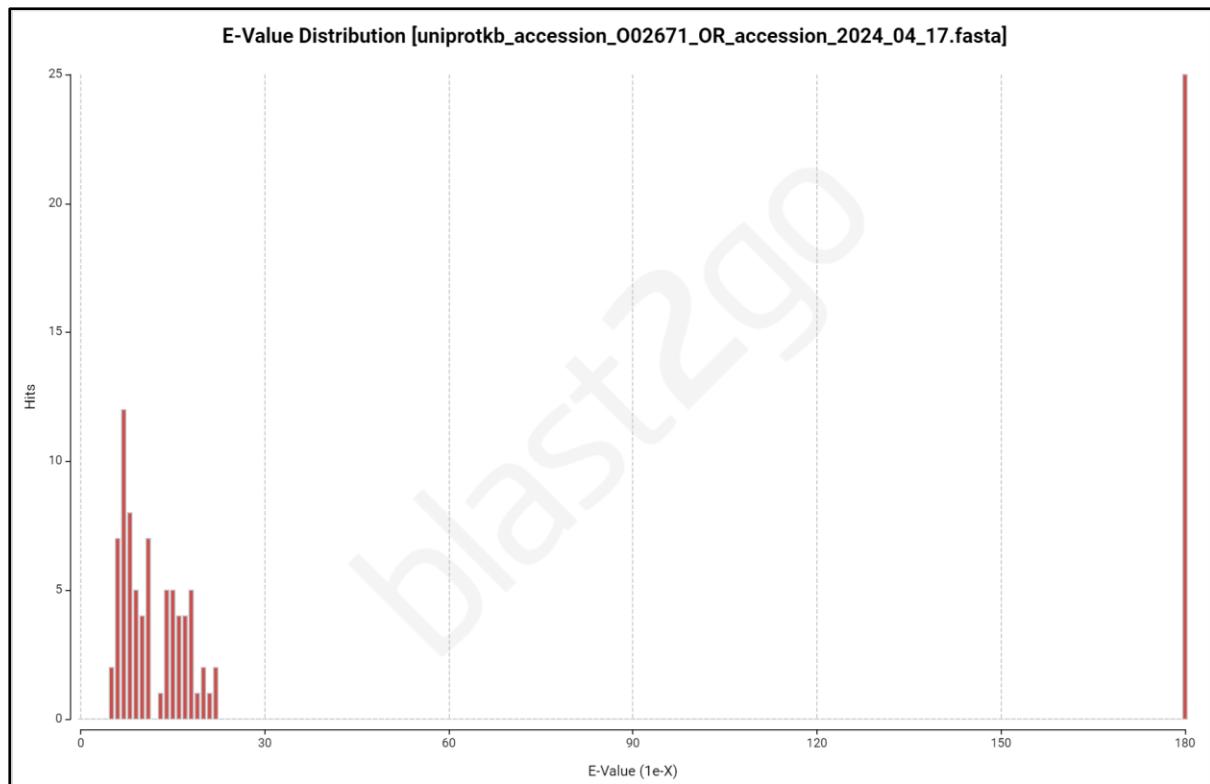


Fig 24: E-value distribution



Fig 25: Top-Hit Species distribution

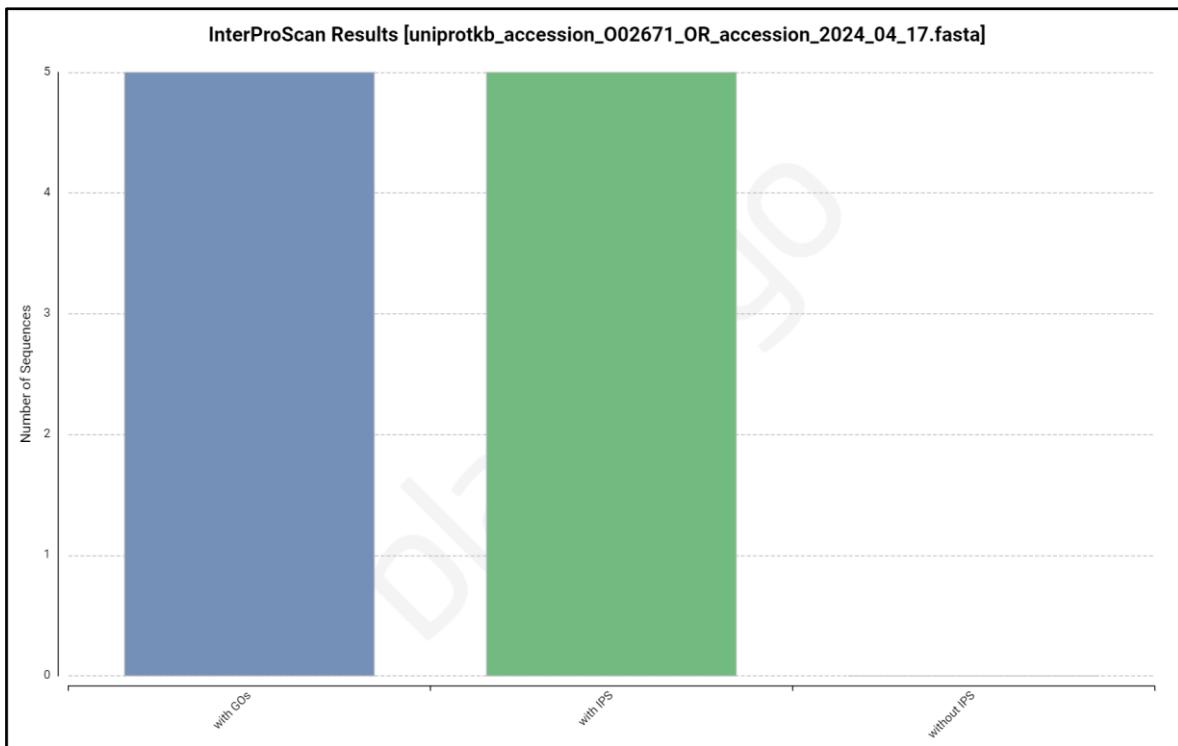


Fig 26: InterProScan results

RESULTS:

In the Blast2GO results that we have gotten via statistics, a sequence similarity chart was obtained and for which alignment length for 40 received the maximum amount of hits which was 15. In species distribution chart, 7 organisms were matched and *Homo sapiens* received 33 hits. A E-value distribution chart was also obtained and a top species was obtained with it. For InterPro, 5 sequences were obtained for both with GOs and with IPS.

CONCLUSION:

Blast2GO database was explored for annotating and analysing our query ‘LEPR’.

REFERENCES:

1. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674–3676. <https://doi.org/10.1093/bioinformatics/bti610>
2. blast2go [Babelomics tutorial]. (2017, May 24). <http://wikis.babelomics.org/babelomicstutorial/blast2go>
3. Robert J. Denver, Ronald M. Bonett, Graham C. Boorse; Evolution of Leptin Structure and Function. Neuroendocrinology 1 July 2011; 94 (1): 21–38. <https://doi.org/10.1159/000328435>
4. Diéguez-Campa, C. E., Angel-Chávez, L. I., Reyes-Ruvalcaba, D., Talavera-Zermeño, M. J., Armendáriz-Cabral, D. A., Torres-Muro, D., & Pérez-Neri, I. (2020). Leptin Levels and Q223R Leptin Receptor Gene Polymorphism in Obese Mexican Young Adults. EJIFCC, 31(3), 197–207.

DATE: 16/03/2024

WEBLEM 6

**INTRODUCTION OF SINGLE NUCLEOTIDE POLYMORPHISM
DATABASE (dbSNP)**

SNP database is indeed a crucial resource in genetics and genomics, serving as a repository for various genetic variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations. This database plays a significant role in understanding genetic diversity within populations, identifying disease associations, and studying variations in genes that can influence health and disease outcomes.

The dbSNP database includes a broad collection of simple genetic polymorphisms, encompassing single-base nucleotide substitutions (SNPs), small-scale multi-base deletions or insertions (DIPs), retroposable element insertions, and microsatellite repeat variations (STRs). Each entry in dbSNP provides essential information such as the sequence context of the polymorphism, the frequency of occurrence (by population or individual), and details about the experimental methods used to assay the variation.

Researchers utilize dbSNP for gene mapping, defining population structure, conducting functional studies, and exploring the genetic basis of complex disorders like heart disease and cancer. By leveraging the data in dbSNP, scientists can gain insights into the genetic underpinnings of diseases, population genetics, and the relationship between genetic variations and phenotypic characteristics.

Based on the provided sources, entries in dbSNP contain detailed information on nucleotide variations, flanking sequences, genotypes, frequencies, and more. Each submission in dbSNP is uniquely identified by an ID starting with "ss" (submitted SNP). The database offers effective search methods through Entrez, including text queries, a Limits page for refined searches, and a Preview/Index page. Additionally, dbSNP integrates submitted and computed data, requiring a minimum sequence length of 100 nucleotides to ensure accurate mapping on the reference genome. Computed information is obtained through a process called the "build," enhancing the database's comprehensiveness and usefulness in genetic research.

Originally intended to support large-scale polymorphism discovery projects like the HapMap project, dbSNP has expanded into a comprehensive repository for various genetic variations. Despite its name, dbSNP houses both common and rare variations, including clinically significant human variations and benign polymorphisms. The database not only stores genetic variant information but also provides data on population-specific allele frequencies, individual genotypes, validation status, and significant genome-wide association results. Furthermore, dbSNP collaborates with locus-specific databases and clinical diagnostic laboratories to integrate genetic variant data with clinical relevance. For those interested in contributing to or accessing information from dbSNP, an automated web submission portal is available along with user support services for inquiries.

dbSNP Reference SNP (rs or RefSNP) number is a locus accession for a variant type assigned by dbSNP. The RefSNP catalog is a non-redundant collection of submitted variants which were clustered, integrated and annotated. RefSNP number is the stable accession regardless of the differences in genomic assemblies. RefSNP numbers facilitate large-scale studies in

association genetics, medical genetics, functional and pharmaco-genomics, population genetics and evolutionary biology, personal genomics, and precision medicine. They provide a stable variant notation for mutation and polymorphism analysis, annotation, reporting, data mining, and data integration.

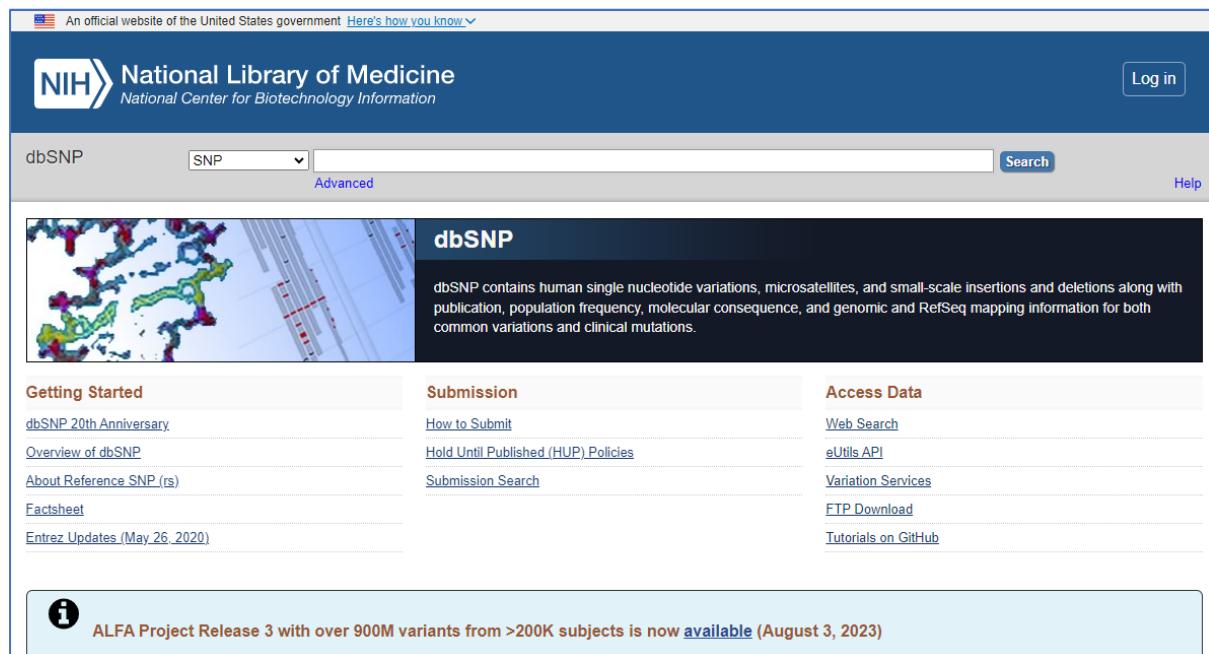


Fig 1: Homepage of dbSNP

Distinguishing RefSNP Features:

1. Non-redundancy and globally unique accession series.
2. Composed from over 2 billion Submitted SNP (ss) from thousands of submitters.
3. More than 20 years of tracking histories for all assigned, merged, and deleted RefSNP.
4. Annotated and linked to the latest human assembly and RefSNP nucleotide and protein sequences.
5. Updates to reflect current knowledge of sequence data and biology data validation.
6. Ongoing curation and annotation by NCBI staff and collaborators.
7. Searchable across variation and genomic databases
8. Supported and reported in open-source and commercial software and tools. Over 400K RefSNP are in CLIN Var
9. Cited in over 51K publications with biological, functional, disease, and clinical information for variants across the genomes.
10. Linked to many NCBI internal and external resources such as CLIN Var, PubMed, PubMed Central, RefSeq, UCSC, EBI, Top Med, and GnomAD.

Supports consistent reporting and non-redundant variation annotations across related sequences including alternate haplotypes, GRC patches, and future graph genomes if the alignment or sequence relationship is known.

dbSNP having two Accession Types:

1. **Submitted SNP (ss):** Submitted variant based on asserted location or flanking sequences.
2. **Reference SNP (rs):** Non-redundant set of variations based on clustering of SS'es of same variant type and sequence position.

SNP database provides various search methods, including text-based and sequence-based searches, enabling researchers to effectively navigate and retrieve genetic information. Detailed data on nucleotide variations, flanking sequences, genotypes, and frequencies are available for in-depth genetic analysis.

This database integrates submissions with other NCBI resources such as GenBank, PubMed, Locus Link, and the Human Genome Project data. This integration enhances the depth and utility of genetic data available for research purposes.

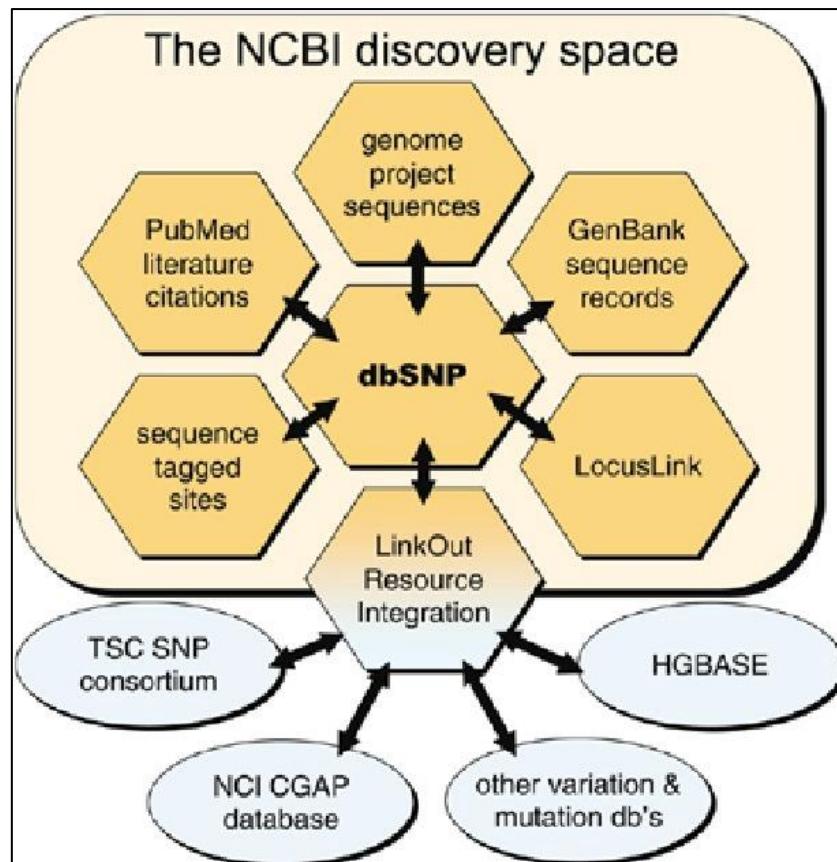


Fig 2: Other NCBI resources and out link resources for dbSNP integration

REFERENCES:

1. Kitts, A., & Sherry, S. (2011). The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. [https://www.semanticscholar.org/paper/The-Single-Nucleotide-Polymorphism-Database-\(dbSNP\)-Kitts-Sherry/c11be17e42b680487bf41949570e1640b30f7ee7](https://www.semanticscholar.org/paper/The-Single-Nucleotide-Polymorphism-Database-(dbSNP)-Kitts-Sherry/c11be17e42b680487bf41949570e1640b30f7ee7)
2. BMC Bioinformatics. (2024, March 21). BioMed Central. <https://bmcbioinformatics.biomedcentral.com/>

DATE: 16/03/2024

WEBLEM 6(A)

dbSNP (Single Nucleotide Polymorphism database)

(URL:<https://www.ncbi.nlm.nih.gov/snp/>)

AIM:

To identify and analyze genetic variations (mutational gene) in the query 'DDIT3' (Reference SNP Report: rs28382352) using dbSNP (SNP database).

INTRODUCTION:

The dbSNP (Single Nucleotide Polymorphism Database) curated by the National Center for Biotechnology Information (NCBI) is indeed a pivotal resource in genetic research. It serves as a comprehensive repository of genetic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations, offering researchers a rich dataset for genetic analysis. DbSNP plays a crucial role in supporting large-scale polymorphism discovery and provides detailed information on population-specific allele frequencies, individual genotypes, validation status, and significant genome-wide association results. This database integrates with other NCBI resources like GenBank, PubMed, LocusLink, and the Human Genome Project data, enhancing the utility and depth of genetic data available for research purpose. Researchers and students exploring genetic diversity, disease association studies, and population genetics benefit significantly from dbSNP's wealth of information. It offers valuable insights into the genetic landscape of populations and the underlying genetic basis of traits and diseases, making it an indispensable tool in the field of genetic research.

DDIT3:

The DDIT3 gene, also known as CHOP or GADD153, encodes a member of the CCAAT/enhancer-binding protein (C/EBP) family of transcription factors. Located on chromosome 12, it regulates diverse biological processes such as transcriptional regulation, blood vessel maturation, and cellular responses to various stressors like DNA damage and endoplasmic reticulum stress. Its involvement extends to cell cycle regulation, Wnt signaling, and modulation of interleukin-8 production, among others.

Functionally, DDIT3 participates in transcriptional regulation by binding to specific DNA sequences near RNA polymerase II promoters and modulating gene expression. It acts as a transcriptional activator or repressor depending on the context, exerting its effects through protein-protein interactions and DNA binding. Moreover, DDIT3 plays roles in cellular processes such as protein binding, homodimerization, and interaction with other transcription factors like cAMP response element-binding protein (CREB). Its leucine zipper domain facilitates dimerization, essential for its transcriptional activity. Overall, DDIT3's multifaceted functions highlight its significance in cellular homeostasis and stress response pathways.

METHODOLOGY:

1. Go to NCBI dbSNP web page.
2. Use the search box in dbSNP to input ‘DDIT3’ query. You can search using text words or phrases related to genetic variations, SNPs, or specific genes.
3. Click on a specific SNP ID or genetic variation from the search results to open an entry.
4. Now interpret the results to understand how the genetic variant found in dbSNP could be linked to biological functions, diseases, or other important factors related to the query.

OBSERVATIONS:

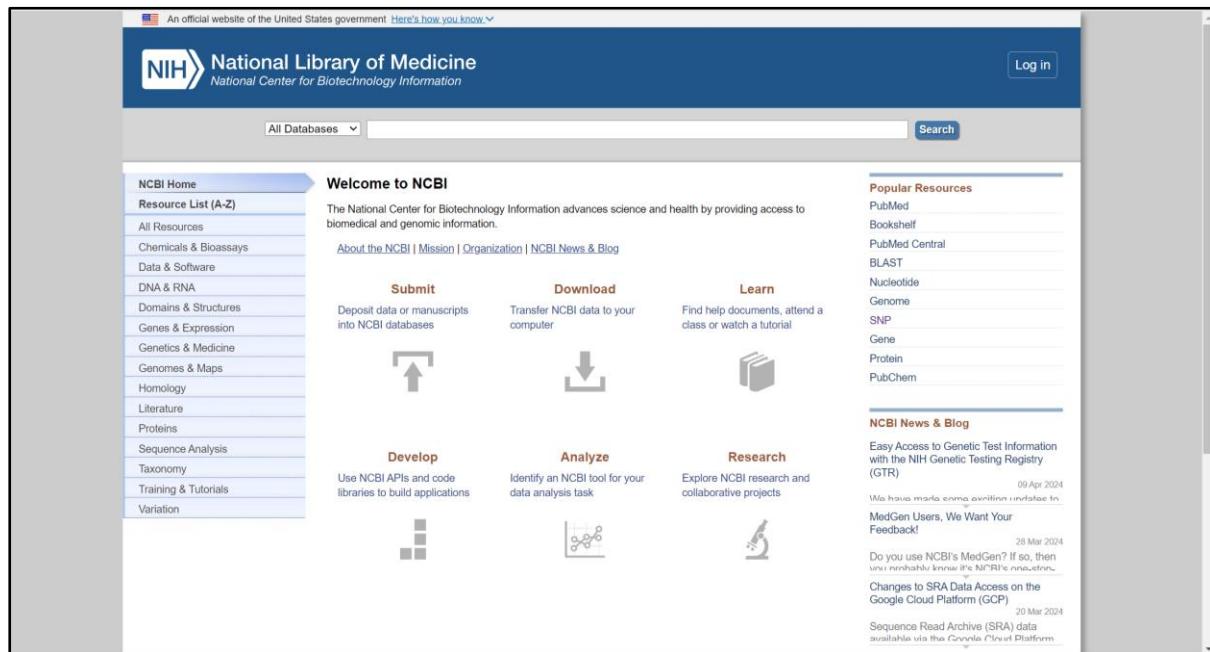


Fig 1: Homepage of NCBI database

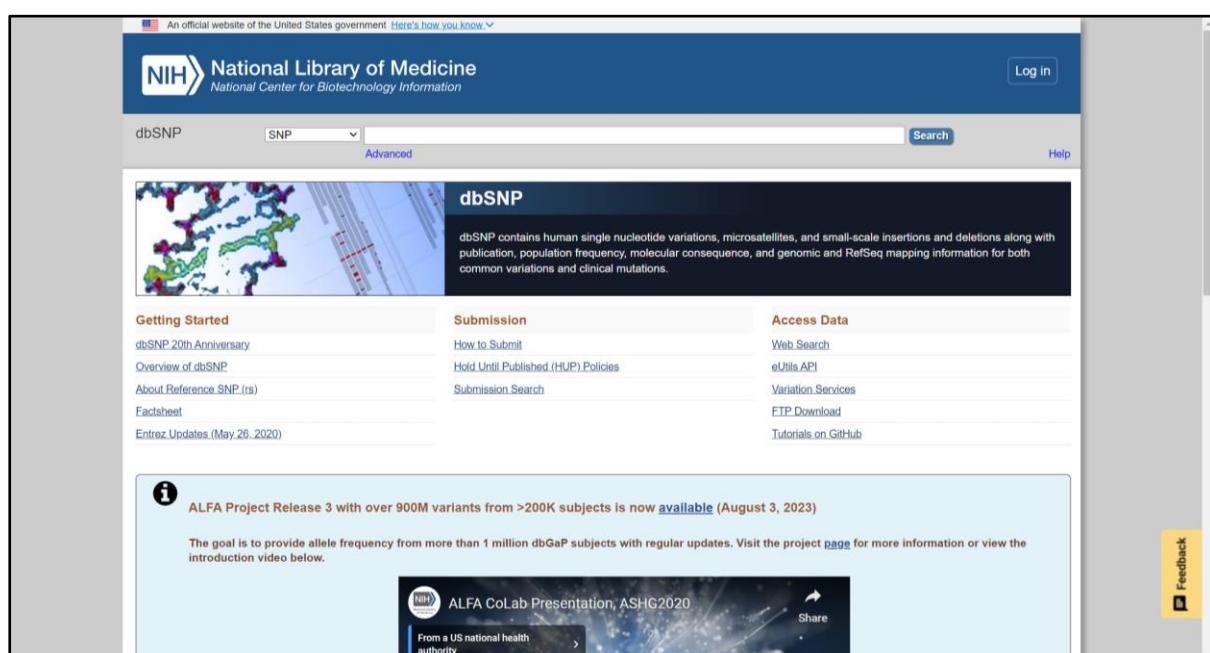


Fig 2: Homepage of dbSNP database

Fig 3: Results obtained after searching for the query ‘DDIT3’

Fig 4: Studying the entry: rs28382352

The screenshot shows the 'Frequency' section of the ALFA project. At the top, there is a navigation bar with tabs: Frequency (highlighted with a red box), Variant Details, Clinical Significance, HGVS, Submissions, History, Publications, and Flanks. Below the navigation bar, the title 'ALFA Allele Frequency' is displayed, followed by a brief description of the aggregate allele frequency from dbGaP. The release version is noted as 20230706150541. A search bar is present at the top right.

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	23482	G=0.99225	C=0.00775
European	Sub	15936	G=0.99969	C=0.00031
African	Sub	3554	G=0.9584	C=0.0416
African Others	Sub	122	G=0.951	C=0.049
African American	Sub	3432	G=0.9586	C=0.0414
Asian	Sub	168	G=1.000	C=0.000
East Asian	Sub	112	G=1.000	C=0.000
Other Asian	Sub	56	G=1.00	C=0.00
Latin American 1	Sub	146	G=0.979	C=0.021
Latin American 2	Sub	610	G=0.997	C=0.003
South Asian	Sub	98	G=1.00	C=0.00

At the bottom, there is a download button labeled 'Download' and a feedback button labeled 'Feedback'. A search bar is also located at the bottom right.

Fig 5: View of the ‘Frequency’ section

The screenshot shows the 'Variant Details' section. At the top, there is a navigation bar with tabs: Frequency, Variant Details (highlighted with a red box), Clinical Significance, HGVS, Submissions, History, Publications, and Flanks. Below the navigation bar, the title 'Genomic Placements' is displayed, followed by a table of sequence names and their changes.

Sequence name	Change
DDIT3 RefSeqGene	NG_027674.1:g.8055C>G
GRCh37.p13 chr 12	NC_000012.11:g.57911246G>C
GRCh38.p14 chr 12	NC_000012.12:g.57517463G>C
MARS1 RefSeqGene	NG_034077.1:g.34511G>C

Below this, a table for the gene 'DDIT3, DNA damage inducible transcript 3 (minus strand)' is shown. It lists variants by molecule type, change, amino acid/codon, and SO term.

Molecule type	Change	Amino acid[Codon]	SO Term
DDIT3 transcript variant 1	NM_001195053.1:c.13C>G	P [CCA] > A [GCA]	Coding Sequence Variant
DDIT3 transcript variant 2	NM_001195054.1:c.13C>G	P [CCA] > A [GCA]	Coding Sequence Variant
DDIT3 transcript variant 3	NM_001195055.1:c.13C>G	P [CCA] > A [GCA]	Coding Sequence Variant
DDIT3 transcript variant 4	NM_001195056.1:c.13C>G	P [CCA] > A [GCA]	Coding Sequence Variant
DDIT3 transcript variant 5	NM_004083.6:c.-32-25C>G	N/A	Intron Variant
DDIT3 transcript variant 6	NM_001195057.1:c.-18-39C>G	N/A	Intron Variant
DDIT3 transcript variant X1	XM_047428446.1:c.86-25C>G	N/A	Intron Variant
DNA damage-inducible transcript 3 protein isoform 1	NP_001181982.1:p.Pro5Ala	P (Pro) > A (Ala)	Missense Variant

At the bottom, there is a feedback button labeled 'Feedback'.

Fig 6: View of the ‘Variant details’ section

Fig 7: View of the ‘Clinical Significance’ section

Fig 8: View of the ‘HGVS’ section

Fig 9: View of the ‘Submission’ section

Fig 10: View of the ‘History’ section

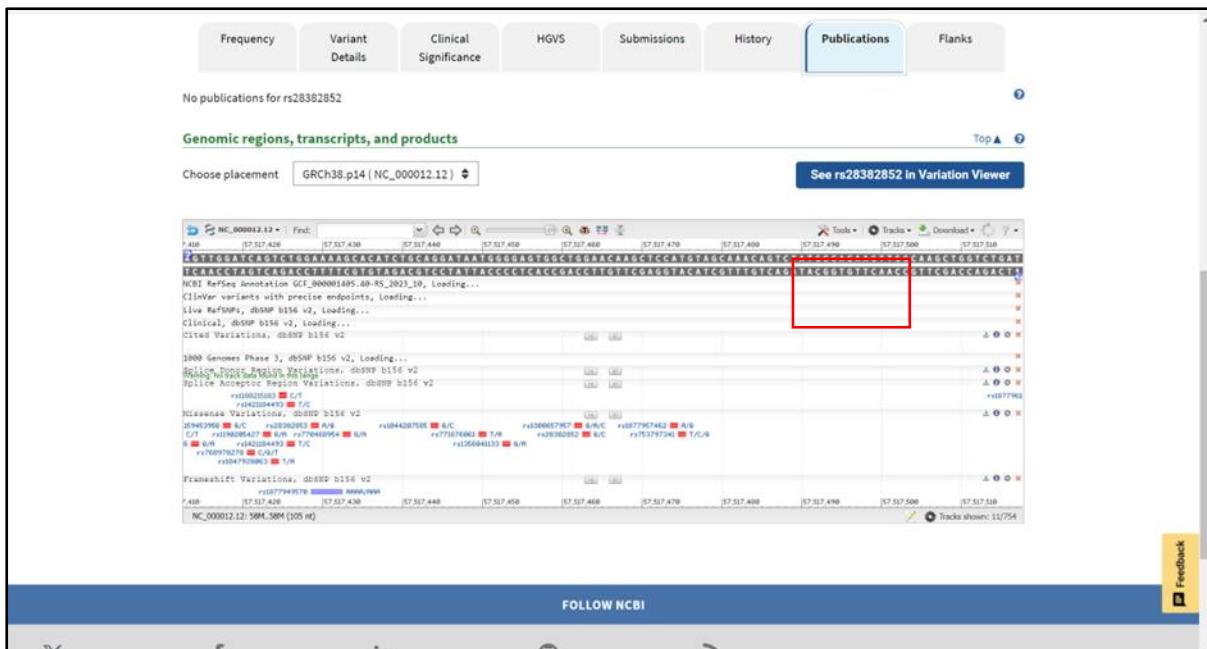


Fig 11: View of the ‘Publication’ section

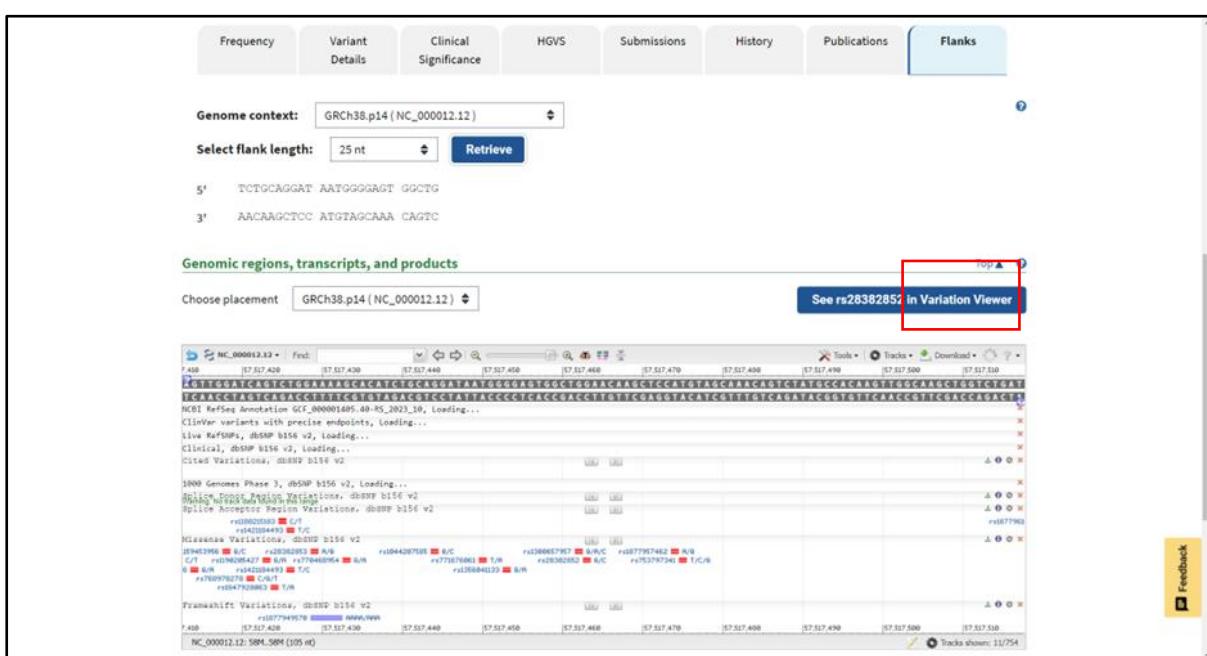


Fig 13: View of the ‘Flanks’ section

RESULTS:

We gained valuable insights into the genetic variations of interest for the query ‘DDIT3’ (Reference SNP Report: rs28382352). We examined the allele frequencies in various populations to understand the variant’s prevalence across ethnicities. Pinpointing the exact genetic change was achieved by identifying the specific location and nucleotide change within the SNP(s) on the chromosome. Additionally, dbSNP annotations and external sources were used to assess the potential impact of the SNP(s) on gene function and human health, providing clues about disease or trait associations. For variants with potential clinical relevance, we

investigated their presence in the Human Disease Variant Submission (HDVS) database. Furthermore, the history section offered a timeline of the SNP(s) within dbSNP, including submission dates, updates, and annotation changes. To delve deeper into the functional significance, we identified relevant scientific publications associated with the SNP(s). Finally, analyzing the flanking DNA sequences surrounding the SNP(s) provided insights into potential regulatory elements or functional motifs affected by the variant.

CONCLUSION:

This comprehensive analysis of the Frequency section, Variant details, Clinical Significance, HDVS, History and Flanks section within dbSNP allowed for a thorough understanding of the SNP(s) for the query ‘DDIT3’ (Reference SNP Report: rs28382352’). This knowledge can be a springboard for further research endeavours in disease association studies, personalized medicine, and functional genomics. However, it’s important to acknowledge that dbSNP annotations may have limitations, and further exploration using external resources is recommended for a complete picture. Depending on the specific research question, future directions could involve functional assays or computational tools to analyze the SNP(s) in greater detail.

REFERENCES:

1. Sherry, S. T., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotnik, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308–311. <https://doi.org/10.1093/nar/29.1.308>
2. RefSNP About. (n.d.). https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/
3. Reščenko, R., Brīvība, M., Atava, I., Rovīte, V., Pečulis, R., Silamiķelis, I., Ansone, L., Megnis, K., Birzniece, L., Leja, M., Xu, L., Xiao-Li, S., Zhou, Y., Slaitas, A., Hou, Y., & Kloviņš, J. (2023). Whole-Genome Sequencing of 502 Individuals from Latvia: The First Step towards a Population-Specific Reference of Genetic Variation. *International Journal of Molecular Sciences*, 24(20), 15345. <https://doi.org/10.3390/ijms242015345>
4. Diaz-Perez, J. A., & Kerr, D. A. (2023, November 22). Gene of the month: DDIT3. *Journal of Clinical Pathology*, 77(4), 211–216. <https://doi.org/10.1136/jcp-2023-208963>
5. Ahluwalia, T. S., Troelsen, J. T., Balslev-Harder, M., Bork-Jensen, J., Thuesen, B. H., Cerqueira, C., Linneberg, A., Grarup, N., Pedersen, O., Hansen, T., & Dalgaard, L. T. (2016, September 14). Carriers of aVEGFAenhancer polymorphism selectively binding CHOP/DDIT3 are predisposed to increased circulating levels of thyroid-stimulating hormone. *Journal of Medical Genetics*, 54(3), 166–175. <https://doi.org/10.1136/jmedgenet-2016-104084>
6. Merk, M., Zierow, S., Leng, L., Das, R., Du, X., Schulte, W., Fan, J., Lue, H., Chen, Y., Xiong, H., Chagnon, F., Bernhagen, J., Lolis, E., Mor, G., Lesur, O., & Bucala, R. (2011, August 4). The D -dopachrome tautomerase (DDT) gene product is a cytokine and functional homolog of macrophage migration inhibitory factor (MIF). *Proceedings of the National Academy of Sciences*, 108(34). <https://doi.org/10.1073/pnas.1102941108>

WEBLEM 7
INTRODUCTION TO WHOLE GENOME SEQUENCING

Whole genome sequencing (WGS) is a comprehensive method that involves analyzing entire genomes, providing detailed genetic information crucial for identifying inherited disorders, characterizing cancer mutations, and tracking disease outbreaks. This technique is not limited to human genomes but can be applied to various species, including livestock, plants, and microbes. Unlike targeted approaches, WGS offers a comprehensive view of the entire genome, making it ideal for various applications such as identifying causative variants and assembling novel genomes. WGS can detect different types of genetic variations like single nucleotide variants, insertions/deletions, copy number changes, and large structural variants. Recent technological advancements have made whole-genome sequencing more efficient than ever before, enabling detailed insights into genetic codes.

WGS has significantly improved disease detection and outbreak investigations by providing precise data for identifying outbreaks early and characterizing bacteria to enhance surveillance efforts. It has become a standard method for detecting foodborne outbreaks associated with various bacteria like *Campylobacter*, *E. coli*, *Salmonella*, *Vibrio*, and *Listeria* since 2019. By comparing sequences from multiple bacteria and identifying differences, scientists can determine the relatedness of bacteria and detect outbreaks more effectively. The implementation of WGS in public health laboratories has enhanced surveillance for foodborne diseases and antimicrobial resistance, leading to quicker diagnoses and improved outbreak prevention.

Types of tools used for Whole Genome Sequencing:

A. MUMmer:

The MUMmer tool is a versatile software package designed for the rapid alignment of large DNA and amino acid sequences, offering a range of functionalities for genome analysis. It consists of various components that work together to produce desired outputs, with the main program being 'mummer,' which can quickly find exact matches of a specified length between sequences. Additionally, MUMmer includes scripts like 'run-mummer1,' 'run-mummer3,' 'nucmer,' and 'promer' that cluster matches and align non-exact regions using a modified Smith-Waterman algorithm, enhancing the analysis of genetic data. The tool is open-source, allowing free access to its source code for both non-profit and for-profit users, making it a valuable resource for genetic analysis and research.

MUMmer is particularly useful for aligning entire genomes, whether complete or in draft form, and can efficiently handle large datasets, making it suitable for various applications such as identifying genetic variations, comparing genomes, and detecting similarities between sequences. The tool's ability to align incomplete genomes, handle numerous contigs from sequencing projects, and generate alignments based on translations of input sequences makes it a powerful resource for genetic analysis and comparison. Overall, MUMmer is a valuable

tool in genomics, offering rapid and accurate alignment capabilities for a wide range of genetic analyses.

MUMmer is a system for rapidly aligning large DNA sequences to one another. It can align:

- whole genomes to other genomes
- large genome assemblies to one another
- partial (draft) genomes sequences to one another
- or (with release 4) a set of reads to a genome

MUMmer is very fast and easy to run. The current version, release 4.x, can find all 20-bp maximal exact matches between two bacterial genomes in just a few seconds on a typical desktop or laptop computer. MUMmer handles the 100s or 1000s of contigs from a draft genome with ease, and will align them to another set of contigs using the nucmer utility included with the system. The promer utility takes this a step further by generating alignments based upon the six-frame translations of both input sequences. Most of the improvements in release 4.x are to nucmer and MUMmer. Promer is unchanged from release 3.0. See the nucmer and promer readme files in the "docs" subdirectory for more details.

MUMmer is free and open source, so all we ask is that you cite our most recent paper in any publications that use it. Here are the relevant publications:

Version 4.0+
MUMmer4 and nucmer4 are described in "[MUMmer4: A fast and versatile genome alignment system](#)" G. Marçais, A.L. Delcher, A.M. Phillippy, R. Coston, S.L. Salzberg, A. Zimin, *PLoS computational biology* (2018), 14(1): e1005944.

Version 3.0
Open source MUMmer 3.0 is described in "[Versatile and open software for comparing large genomes](#)" S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg, *Genome Biology* (2004), 5:R12.

Version 2.1
MUMmer 2.1, NUCmer, and PROMer are described in "[Fast Algorithms for Large-scale Genome Alignment and Comparison](#)" A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, *Nucleic Acids Research* (2002), Vol. 30, No. 11 2478-2483.

Version 1.0
MUMmer 1.0 is described in "[Alignment of Whole Genomes](#)" A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg, *Nucleic Acids Research*, 27:11 (1999), 2369-2376.

Fig 1: Homepage of MUMmer tool (Command Line tool, download from here)

B. PipMaker:

PipMaker is a web-based tool designed for aligning and comparing two long genomic DNA sequences to identify conserved segments between them. This tool, available at the World-Wide Web site bio.cse.psu.edu, offers an efficient method for aligning genomic sequences and produces informative output in the form of percent identity plots (pip). The percent identity plot generated by PipMaker displays aligning segments between two sequences, showing both the position relative to the first sequence and the degree of similarity, aiding in the interpretation of alignments. Optional annotations on the plot provide additional information to enhance alignment interpretation, making PipMaker a valuable resource for researchers analyzing genomic sequences. PipMaker utilizes default parameters from the blastz alignment program, which are optimized for human-mouse alignments, providing users with a reliable and accurate method for comparing large genomic sequences. This tool is particularly useful for identifying conserved regions between genomes, facilitating comparative genomics studies and aiding in the understanding of genetic similarities and differences across species. Overall, PipMaker serves as a user-friendly platform for researchers to perform detailed genomic sequence alignments efficiently and interpret the results effectively.

← → ⌂ Not secure pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic

PipMaker ([instructions](#)) aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment.

- First sequence (FASTA format):

or filename (file must be plain text only):
 No file chosen
- Second sequence (FASTA format):

or filename (file must be plain text only):
 No file chosen
- Your email address:
- Optional features:
 - First sequence mask:

 No file chosen
 - First sequence exons:

 No file chosen

[Privacy policy](#)

Email the authors at <pipmaster@bio.cse.psu.edu>

Development and maintenance of PipMaker is supported by grant HG02238 from the National Human Genome Research Institute.

If you publish results obtained using PipMaker, please cite [Schwartz et al., Genome Research 10:577-586, April 2000](#).

Fig 2: Homepage of Pipmaker tool

C. VISTA:

The VISTA tool is a comprehensive suite of programs and databases designed for comparative analysis of genomic sequences, offering rich capabilities for researchers in the field of genomics. It provides users with the ability to browse pre-computed whole-genome alignments of large vertebrate genomes and other organisms, submit their sequences for various types of comparative analysis, and obtain detailed comparative analysis results for specific genes of interest. VISTA offers a range of servers like mVISTA, rVISTA, GenomeVISTA, and VISTA Browser, each serving different purposes such as aligning sequences, identifying regulatory transcription factor binding sites, comparing whole genome assemblies, and examining pre-computed alignments among various species. The tool allows researchers to analyze multiple DNA sequence alignments from different species while considering their phylogenetic relationships, align sequences up to 10Mb long, and compare sequences with whole genome assemblies, making it a versatile resource for genetic analysis and comparison. With more than 28 searchable genomes, including vertebrates, non-vertebrates, plants, fungi, bacteria, and others, VISTA continues to expand its database to provide a comprehensive platform for genomic research and analysis. Developed and hosted at the Genomics Division of Lawrence Berkeley National Laboratory, VISTA is supported by various grants and collaborations, making it a valuable tool for researchers in the genomics field.



Fig 3: Homepage of VISTA tool

REFERENCES:

1. Ng, P. C., & Kirkness, E. F. (2010). Whole Genome Sequencing. Methods in Molecular Biology, 215–226. https://doi.org/10.1007/978-1-60327-367-1_12
2. Homepage for MuMMer tool- <https://mummer.sourceforge.net/>
3. Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics, 16, 1046–1047.
4. Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018, January 26). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1), e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
5. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., & Miller, W. (2000, April 1). PipMaker—A Web Server for Aligning Two Genomic DNA Sequences. *Genome Research*, 10(4), 577–586.
<https://doi.org/10.1101/gr.10.4.577>

DATE:18/03/2024

WEBLEM 7(A)

PipMaker Tool

(URL: <http://pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic>)

AIM:

To compare genomic sequences, identify conserved regions, and visualize sequence alignments for query ‘ADD1’ (Accession ID: NM_001354759.2 and NM_001354756.2) using PipMaker Tool.

INTRODUCTION:

PipMaker is a web server used for aligning two genomic DNA sequences, allowing for the comparison of DNA sequences and identification of conserved segments between them. It provides functionalities such as generating alignments and percent identity plots (pips) to visualize similarities between sequences.

PipMaker is a World-Wide Web site for comparing two long DNA sequences to identify conserved segments and for producing informative, high-resolution displays of the resulting alignments. One display is a percent identity plot (pip), which shows both the position in one sequence and the degree of similarity for each aligning segment between the two sequences in a compact and easily understandable form. Positions along the horizontal axis can be labeled with features such as exons of genes and repetitive elements, and colors can be used to clarify and enhance the display. The web site also provides a plot of the locations of those segments in both species (similar to a dot plot). PipMaker is appropriate for comparing genomic sequences from any two related species, although the types of information that can be inferred (e.g., protein-coding regions and cis-regulatory elements) depend on the level of conservation and the time and divergence rate since the separation of the species. Gene regulatory elements are often detectable as similar, noncoding sequences in species that diverged as much as 100–300 million years ago, PipMaker supports analysis of unfinished or “working draft” sequences by permitting one of the two sequences to be in unoriented and unordered contigs.

ADD1 gene:

The ADD1 gene, encoding the α -adducin protein, plays significant roles in both animal and human physiology. In mice, it influences oocyte chromosome meiosis, indicating its involvement in embryonic development. In cattle, polymorphisms within the gene have been linked to growth traits, suggesting its potential as a marker for selective breeding in beef cattle populations.

In humans, genetic variations in ADD1 have been associated with hypertension and economic traits. Studies in North Indian populations and cattle-derived populations in Jeju-do revealed correlations between ADD1 polymorphisms and hypertension, along with economic traits such as meat color and carcass weight. Additionally, in essential hypertension, the G460T polymorphism in the ADD1 gene has been implicated in the new onset of diabetes, especially under the influence of diuretic and other antihypertensive therapies. A meta-analysis further

confirmed the association of the T allele in the ADD1 gene with essential hypertension susceptibility in Asian populations.

These findings underscore the diverse roles of the ADD1 gene in regulating physiological processes across species and its potential implications for health and breeding programs in both animals and humans.

METHODOLOGY:

1. Open Homepage of PipMaker Tool.
2. Take 2 FASTA Sequence from NCBI.
3. Paste the sequence in PipMaker Tool.

OBSERVATIONS:

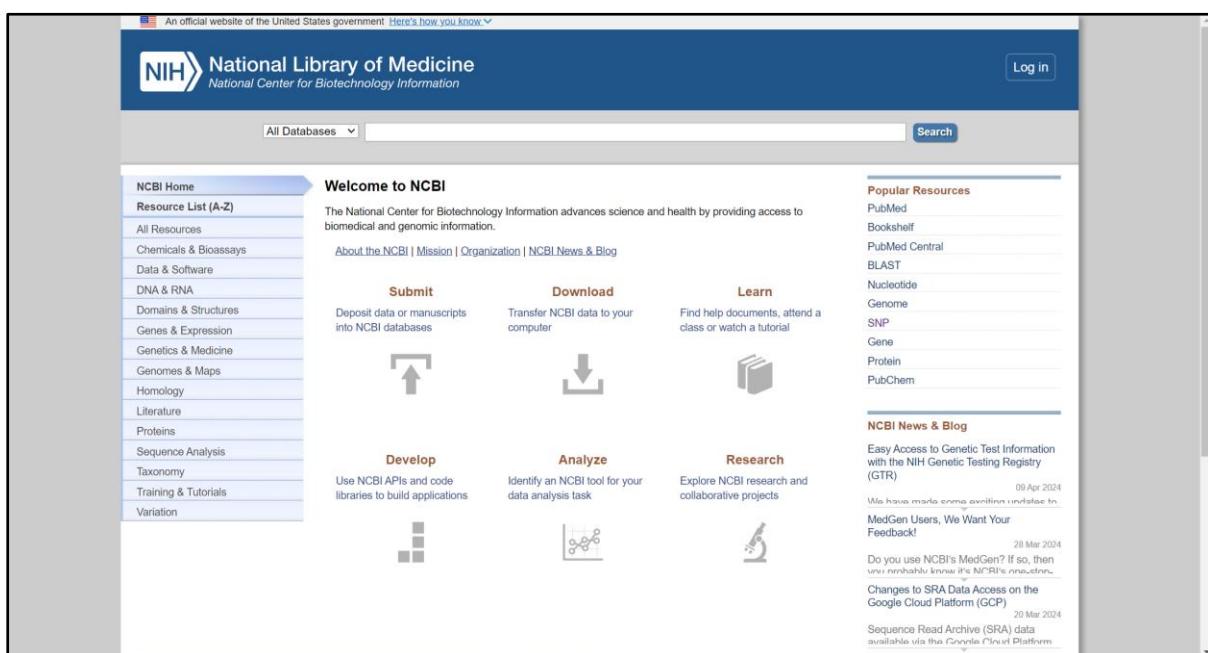


Fig 1: Homepage of NCBI database

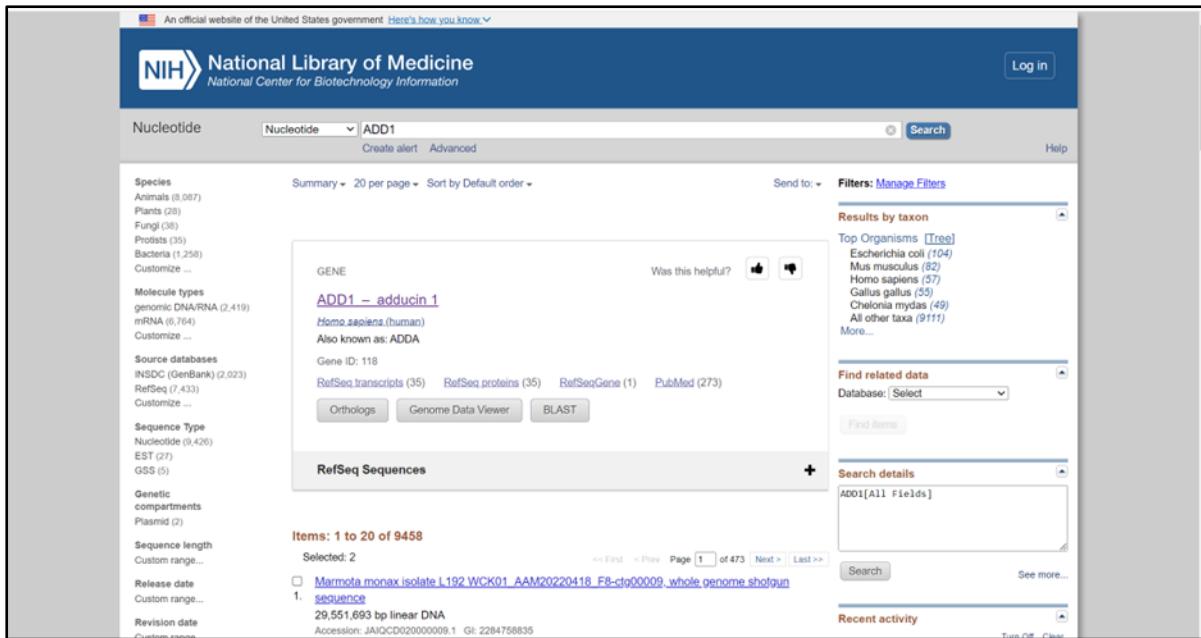


Fig 2: Results obtained for the query ‘ADD1’

Fig 3: FASTA Sequence for the query ‘ADD1’ with Accession ID: NM_001354759.2

Fig 4: FASTA Sequence for the query ‘ADD1’ with Accession ID: NM_001354756.2

PipMaker ([instructions](#)) aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment.

- First sequence (FASTA format):

or filename (file must be plain text only):
 Choose File | No file chosen
- Second sequence (FASTA format):

or filename (file must be plain text only):
 Choose File | No file chosen
- Your email address:
- Optional features:
 - First sequence mask:

Choose File | No file chosen
 - First sequence exons:

Choose File | No file chosen

[Privacy policy](#)

Email the authors at <pipmaster@bio.cse.psu.edu>

Development and maintenance of PipMaker is supported by grant HG02238 from the National Human Genome Research Institute.

If you publish results obtained using PipMaker, please cite [Schwartz et al., Genome Research 10:577-586, April 2000](#).

Fig 5: Homepage of Pipmaker tool

← → ⌂ Not secure pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic

PipMaker ([instructions](#)) aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment.

- First sequence (FASTA format):


```
>NP_001354759.2 Homo sapiens adducin 1 (ADD1), transcript variant 11, mRNA
AGC CGCC CAAGG CCG CACCC AGG TC GGG CGGT GGG GGC GAG CGGA GGG GCT GA GGG CGG AGA GGC
```

 or filename (**file must be plain text only**):
 Choose File [No file chosen]
- Second sequence (FASTA format):


```
>NP_001354756.2 Homo sapiens adducin 1 (ADD1), transcript variant 8, mRNA
AGC CGCC CAAGG CCG CACCC AGG TC GGG CGGT GGG GGC GAG CGGA GGG GCT GA GGG CGG AGA GGC
```

 or filename (**file must be plain text only**):
 Choose File [No file chosen]
- Your email address:
 kothariparthih@gmail.com
- Optional features:
 - First sequence mask:

 Choose File [No file chosen]
 - First sequence exons:

 Choose File [No file chosen]

Submit Reset

[Privacy policy](#)

Email the authors at <pipmaster@bio.cse.psu.edu>

Development and maintenance of PipMaker is supported by grant HG02238 from the National Human Genome Research Institute.

If you publish results obtained using PipMaker, please cite [Schwartz et al., Genome Research 10:577-586, April 2000](#).

Fig 6: Sequences pasted in PipMaker tool

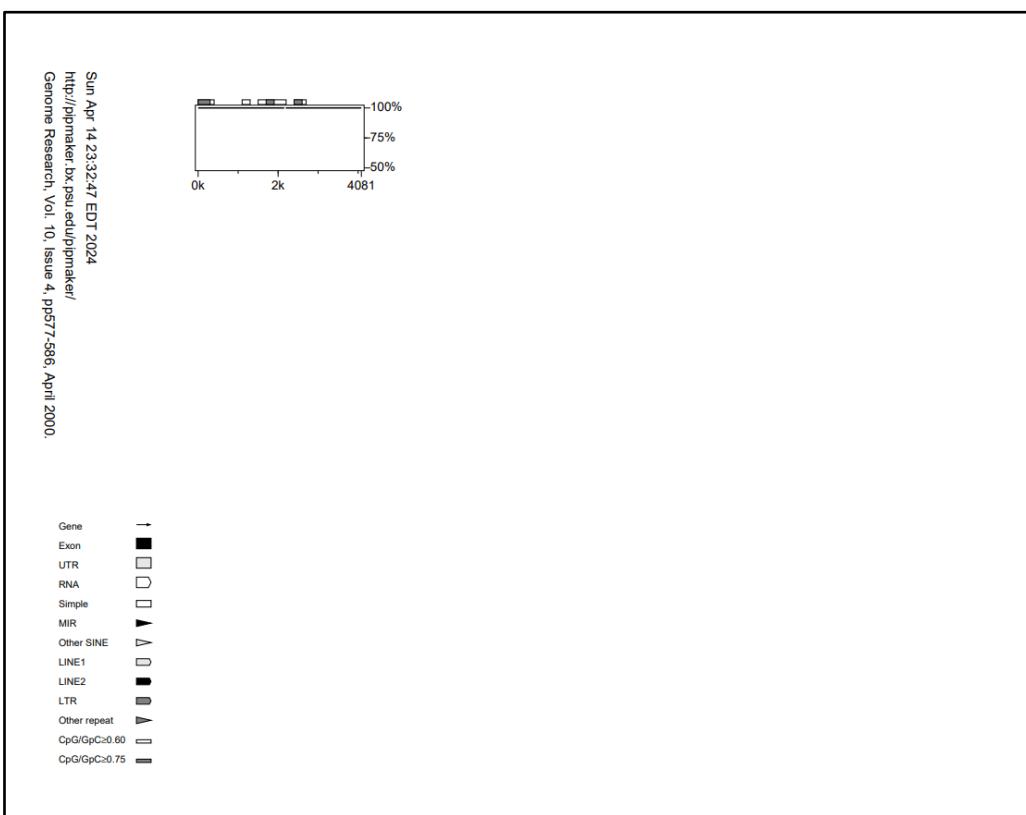


Fig 7: Results obtained for ‘Sequence alignment’

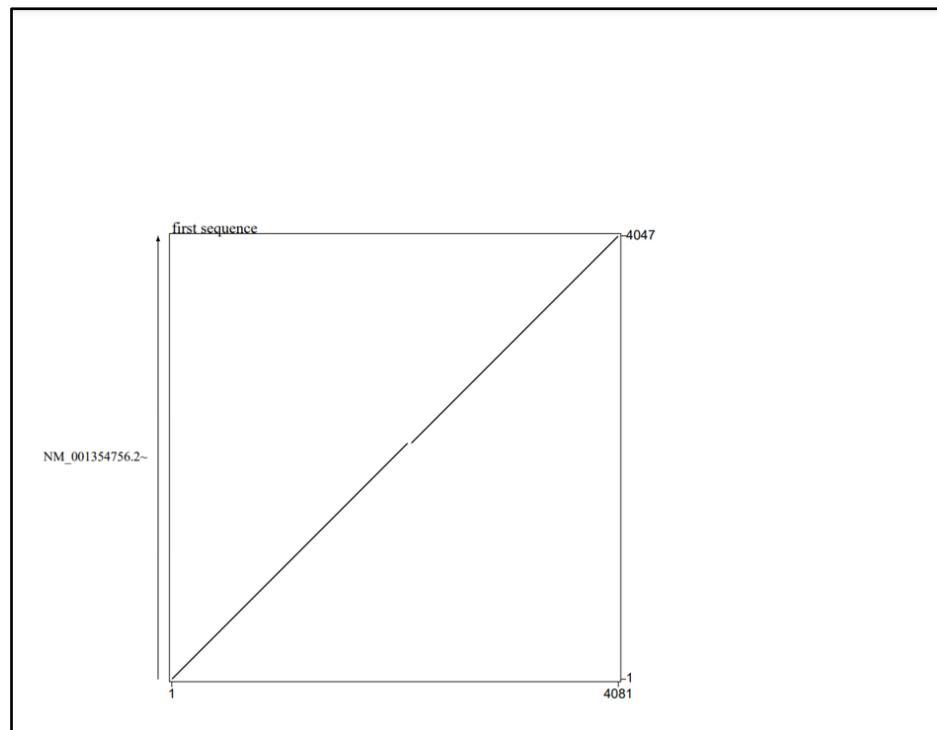


Fig 8: Results obtained for ‘Dot plot’

A screenshot of the "concise" software interface. The window title is "concise". The menu bar includes "File", "Edit", "View", and a settings gear icon. The main text area displays the following sequence alignment information:

```
Aligned to ">NM_001354756.2 Homo sapiens adducin 1 (ADD1), transcript variant 8, mRNA":  
:  
3-2155 <-> 3-2155 100% (2153 nt)  
2190-4079 <-> 2156-4045 100% (1890 nt)
```

The status bar at the bottom shows "Ln 1, Col 1" and "166 characters" on the left, "100%" and "Windows (CRLF)" in the center, and "UTF-8" on the right.

Fig 9: Results obtained for ‘concise’

A screenshot of a terminal window titled "text". The window shows an error message from the PipMaker tool:

```
pipmaker: lat failed; its incomplete output has been deleted.  
+make_blastz_txt out.23536.lav  
/afs/bx.psu.edu/service/web/sites/pipmaker.bx.psu.edu/pipmaker/bin/lat: 9: exec: java: not found
```

The terminal also displays status information at the bottom: "Ln 4, Col 1", "190 characters", "100%", "Windows (CRLF)", and "UTF-8".

Fig 10: Results obtained for ‘text’

A screenshot of a terminal window titled "parameters". The window shows the command-line parameters used for the PipMaker analysis:

```
email: kothariprarthi@gmail.com  
seq1data: bp 4081: >NM_001354759.2 Homo sapiens adducin 1 (ADD1), transcript variant 11, mRNA  
seq2data: bp 4047: >NM_001354756.2 Homo sapiens adducin 1 (ADD1), transcript variant 8, mRNA  
seq1mask: 0 lines  
exons: 0 lines  
underlay: 0 lines  
search strand: both  
coverage: show all matches  
pip title:  
generate pip: yes  
generate dotplot: yes  
data format: PDF  
generate concise text: yes  
generate traditional text: yes  
generate analysis of exons: no  
return raw blastz output: no
```

The terminal also displays status information at the bottom: "Ln 1, Col 1", "503 characters", "100%", "Windows (CRLF)", and "UTF-8".

Fig 11: Results obtained for ‘parameters’

RESULTS:

By exploring PipMaker tool for the query ‘ADD1’ (Accession ID: NM_001354759.2 and NM_001354756.2), files were generated as result of sequence of two entries for the query ‘ADD1’. 100% ,75% 50% sequence alignment were observed. The dot plot showed a long diagonal line indicating similarity with a gap in between. The parameter file showed the information about the program. The query ‘ADD1’ gene sequence showed similarity and number of nucleotides aligned highest percent similarity was found to be 100% with nucleotides aligned.

CONCLUSION:

Genome sequence was compared and conserved regions were identified and sequence alignments were visualized for the query ‘ADD1’ (Accession ID: NM_001354759.2 and NM_001354756.2) by exploring PipMaker tool.

REFERENCES:

1. PipMaker-A Web Server for Aligning Two Genomic DNA Sequences. (n.d.). PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC310868/>
 2. PipMaker: A World Wide Web Server for Genomic Sequence Alignments (2003, February). <https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1002s00>
 3. PipMaker and MultiPipMaker (n.d.). <http://pipmaker.bx.psu.edu/pipmaker/>
 4. Han, S. H., Oh, H. S., Lee, J. B., Jwa, E. S., Kang, Y. J., Kim, S. G., Yang, S. N., Kim, Y. K., Cho, I. C., Cho, W. M., Ko, M. S., & Baek, K. S. (2015, January 30). Effects of Genetic Polymorphisms of ADD1 Gene on Economic Traits in Hanwoo and Jeju Black Cattle-derived Commercial Populations in Jeju-do. *Journal of Life Science*, 25(1), 21–28. <https://doi.org/10.5352/jls.2015.25.1.21>
 5. Huang, Y. Z., Qian, L. N., Wang, J., Zhang, C. L., Fang, X. T., Lei, C. Z., Lan, X. Y., Ma, Y., Bai, Y. Y., Lin, F. P., & Chen, H. (2018, March 12). Genetic Variants in ADD1Gene and their Associations with Growth Traits in Cattle. *Animal Biotechnology*, 30(1), 7–12. <https://doi.org/10.1080/10495398.2017.1398754>
 6. Gupta, S., Jhawat, V., Agarwal, B. K., Roy, P., & Saini, V. (2019, May 29). Alpha Adducin (ADD1) Gene Polymorphism and New Onset of Diabetes Under the Influence of Selective Antihypertensive Therapy in Essential Hypertension. *Current Hypertension Reviews*, 15(2), 123–134. <https://doi.org/10.2174/1573402114666180731111453>
-

DATE:18/03/2024

WEBLEM 7(B)
VISTA Tool
(URL: <https://genome.lbl.gov/vista/index.shtml>)

AIM:

To perform comparative genomics analysis to identify conserved regions and its functional elements for query 'ADD1' (Accession ID: NM_001354759.2) using VISTA tool.

INTRODUCTION:

The VISTA tool in bioinformatics is a powerful set of computational tools designed to facilitate comparative genomics studies. Developed to assist biologists in analyzing DNA sequences from different species, VISTA offers a range of tools for aligning sequences, visualizing conservation patterns, and identifying functional elements in genomes. The VISTA family of tools includes various servers like mVISTA, rVISTA, GenomeVISTA, and VISTA Browser, each serving different purposes such as aligning sequences, discovering regulatory transcription factor binding sites, comparing sequences with whole genome assemblies, and analyzing multiple DNA sequence alignments while considering phylogenetic relationships. VISTA has been widely utilized by the biological community for tasks like comparing gene families, identifying non-coding functional elements, and studying conservation patterns on a whole-genome scale. This tool integrates global alignment strategies and curve-based visualization techniques, making it a valuable resource for researchers in the field of bioinformatics.

There are multiple VISTA servers, each allowing different types of searches.

1. mVISTA can be used to align and compare your sequences to those of multiple other species
2. rVISTA (regulatory VISTA) combines transcription factor binding sites database search with a comparative sequence analysis, the discovery of possible regulatory transcription factor binding sites in regions of their genes of interest. It can be used directly or through mVISTA, Genome VISTA, or VISTA Browser. A database of tissue-specific human enhancers is available through VISTA Enhancer Browser.
3. GenomeVISTA allows the comparison of sequences with whole genome assemblies. It will automatically find the ortholog, obtain the alignment and VISTA plot. It allows the viewing of an alignment together with pre-computed alignments of other species in the same interval.
4. Phylo-VISTA allows the analysis of multiple DNA sequence alignments of sequences from different species while considering their phylogenetic relationships.
5. wgVISTA allows the alignment of sequences up to 10Mb long (finished or draft) including microbial whole-genome assemblies.

Researchers can use the VISTA Browser:

1. To examine pre-computed alignments among a variety of species
2. To submit sequences of their own (not limited by the species collection already in the database)

ADD1 gene:

The ADD1 gene, encoding the α -adducin protein, plays significant roles in both animal and human physiology. In mice, it influences oocyte chromosome meiosis, indicating its involvement in embryonic development. In cattle, polymorphisms within the gene have been linked to growth traits, suggesting its potential as a marker for selective breeding in beef cattle populations.

In humans, genetic variations in ADD1 have been associated with hypertension and economic traits. Studies in North Indian populations and cattle-derived populations in Jeju-do revealed correlations between ADD1 polymorphisms and hypertension, along with economic traits such as meat color and carcass weight. Additionally, in essential hypertension, the G460T polymorphism in the ADD1 gene has been implicated in the new onset of diabetes, especially under the influence of diuretic and other antihypertensive therapies. A meta-analysis further confirmed the association of the T allele in the ADD1 gene with essential hypertension susceptibility in Asian populations.

These findings underscore the diverse roles of the ADD1 gene in regulating physiological processes across species and its potential implications for health and breeding programs in both animals and humans.

METHODOLOGY:

1. Go to GenBank database and search for the query ‘ADD1’ gene and copy the fast sequence.
2. Open Homepage of Vista tool.
3. Paste the FASTA sequence of query ‘ADD1’ on gVISTA tool in the search bar.
4. Click on submit and wait for the results.
5. Interpret the results.

OBSERVATIONS:

The screenshot shows the search interface for the National Library of Medicine's GenBank database. The search term '(ADD1) AND "Homo sapiens"[porgn:__txid9606]' is highlighted with a red box. The results list five entries for Homo sapiens adducin 1 (ADD1) transcript variants 8, 9, 11, 13, and 15. Each entry includes the accession number, protein ID, and taxonomy information. The right sidebar displays search details, recent activity, and filter options.

Fig 1: Query searched for the ADD1 gene in the GenBank database

The screenshot shows the detailed view of the selected ADD1 transcript variant 11 mRNA entry. The sequence is displayed in FASTA format. The right sidebar provides options for changing the region shown, customizing the view, analyzing the sequence, and viewing related articles and reference sequence information.

Fig 2: Copy the FASTA sequence of selected entry with Accession ID: NM_001354759.2

This screenshot shows the elixir bio.tools search results page. At the top, there is a search bar with the placeholder "Search bio.tools". To the right of the search bar are buttons for "Explore", "Login", and "Sign-up". Below the search bar, a message states "This site uses cookies. By continuing to browse the site you are agreeing to our use of cookies. Find out more here." On the left, the "VISTA" tool is listed with its URL "http://genome.lbl.gov/vista/index.shtml" highlighted with a red box. The tool's logo, a blue gear icon with the text "OPENBENCH", is also present. Below the tool entry, there are several categories and filters: "Gene transcripts", "Mapping", "Transcription factors and regulatory sites", "DNA", "Sequence assembly", and "Rare diseases"; "Mature", "Free of charge", and "Open access"; and "Web application", "Database portal", "Java", and icons for Linux, Windows, and Mac OS. A large text block describes VISTA as a comprehensive suite of programs and databases for comparative analysis of genomic sequences. It mentions two ways to use VISTA: submitting own sequences or examining pre-computed whole-genome alignments. A sidebar on the right lists alignment types: Pairwise sequence alignment, Multiple sequence alignment, Local alignment, and Global alignment.

Fig 3: Link for VISTA tool

This screenshot shows the VISTA homepage. The header features the "VISTA" logo and the tagline "Tools for Comparative Genomics". Navigation links include "VISTA Home", "Custom Alignment", "Browser", "Enhancer DB", "Downloads", "Publications", and "Help". Social media links for "About Us", "Cite Us", and "Contact Us" are also present. The main content area contains a brief description of VISTA's purpose: "VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species." Below this, there are four main sections: "Submit Your Sequences" (with "mVISTA" and "gVISTA" options), "Precomputed Alignments" (with "VISTA Browser" and "VISTA-Point" options), "Enhancer DB" (with "Enhancer DB" and "JGI Genome Portal" options), and "Other Projects" (with "Phylo-VISTA", "TreeQ-Vista", and "PGA" options). A "Updates" sidebar on the right tracks recent changes, such as updates to the Sorghum, Monkey flower, Moss, Maize, Medicago, Sainfoingrass, and Soybean assemblies, and the addition of new plants like C. grandiflora, Drunmond's rockcress, Turnip mustard A. halleri, and Hair grass panicograss. It also mentions 180 new whole-genome plant alignments added to the VISTA Browser. Other updates from August 2013 include updates to C. elegans and C. briggsae assemblies, and the addition of new worms like C. brenneri, C. remanei, C. japonica, C. so. 11, and C. angelaria. A link to the "Vista News Archive" is also provided. The footer includes the "JOINT GENOME INSTITUTE" logo and the "U.S. DEPARTMENT OF ENERGY" logo, along with the copyright notice "© 1997-2013 The Regents of the University of California".

Fig 4: Homepage of VISTA tool

VISTA Tools for Comparative Genomics

[About Us](#) [Cite Us](#) [Contact Us](#)

[VISTA Home](#) [Custom Alignment](#) [Browser](#) [Enhancer DB](#) [Downloads](#) [Publications](#) [Training](#) [Help](#)

[miVISTA](#) [gVISTA](#) [wgVISTA](#) [Phylo-VISTA](#)

gVISTA

- [Submit](#)
- [About gVISTA](#)
- [Cite](#)

gVISTA Submit:

Query Sequence (choose one of the three options)

Sequence:
NM_001354759.2 Homo sapiens adducin 1 (ADD1), transcript variant 11, mRNA
AGCGCCGCAAGCCGACCCCAGGTCTGGCGGGTGGGGCGAGCGGAGGGCTGAGGGCGGA
GAAGGCCCTGGC
GGGGCGCTGCTGCAGGGCCAGGGGACGGGGCGGAGCCGGAGCCGAGCGACGGCGGTG
GCCGCACTGG

FASTA File: No file chosen

OR

GENBANK Identifier:

Treat lower-case letters as repeats

Target Sequence

Base Genome: Human Mar. 2006

Submit

Advanced Options

Your email address: Enter your address to get a notification of the results via email

Name of request: Enter something to identify the data set

Fig 5: Paste the FASTA sequence on gVISTA tool

You are browsing Human Mar. 2006

aligned with:
sequence1
using the **AVID** alignment program

Chromosome 4

Total Groups: 2 (sorted by alignment size)

chr4:2,865,510-2,870,961 (5451bp)	VISTA-Point	VISTA Browser
chr4:2,897,531-2,901,587 (4056bp)	VISTA-Point	VISTA Browser

Fig 6: Result page for query ALPK1

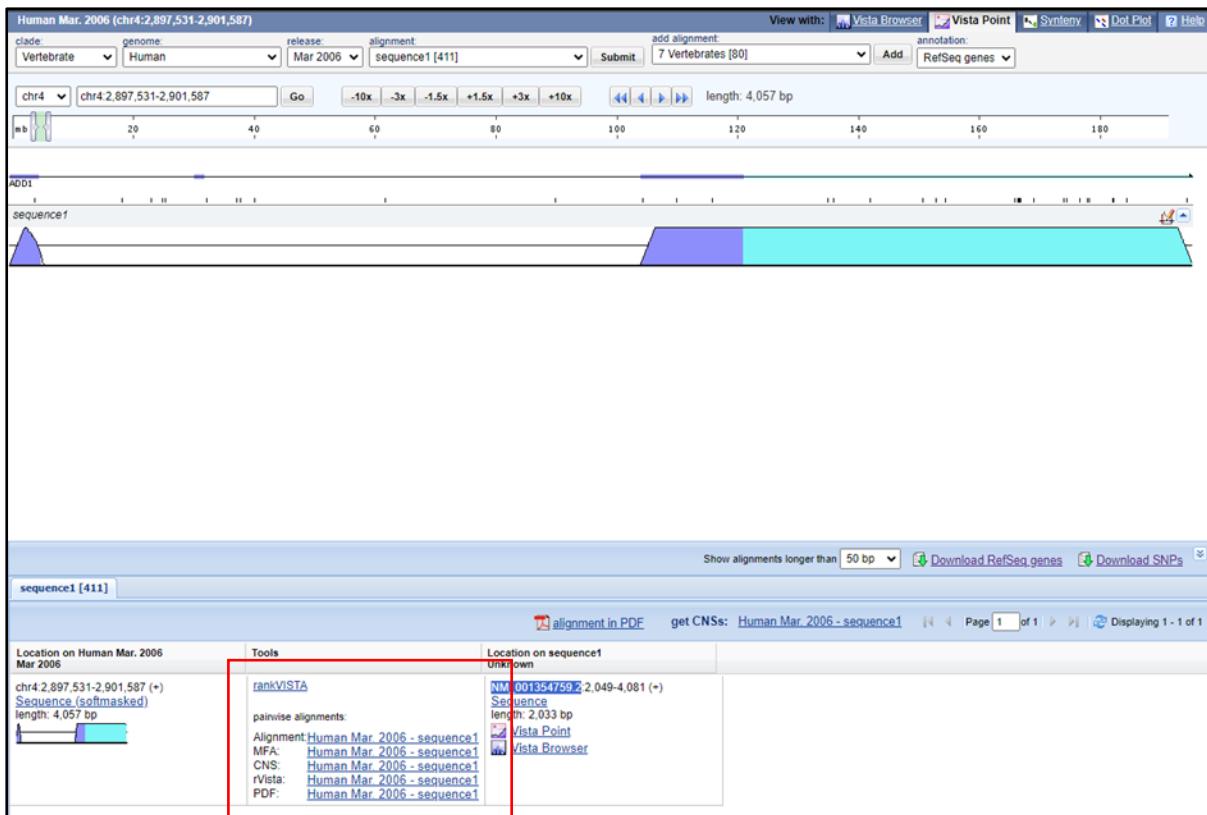


Fig 7: VISTA Point

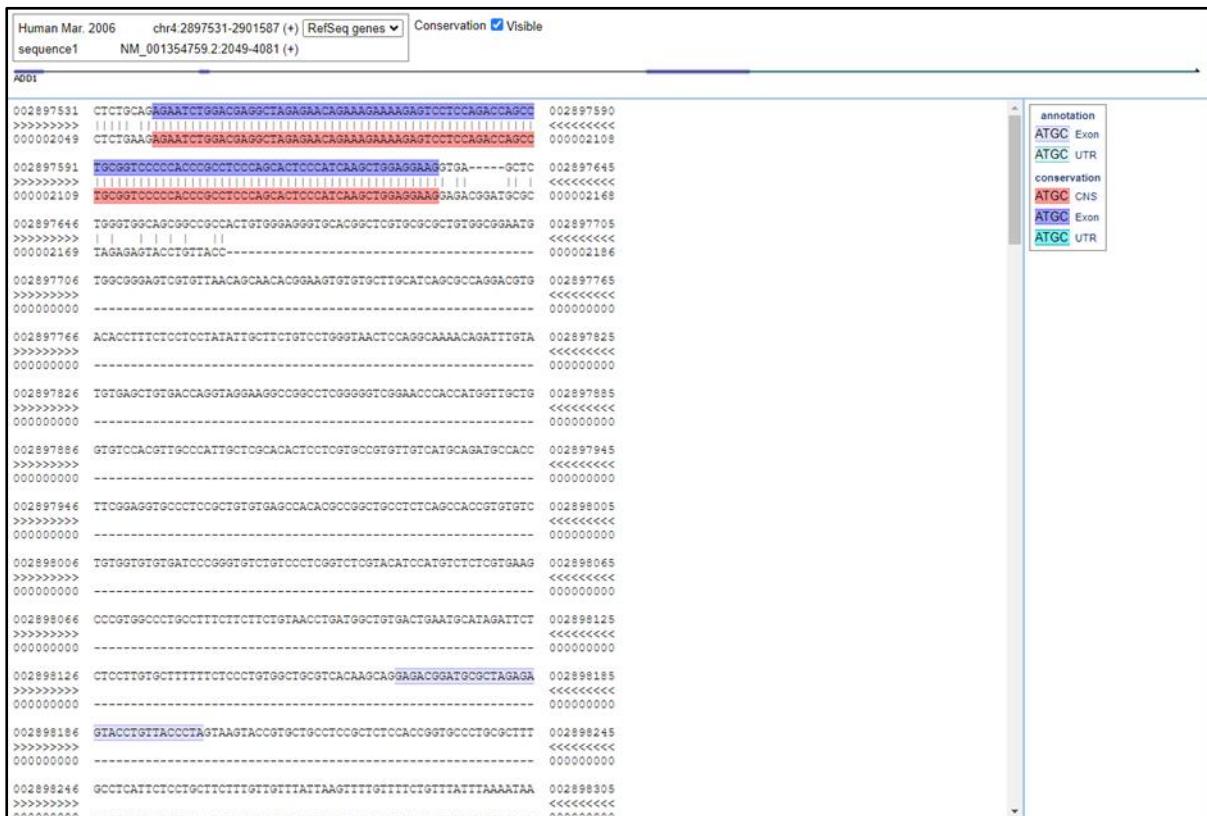


Fig 8: Sequence Alignment for query ADD1

Fig 9: MFA (Multiple Feature Alignment)

***** Conserved Regions - Human Mar. 2006 chr4 (sequence1 NM_001354759.2) *****

2897539	(2057)	to	2897637	(2155)	=	<u>99bp</u>	at 100.0%	exon
2899696	(2190)	to	2900048	(2542)	=	<u>353bp</u>	at 100.0%	exon
2900049	(2543)	to	2901587	(4081)	=	<u>1539bp</u>	at 100.0%	UTR

Fig 10: CNS (Conserved Regions of Sequence)

CNS retrieval options:

Remove gaps Add extra bases upstream (5') and downstream (3') of CNS

```
>Human Mar. 2006 chr4:2899696-2900048 (+)
ACCTTGCGGGAGCCGACTACTGGAGATGACAGTGTGCTGCCACCTTAAGCCAACTC
TCCCCGATCTGCCCCCTGATGAACCTTCAAGGGACTCGGCTTCCCAATGTTAGAGAAGG
AGGAGGAAGCCCATAGACCCCCAAGCCCCACTGAGGCCCCCTACTGAGGCCAGCCCCGAGC
CAGCCCCAGCACCGAGCCCCGGTGGCTGAAGAGGCTGCCCCCTCAGCTGTGAGGGAGGGGG
CCGCCGCGGACCTGGCAGCGATGGGCTCTCAGGCAAGTCCCCGTCAAAAAGAAGAAGAAGA
AGTCCGTACCCGGTCTTCTGAAGAAGGCAAGAAGAAGAGTGAECTCTGA
>sequence1 NM_001354759.2 :2190-2542 (+)
ACCTTGCGGGAGCCGACTACTGGAGATGACAGTGTGCTGCCACCTTAAGCCAACTC
TCCCCGATCTGCCCCCTGATGAACCTTCAAGGGACTCGGCTTCCCAATGTTAGAGAAGG
AGGAGGAAGCCCATAGACCCCCAAGCCCCACTGAGGCCCCCTACTGAGGCCAGCCCCGAGC
CAGCCCCAGCACCGAGCCCCGGTGGCTGAAGAGGCTGCCCCCTCAGCTGTGAGGGAGGGGG
CCGCCGCGGACCTGGCAGCGATGGGCTCTCAGGCAAGTCCCCGTCAAAAAGAAGAAGAAGA
AGTCCGTACCCGGTCTTCTGAAGAAGGCAAGAAGAAGAGTGAECTCTGA
= length = 353bp, identity = 100.0%, type = exon
```

Fig 11: An entry of length 353bp with 100% identity of CNS



Fig 12: PDF format of result page

RESULTS:

The VISTA tool was used for the query ‘ADD1’ (Accession ID: NM_001354759.2) to interpret results by analyzing alignment plots, which highlighted regions of conservation or divergence among sequences. Areas of high similarity were identified by long stretches of aligned sequences or conserved regions, while areas of divergence indicated potential functional differences or evolutionary changes. Metrics or scores provided quantified the similarity or conservation levels between sequences, aiding in understanding evolutionary relationships, identifying functional elements within genomes, or studying genetic variation within populations.

CONCLUSION:

VISTA tool was used for the query ‘ADD1’ (Accession ID: NM_001354759.2) for the comparison of sequences against whole-genome assemblies. This analysis helps in understanding the similarities and differences between the sequences and provides valuable insights into their functional significance.

REFERENCES:

1. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004, July 1). VISTA: computational tools for comparative genomics. *Nucleic Acids Research*, 32(Web Server), W273–W279. <https://doi.org/10.1093/nar/gkh458>
 2. VISTA tools. (n.d.). <https://genome.lbl.gov/vista/index.shtml>
 6. Han, S. H., Oh, H. S., Lee, J. B., Jwa, E. S., Kang, Y. J., Kim, S. G., Yang, S. N., Kim, Y. K., Cho, I. C., Cho, W. M., Ko, M. S., & Baek, K. S. (2015, January 30). Effects of Genetic Polymorphisms of ADD1 Gene on Economic Traits in Hanwoo and Jeju Black Cattle-derived Commercial Populations in Jeju-do. *Journal of Life Science*, 25(1), 21–28. <https://doi.org/10.5352/jls.2015.25.1.21>
 7. Huang, Y. Z., Qian, L. N., Wang, J., Zhang, C. L., Fang, X. T., Lei, C. Z., Lan, X. Y., Ma, Y., Bai, Y. Y., Lin, F. P., & Chen, H. (2018, March 12). Genetic Variants in ADD1 Gene and their Associations with Growth Traits in Cattle. *Animal Biotechnology*, 30(1), 7–12. <https://doi.org/10.1080/10495398.2017.1398754>
 8. Gupta, S., Jhawat, V., Agarwal, B. K., Roy, P., & Saini, V. (2019, May 29). Alpha Adducin (ADD1) Gene Polymorphism and New Onset of Diabetes Under the Influence of Selective Antihypertensive Therapy in Essential Hypertension. *Current Hypertension Reviews*, 15(2), 123–134. <https://doi.org/10.2174/1573402114666180731111453>
-

DATE: 20/03/2024

WEBLEM 8

INTRODUCTION TO IDENTIFICATION OF REPETATIVE ELEMENTS: REPEATMASKER TOOL

(URL: <http://galaxy.org/>)

INTRODUCTION:

“Repeats” represent a mosaic of nucleotide patterns that have intrigued biologists for decades. Often seemingly mundane in their nature, “Repeats” also known as “Repetitive elements” hold profound significance in shaping and offering insights into genetic architecture, genomic stability, and evolutionary relationships across all forms of life.

The analysis of genetic diversity and relatedness within and between the different species and populations has been a major theme of research for many biologists. With the availability of whole-genome sequencing for an increasing number of species, focus has been shifted to the development of molecular markers based on DNA or protein polymorphism. DNA sequences originate and undergo evolutionary metamorphoses’ and thus may be used as powerful genetic markers to characterize genomes of wide range of species.

The mammals have approximately 3 billion base pairs per haploid genome harbouring about 20,000-25000 genes. A minor part of the genome (5-10%) is coding sequences and the remaining part is non-coding (Heterochromatin) representing repetitive DNA. Comparison of the genome size of different eukaryotes shows that the amount of non-coding DNA is highly variable and constitutes 30% to about 99% of the total genome.

Repetitive sequences are dynamic elements exhibiting a high degree of polymorphism due to variation in the number of their repeat units caused by mutations involving several mechanisms. They reshape their host’s genome by generating rearrangements, shuffling of genes and modulating pattern of expression. This dynamism of repeats leads to evolutionary divergence that can be used in species identification, phylogenetic inference and for studying process of sporadic mutations and natural selection. These repetitive sequences are mainly composed of “Interspersed” and “Tandem Repeats”. However, the majority of repetitive DNA sequences have not been sequenced and/or are not identifiable by currently applied methods.

TYPES OF REPETITIVE ELEMENTS:

Tandem Repeats are small sequences (<60 base pairs) arranged one after the other in the genome in a head-to-tail organization. Tandemly repeated DNA is common in eukaryotic genomes, in some cases in short sequences 1–10 bp long and in other cases associated with genes and in much longer sequences. Tandem repeats may be further classified according to the length and copy number of the basic repeat units as well as its genomic localization. The greatest amount of tandemly repeated DNA is associated with centromeres and telomeres.

Interspersed Repeats or Dispersed Repeats consist of families of repeated sequences interspersed through the genome with unique-sequence DNA. Each family consists of a set of related sequences characteristic of the family. Often, small numbers of families have very high copy numbers and make up most of the dispersed repeated sequences in the genome. By the mechanism of their transposition, interspersed repeats are classified into two classes:

- RNA transposons:** RNA transposons also known as retroelements found in eukaryotic genome require reverse transcription for their activity. Based on their structural relationship, RNA transposons are divided into two general categories: LTR elements and non-LTR elements. The non-LTR elements consist of Long Interspersed Elements (LINEs), in which the sequences in the families are about 1,000–7,000 bp long and Short Interspersed Elements (SINEs), in which the sequences in the families are 100–400 bp long.
- DNA Transposons:** DNA transposons do not require RNA intermediate and transpose in a direct DNA-to-DNA manner.

All eukaryotic organisms have LINEs and SINEs, with a wide variation in their relative proportions.

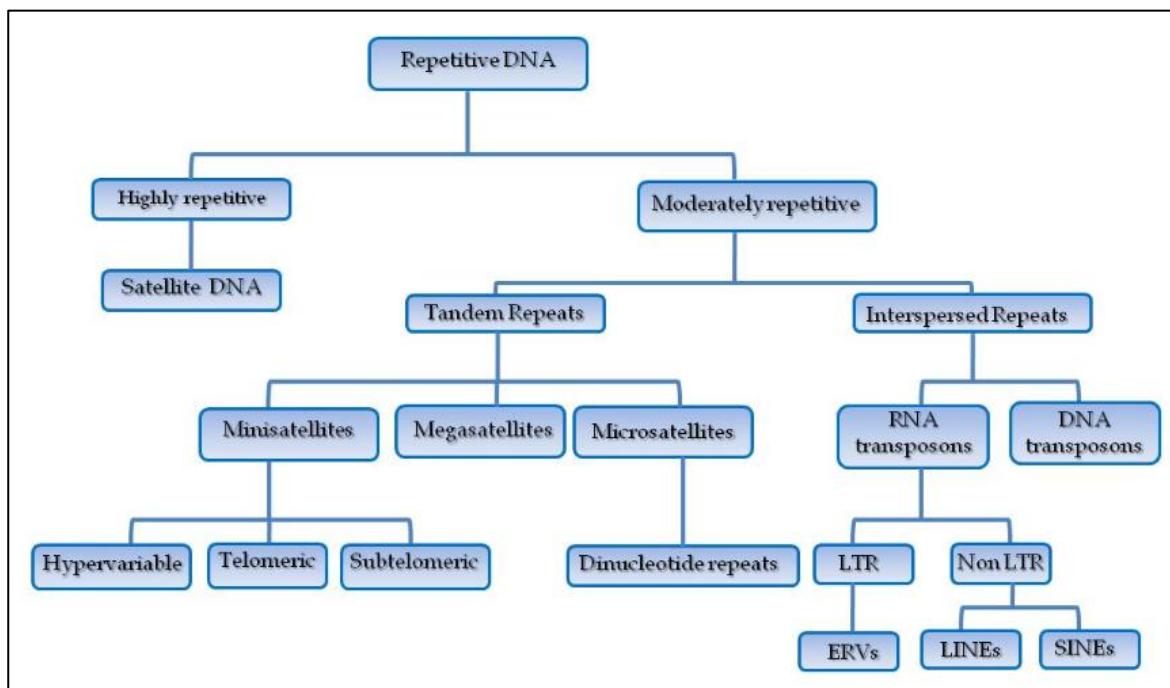


Fig 1: Schematic diagram showing biological categories of the different repetitive sequences

Repeats, however important, the source of a lot of trouble when you work on genomics data. The widespread occurrence of repetitive stretches of DNA in genomes of imposes fundamental challenges for sequencing, genome assembly, and automated annotation of genes and proteins. This multi-level problem can lead to errors in genome and protein databases that are often not recognized or acknowledged. Repetitive elements complicate the accurate determination of gene boundaries, regulatory regions, and structural variations. In sequencing efforts, repetitive sequences can confound read alignment algorithms, leading to misassembly or fragmentation of genomic contigs. This, in turn, undermines the reliability of downstream genomic analyses and annotations. To address these challenges, tools like RepeatMasker are employed.

REPEATMASKER TOOL:

RepeatMasker is a program that screens DNA sequences and detects transposable elements, satellites, and low-complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version

of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). Currently over 56% of human genomic sequence is identified and masked by the program.

There are two types of masking, producing slightly different FASTA output:

1. **Soft-masking:** Repeat elements are written in lower case, e.g. ATATA = atata.
2. **Hard-masking:** Repeat elements are replaced by stretches of the letter N

The non-repeated sequences are always in uppercase. Hard-masking is destructive as large parts of the sequence which are replaced by stretches of N, therefore, soft masking is preferred as it allows to perform annotation.

The tool can be accessed from Galaxy Server's tool repository among the available options. Galaxy is an open source, web-based platform for data intensive biomedical research. It is designed to facilitate accessible, reproducible, and collaborative computational analysis in the field of bioinformatics. It provides a user-friendly interface for researchers, enabling them to perform a wide range of genomic analyses without requiring advanced computational expertise or software installation. The development and maintenance of the site are supported by NIH NHGRI award U24 HG006620.

Repeat identification and masking is usually a previous step to the gene prediction and annotation phase. The masking step signals to downstream sequence alignment and gene prediction tools that these regions are repeats. Identifying repeats is complicated by the fact that repeats are often poorly conserved; thus, accurate repeat detection usually requires a repeat library for the species of interest. RepeatMasker relies on existing databases of repeated elements signatures like Dfam and RepBase.

The Dfam database is an open collection of DNA Transposable Element sequence alignments, Hidden Markov Models (HMMs), consensus sequences, and genome annotations. It represents a collection of multiple sequence alignments, each containing a set of representative members of a specific transposable element family. These alignments (seed alignments) are used to generate HMMs and consensus sequences for each family. While RepBase is a database of representative repetitive sequences from eukaryotic species.

Sequences can be pasted in or uploaded as files, both in FASTA format. The program returns four or five output files for each query:

1. **Masked Sequence:** This is the FASTA file that is used for future analysis. When displayed it is observed that some portions of the sequence are in lowercase or series of Ns.
2. **Repeat Statistics:** This one contains some statistics on the number of repeats found in each category, and the total number of base pairs masked.
3. **Output Log:** This is a tabular file listing all repeats.
4. **Repeat Catalogue:** This one contains the list of all repeat sequences that were identified, with their position, and their similarity with known repeats from the Dfam database.
5. **Repeat Annotation:** This one contains the coordinate of each repeat element in GFF2 format.

REFERENCES:

1. Dfam. (n.d.). <https://www.dfam.org/home>
 2. Galaxy Training Network. (2024, January 8). Genome Annotation / Hands-on: Masking repeats with RepeatMasker. <https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/repeatmasker/tutorial.html>
 3. Liehr, T. (2021). Repetitive elements in humans. International Journal of Molecular Sciences, 22(4), 2072. <https://doi.org/10.3390/ijms22042072>
 4. Pathak, D., & Ali, S. (2012). Repetitive DNA: a tool to explore Animal Genomes/Transcriptomes. In InTech eBooks. <https://doi.org/10.5772/48259>
 5. Repeat masking. (n.d.). <https://manual.omicsbox.biobam.com/user-manual/omicsbox-modules/module-genome-analysis/repeat-masking/>
 6. RepeatMasker home page. (n.d.). <https://www.repeatmasker.org/>
 7. Russell, P. (2005). IGenetics: A Molecular Approach. <http://ci.nii.ac.jp/ncid/BB00859532>
 8. Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Research, 47(21), 10994–11006. <https://doi.org/10.1093/nar/gkz841>
-

DATE: 20/08/2024

WEBLEM 8(A)
REPEATMASKER TOOL
(URL: <http://galaxy.org/>)

AIM:

To mask repeats using RepeatMasker tool on dataset imported from Zenodo.

INTRODUCTION:

Repetitive sequences, refer to segments of DNA that occur multiple times within a genome. These sequences can range in length from a few base pairs to thousands of base pairs and can be categorized into different types based on their structure, function, and evolutionary origin. Repeats are ubiquitous across all domains of life and play diverse roles in genome organization, stability, and evolution.

They can contribute to genome plasticity and adaptation by promoting genetic diversity through recombination, mutation, and gene duplication. However, repeats can also pose challenges for genome assembly, annotation, and stability, as they can lead to errors in sequencing and alignment and can contribute to genomic instability and disease. Therefore, tools like RepeatMasker are used before downstream analysis of sequences and gene prediction.

RepeatMasker is a popular software tool widely used in computational genomics to identify, classify, and mask repetitive elements, including low-complexity sequences and interspersed repeats. The tool utilizes the Dfam database as a reference library for identifying known repetitive element families, improving the accuracy of repeat annotation in genomic analyses. RepeatMasker can be accessed from the toolkit of Galaxy Server.

Dfam is a comprehensive database of repetitive DNA elements, specifically focusing on transposable elements (TEs) and other repetitive sequences that proliferate within genomes. Created and maintained by the Gardner Lab at the University of California, Santa Cruz, Dfam serves as a valuable resource for researchers studying the structure, function, and evolution of repetitive elements in genomes.

The following weblem is based on a tutorial given by Trainings in Galaxy Project. The data is imported from Zenodo (link: https://zenodo.org/record/7085837/files/genome_raw.fasta).

Zenodo is an open-access digital repository and platform designed to facilitate the sharing, preservation, and citation of research outputs across all fields of science. Developed and maintained by CERN (the European Organization for Nuclear Research), Zenodo serves as a centralized repository where researchers, scientists, and scholars can deposit and archive a wide range of scholarly outputs, including datasets, software, papers, posters, presentations, and other research artifacts.

METHODOLOGY:

1. Open Galaxy Server Webpage (<http://galaxy.org/>) and log into your account.
2. Search for RepeatMasker Tool in the Galaxy Toolkit
3. Select “masked sequence on data 1”.

4. Go to Galaxy Trainings and click on Genome annotation, under that click on Repeat Masking with RepeatMasker.
5. Copy the Zenodo link for the dataset and import the files on RepeatMasker.
6. Using the genome sequence run the tool.
7. Six files are generated, interpret the results.

OBSERVATIONS:

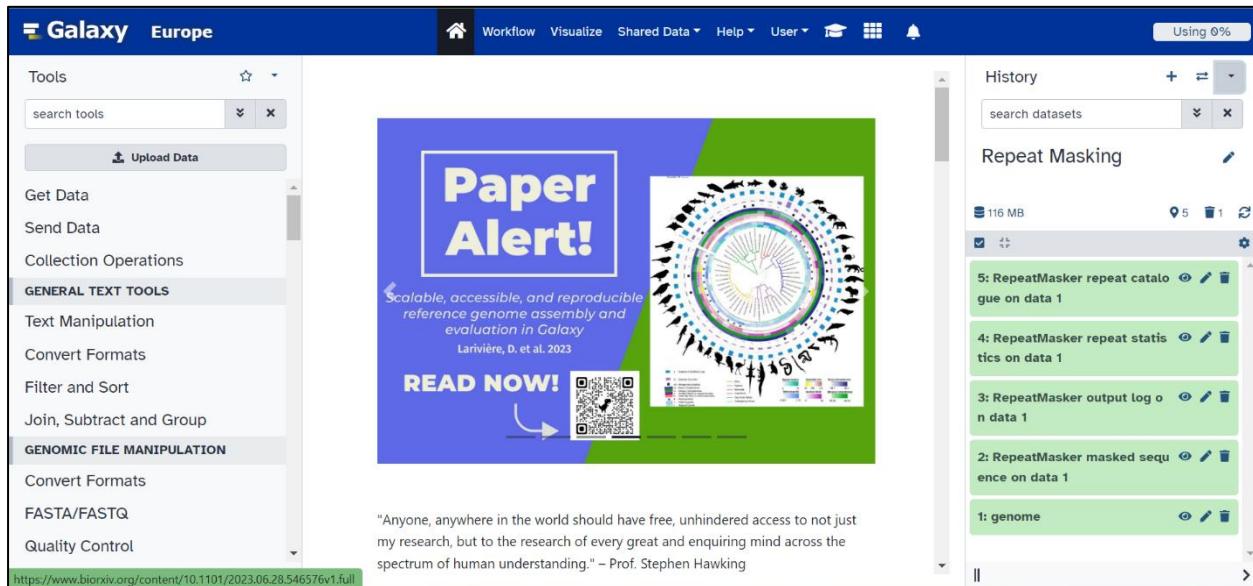


Fig 1: Homepage of Galaxy Server

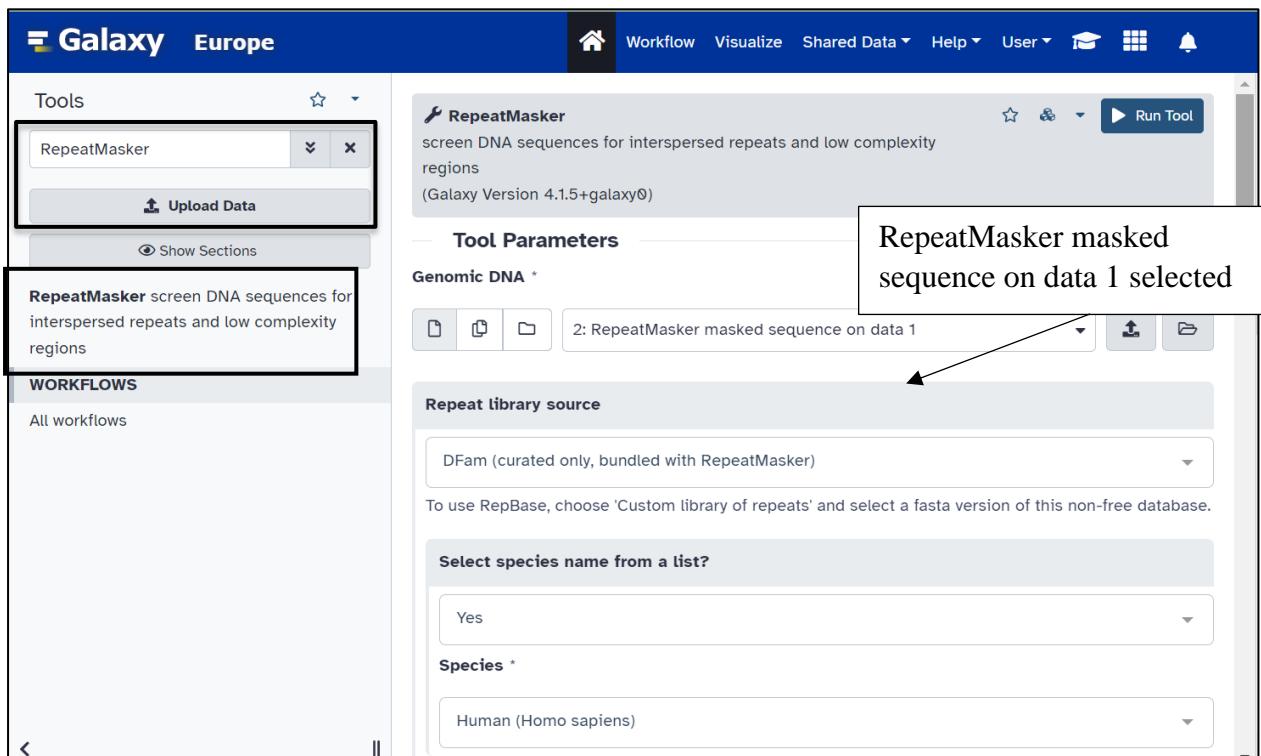


Fig 2: RepeatMasker Searched in Galaxy Toolkit

Galaxy Europe

Tools

RepeatMasker

Upload Data

Show Sections

RepeatMasker screen DNA sequences for interspersed repeats and low complexity regions

WORKFLOWS

All workflows

RepeatMasker screen DNA sequences for interspersed repeats and low complexity regions

Output annotation of repeats in GFF format

No
(-gff)

Yes
Scaffolds are sometimes joined with stretches of 25 or more Ns. This option ignores them when calculating repeat statistics (-excln)

Perform softmasking instead of hardmasking
Output repetitive regions as lowercase, non-repetitive regions as uppercase (instead of replacing repetitive regions with 'N's) (-xsmall)

Advanced options

Additional Options

Email notification

Fig 3: Default Criteria before Running tool

Galaxy

Tools

repeat

Upload Data

Show Sections

RepeatMasker screen DNA sequences for interspersed repeats and low complexity regions

RepeatModeler Model repetitive DNA

equicktandem Finds tandem repeats

tandem Looks for tandem repeats in a nucleotide sequence

einverted Finds DNA inverted repeats

palindrome Looks for inverted repeats in a nucleotide sequence

Red repeat masking

PRESTO FilterSeq Filters and/or masks reads based on length, quality, missing bases and repeats.

WORKFLOWS

All workflows

RepeatMasker screen DNA sequences for interspersed repeats and low complexity regions
(Galaxy Version 4.1.5+galaxy0)

Ignore stretches of Ns when computing statistics
Scaffolds are sometimes joined with stretches of 25 or more Ns. This option ignores them when calculating repeat statistics (-excln)

Perform softmasking instead of hardmasking
Output repetitive regions as lowercase, non-repetitive regions as uppercase (instead of replacing repetitive regions with 'N's) (-xsmall)

Advanced options

Additional Options

Email notification

Yes
Send an email notification when the job completes.

Run Tool

Help

RepeatMasker is a program that screens DNA for interspersed repeats and low complexity DNA sequences. The database of repeats to screen for can be provided as a FASTA file or downloaded from RepBase. If the RepBase option is chosen the RepBaseRepeatMaskerEdition file should be downloaded and unpacked, and the enclosed EMBL format file ('RMRBSeqs.embl') should be uploaded to Galaxy for use with this tool.

Further documentation is available on the RepeatMasker homepage.

Fig 4: E-mail Notification Option Turned On

 Galaxy Training!

Contributors Learning Pathways Help ▾ Settings ▾ Search Tutorials

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

We have separated the tutorials into several categories based on field and technology. We are exploring other ways to organise the tutorials going forward!

Introduction

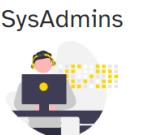
Topic	Tutorials
Introduction to Galaxy Analyses	12

Quickstart



Learning Pathways





Galaxy for SysAdmins



Fig 5: Galaxy Training Page Opened

Foundations of Data Science	48
Ecology	20
Evolution	2
FAIR Data, Workflows, and Research	14
Genome Annotation	19
Imaging	7
Materials Science	1
Microbiome	15
One Health	9
Statistics and machine learning	17

Feb 29, 2024
↗ GTN ❤️ GMOD

Jan 31, 2024
↗ 🚧 Hosting Galaxy at the Edge: Directly in Your Pocket!

Jan 30, 2024
↗ GTN in Discourse

[See all news](#)

GTN Toots

3/18/2024, 7:24:10 PM

 **Galaxy Training Network**
@gtn

New GTN post: Cool URLs Don't Change, GTN

Fig 6: Click on Genome Annotation

The screenshot shows the Galaxy Training! interface. At the top, there is a navigation bar with links for 'Learning Pathways', 'Help', 'Settings', and a search bar labeled 'Search Tutorials'. Below the navigation bar, the main content area has a title 'Genome Annotation'. A descriptive text follows, stating: 'Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.' Below this text is a note: 'You can view the tutorial materials in different languages by clicking the dropdown icon next to the slides (畏惧) and tutorial (畏惧) buttons below.' Under the heading 'Requirements', there is a list: '• [Introduction to Galaxy Analyses](#)'.

Fig 7: Genome Annotation Page Opened

The screenshot shows the 'Eukaryotes' learning pathway page. The title 'Eukaryotes' is at the top, followed by a subtitle 'Annotation of eukaryotic genomes.' Below this is a table with columns: 'Lesson', 'Slides', 'Hands-on', 'Recordings', 'Input dataset', and 'Workflows'. There are four rows of lessons:

Lesson	Slides	Hands-on	Recordings	Input dataset	Workflows
Masking repeats with RepeatMasker graduation cap icons, eukaryote	Slides icon with dropdown arrow	Hands-on icon with dropdown arrow	Recordings icon with dropdown arrow	Input dataset icon with dropdown arrow	Workflows icon with dropdown arrow
Genome annotation with Funannotate graduation cap icons, gmod, eukaryote, jbrowse1	Slides icon with dropdown arrow	Hands-on icon with dropdown arrow	Recordings icon with dropdown arrow	Input dataset icon with dropdown arrow	Workflows icon with dropdown arrow
Genome annotation with Maker (short) graduation cap icons, gmod, eukaryote, maker, jbrowse1	Slides icon with dropdown arrow	Hands-on icon with dropdown arrow	Recordings icon with dropdown arrow	Input dataset icon with dropdown arrow	Workflows icon with dropdown arrow
Genome annotation with Maker graduation cap icons, gmod, eukaryote, jbrowse1, maker	Slides icon with dropdown arrow	Hands-on icon with dropdown arrow	Recordings icon with dropdown arrow	Input dataset icon with dropdown arrow	Workflows icon with dropdown arrow

Fig 8: Click on “Masking Repeats with RepeatMasker” tutorial

The screenshot shows the Galaxy Training! website interface. At the top, there is a navigation bar with links for "Galaxy Training!", "Genome Annotation", "Learning Pathways", "Help", "Settings", and "Search Tutorials". Below the navigation bar, the title of the tutorial is displayed: "Masking repeats with RepeatMasker". Underneath the title, there is a section for "Author(s)" featuring six individuals with their names and profile pictures. Below this, there is a "Overview" section, a "Questions" section with two items, and a "Objectives" section with one item. A Creative Commons BY license logo is also present.

Fig 9: “Masking Repeats with RepeatMasker” tutorial opened

The screenshot shows the "Get data" section of the tutorial. On the left, there is a sidebar with links for "Get data", "Soft-masking using Red", "Soft-masking using RepeatMasker", "Conclusion", "Frequently Asked Questions", "References", "Feedback", and "Citing this Tutorial". The main content area is titled "Get data" and contains a "Hands-on: Data upload" section. It provides instructions for creating a new history and importing files from Zenodo or a shared data library. A list of three URLs is shown, each with a "Copy" button. Below the URLs, there are two tips: "Tip: Importing via links" and "Tip: Importing data from a data library".

Fig 10: Copy the first link to import dataset

The screenshot shows the Galaxy Europe web interface. At the top, there's a navigation bar with the Galaxy logo, 'Galaxy Europe', and links for 'Workflow', 'Visualize', and 'Share'. On the left, under the 'Tools' section, 'RepeatMasker' is selected. Below it are buttons for 'Upload Data' and 'Download from URL or upload files from disk'. A callout points to the 'Upload Data' button. To the right, the 'RepeatMasker' tool details are shown, including its purpose ('screen DNA sequences for interspersed repeats and low complexity regions'), version ('Galaxy Version 4.1.5+galaxy0'), and parameters like 'Genomic DNA *' and 'Repeat library source' (set to 'DFam').

Fig 11: Click on Upload Data

This screenshot shows the 'Upload from Disk or Web' dialog box. At the top, it says 'Upload from Disk or Web'. Below that, there are tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based', with 'Regular' selected. A message indicates 'You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.' The main area shows a table with columns: Name, Size, Type, Genome, Settings, and Status. One row is listed: 'Name' is 'New File', 'Size' is '56 b', 'Type' is 'Auto-det...', 'Genome' is 'unspecified (?)', 'Settings' has a gear icon, and 'Status' is '0%'. Below the table, a note says 'Download data from the web by entering URLs (one per line) or directly paste content.' A text input field contains the URL 'https://zenodo.org/record/7085837/files/genome_raw.fasta'. At the bottom, there are filters for 'Type (set all):' (Auto-detect), 'Q', 'Genome (set all):' (unspecified (?)), and buttons for 'Choose local files', 'Choose remote files', 'Paste/Fetch data', 'Start', 'Pause', 'Reset', and 'Close'.

Fig 12: Paste the link and click “Start”

Upload from Disk or Web

Regular Composite Collection Rule-based

Name	Size	Type	Genome	Settings	Status
New File	56 b	Auto-det...	unspecified (?)	⚙️	100% ✓

Download data from the web by entering URLs (one per line) or directly paste content.

https://zenodo.org/record/7085837/files/genome_raw.fasta

Type (set all): Auto-detect Genome (set all): unspecified (?)

Choose local files Choose remote files Paste/Fetch data Start Pause Reset Close

The screenshot shows a user interface for uploading files. At the top, there are tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. Below this is a table with columns for Name, Size, Type, Genome, Settings, and Status. A single row is present, labeled 'New File' with a size of '56 b', type 'Auto-det...', genome 'unspecified (?)', settings icon, and a status bar showing '100%' with a checkmark. Below the table is a green box containing a URL: 'https://zenodo.org/record/7085837/files/genome_raw.fasta'. At the bottom, there are filters for 'Type (set all)' and 'Genome (set all)', and several buttons: 'Choose local files', 'Choose remote files', 'Paste/Fetch data', 'Start', 'Pause', 'Reset', and 'Close'.

Fig 13: New File successfully loaded

The screenshot shows a list of generated files. At the top, it says '166 MB' and has icons for location (6), trash (1), and refresh. Below is a list of files:

- 7: genome_raw.fasta
- 5: RepeatMasker repeat catalog on data 1
- 4: RepeatMasker repeat statistics on data 1
- 3: RepeatMasker output log on data 1
- 2: RepeatMasker masked sequence on data 1
- 1: genome

Each item has three icons: eye, edit, and trash.

Fig 14: Six Files are generated

Tool Parameters

Genomic DNA *

1: genome

Repeat library source

DFam (curated only, bundled with RepeatMasker)

To use RepBase, choose 'Custom library of repeats' and select a fasta version of this non-free database.

Select species name from a list?

Yes

Species *

Human (Homo sapiens)

Fig 15: Import the Genome file

Perform softmasking instead of hardmasking

No
Output repetitive regions as lowercase, non-repetitive regions as uppercase (instead of replacing repetitive regions with 'N's) (-xsmall)

Advanced options

Additional Options

Email notification

Yes
Send an email notification when the job completes.

Run Tool

Help

RepeatMasker is a program that screens DNA for interspersed repeats and low complexity DNA

Fig 16: Turn On e-mail notification and Run Tool

>contig_1000
GTCTTGACAGTCCCCCAAGATAAAAGTATCATTTCACAACCGCGATACATCACAG
ATGTATGATTAATATACTTGAAGGATATACCGACATTCAAATTTAAAATTAGTGAG
ATGTATATGGAATCCATAAGGAACCTTACTAAATGTGTTCATACAATTTCACAAAAA
AAAATAAAAAAATCCATTTCATTTACTCAATAATCGATGGGTATCCTTACTCGG
ATCACATCCTTCATTGTTGGCATTAATCGATTAACCTTAAATCTTCACGCC
ATTAAGTCCATTAGATTCTACGAGAACCTCAATGATTCTATGTTGAACATACTA
AACACGTCTTCTTGTCTGTTCAAGATCATCCAAGATTCTGATTGGTTCAAATCT
GGTGATTGAGGTGGCCAATTCAATCGAAGAATTGGTATTGCTTCCATTACAGAGCA
ACTTTGCAGTATGGATTGGAGCATTATCTCCATCAATACTGGTGTGCTGAGCTTGTG
AAAAAAATCCAACAAACACTGGTTCATAAACCTGCTTATGAAGTCGTTCTAGATCCCTGG
CCTTGTGAAAAAAACCAAACGGTGCAGCATTAGGTTATTCTCCCCAAATAGAC
AACGACGACCTCCACTTTAAATGGTGGCTAAACAACTAGGAAGCATTGGTCTTCTT
GGGCCTCGTGAATTCTAATAGGACAAGAAGAAGTCAAGTTGGAGCTGCATTCT
GTCCAGATTACTCGACGCCAGTCTTCCAAAGTCAACTTCGTTGGCCCTGGCAAACCTGA
AGACGTCTTCTCTGTTCTGTTCAACAAAGGCTTTCTCGGATACGGCTCTC
AAGTTGAGGGCGGTGAAGCTTTCTAAAGGTTCTTGTAGCATTAACGCCACATTA
GCCATGATCGGCAAGAGTACTCGTCTATGAGTAATCACATAGTATCGTAAGGCTCGT
CTGTCCTTCAAGATCGTACGTCACCAGTGCCTTTCTCTCAAATGAACCT
GATTCCTCAACGTTCAATCGCACGCTTAATTGAGCATGGTAAACTCAGTTCGCGA
GCAATTGGCAATAGTACGCCATTCTTCTCGGACTGCATAGGCTTTGAGTG
GGACTTAATTGAGGCTTGAAGATCTTATTAAATGAGAAACTAACGAAATA
AAGCTTCTTATTGCGGGGGCTTGAAGAAAATGACGCCAGTGCACAAATGATACT
TTATCTGGAGGGGACTGTACATCCTTGTCAATATGCCCTGATCGGACCTGGT

Fig 17: Genome File

>contig_1001
AGTTGTCAGCAAAAAATCTCAGATCTGATCTGATCTGAAGTG
CGGAGTACCTCTATGGGTTGTATAAAACTCAGTAACTC
AGATGAAAAAATCCATCTGAAGTCTACCGAAGTGATAAAAAGGAATT
TCCCAGTATTCTAATTCAAGATATTCCAATTTTTAAATCC
AAATTCTTCTTCTTCTTCTATAAACATTAACATTACCATGGAGCAACAA
ACATATTGAAACAAAGGAAAGATTACTAAGAACAGTATACCAA
GATGATATGACCATTGCAGAAATTATGGACGCTACTATTGGCTAT
CAATACTCACAAATATTGGCAGAACAGAGTCAAATCAACAACTC
AGCATTGGAAAGCGTGTGACGTTGGAATACAACGAGGCCAACGCTCCT
CCCAACATGAAGGTTGTTAGGTCCGAAAGCTTCTCGAGAAATA
TGGTGTATGAAAACAAAGTATCAGAACAAAAGGAGACATTAAAGTCGA
CCGGTTCGGGAGGTGGCCACCTTCTGGTACCTTCAAAATGCTGAT
TATTTGCTAAAGGACCCCGACTCACCACCCAAAGGACCTATAGCTCGGT
CAAAGAGCTTATAGTGGAGATGACGAGGAGAACAGTGTGATCTG
TCGTCTCGAAGATTCTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GCCTCC
ATAGCGACAAAAAAGAGGAAATCTGCAAGGAGGAAAGGAGCT
AGCTCGAGTGAAGTCTAGAGATGAAGCTTGGAACAGATGAAGCAAATCA
CGGAAGGATCTGACAAGCAGTCAGGAGGATATTAGCAAGGCTCTGTT
GTTCAACACGTTGAATCTGAAGAGAAAAGGAGTTTACAAGAACGTTCT
TGAATTGATGGAAAATGGTTAAAATAATTCTGTTAAATTTTTTA
TTGAAACTAGTCAAGAACGTTAACATCACGTTAAAGATCTCGTC
GTTGTCCTCCAGCAGTTCTCAAGGAGGTGTTCTCAATGTTGCT
GGACGATCACTCTGAAGCAAGTGACAACTCGACTAATTCTCATCAGT
TATACCAAAAGTATCATTGACATTGATGCAAAAGTATGCAAAATACAGC
TGGTATTAAAGATTTGACCAAGCTAATTGGCTTGGATGACACATAC
TTCAAAAGGAATCTCCATCGCTTACAAGGATACCAATGCTTTCAAC
AACCTGGCGATTGCAACTAAACTGGCGTTGAATCTGCTTGGGGAT

Hard-masking

Fig 18: Masked Sequence File

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9
SW score	% div.	% del.	% ins.	query sequence	pos in query: begin	end (left)	repeat	
13	13.7	0.0	0.0	contig_1001	720	744 (2078)	(CCTC)n	
11	7.3	10.3	0.0	contig_1008	2423	2451 (564)	(GATGAC)n	
12	7.5	6.7	3.2	contig_101	1447	1476 (126812)	(TACTTT)n	
15	10.0	0.0	8.6	contig_101	13270	13307 (114981)	A-rich	
37	0.0	0.0	0.0	contig_101	15274	15307 (112981)	(CAT)n	
18	14.9	0.0	0.0	contig_101	18515	18544 (109744)	(T)n	
18	33.6	0.0	0.0	contig_101	18894	18960 (109328)	(CTG)n	
15	24.0	0.0	0.0	contig_101	19841	19879 (108409)	(ACA)n	
37	0.0	0.0	0.0	contig_101	22257	22290 (105998)	(T)n	
29	23.6	2.1	2.1	contig_101	23907	23999 (104289)	(AGAAGT)n	
26	0.0	0.0	0.0	contig_101	24455	24479 (103809)	(CATCCT)n	
25	n/a	n/a	n/a	contig_101	25516	25542 (102746)	(Δ)n	

Fig 19: Sequence Output Log File

```
=====
file name: rm_input.fasta
sequences: 1461
total length: 48645285 bp (48645285 bp excl N/X-runs)
GC level: 36.60 %
bases masked: 1173324 bp ( 2.41 %)
=====
          number of      length      percentage
          elements*    occupied    of sequence
-----
SINEs:           26       1259 bp   0.00 %
    ALUs          0        0 bp    0.00 %
    MIRs          5       265 bp   0.00 %

LINEs:          160      10580 bp  0.02 %
    LINE1         6       321 bp   0.00 %
    LINE2         39      2382 bp  0.00 %
    L3/CR1        59      4081 bp  0.01 %

LTR elements:   15       2216 bp  0.00 %
    ERVL          2       106 bp   0.00 %
    ERVL-MaLRs   0        0 bp   0.00 %
    ERV_classI   0        0 bp   0.00 %
    ERV_classII  0        0 bp   0.00 %

DNA elements:   34       2431 bp  0.00 %
    hAT-Charlie  3       149 bp   0.00 %
    TcMar-Tigger 4       227 bp   0.00 %

Unclassified:   6       395 bp   0.00 %

Total interspersed repeats: 16881 bp  0.03 %
=====
```

Fig 20: Sequence Statistics File

```

13 13.72 0.00 0.00 contig_1001 720 744 (2078) (CCTC)n#Simple_repeat 1 25 (0) m_b1s252i1

contig_1001          720 CCCCCCTCCCTCTTCCCTCCCTCC
                     i      ii
(CCTC)n#Simpl       1 CCTCCCTCCCTCCCTCCCTCCCTCC 25

Matrix = Unknown
Transitions / transversions = 1.00 (3/0)
Gap_init rate = 0.00 (0 / 24), avg. gap size = 0.0 (0 / 0)

11 7.28 10.34 0.00 contig_1008 2423 2451 (564) (GATGAC)n#Simple_repeat 1 32 (0) m_b1s252i0

contig_1008          2423 GATGACGATGAGGATGTC-ATGA-G-TGACGA 2451
                     V      V -   -
(GATGAC)n#Sim       1 GATGACGATGACGATGACGATGACGATGACGA 32

Matrix = Unknown
Transitions / transversions = 0.00 (0/2)
Gap_init rate = 0.11 (3 / 28), avg. gap size = 1.00 (3 / 3)

12 7.52 6.67 3.23 contig_101 1447 1476 (126812) (TACTTT)n#Simple_repeat 1 31 (0) m_b2s252i0

contig_101           1447 TACTTT-CTTTA-GTTTAGTTAACTTT 1476
                     -      -V    V   -
(TACTTT)n#Sim       1 TACTTTACTTTACTTTACTTT-ACTTTT 31

Matrix = Unknown
Transitions / transversions = 0.00 (0/2)

```

Fig 21: Sequence Catalog File

RESULTS:

RepeatMasker tool was used (via Galaxy Server toolkit) to mask repeats in DNA sequences of dataset imported from Zenodo. The tool generated an output of 4 files namely: Masked Sequence file, Output Log File, Statistics file and Catalog file. The masked sequence file contains the input sequence with the identified repetitive elements masked, replacing them with Ns for nucleotides (hard-masking was carried out). The masked sequences are crucial for downstream analyses where repetitive elements might interfere, such as gene prediction or genome assembly. The output log file records the procedural events that occurred during the execution of RepeatMasker. This includes the list of repeats in the contigs, and their positions along with the SW score (Smith-Waterman score, represents a measure of similarity between the sequence being analyzed and the repetitive elements identified in the Dfam database). The statistics file summarizes the process of masking repeats by stating the number of sequences (1461), total number of base pairs (48645285), length occupied by LINEs, SINEs, LTR elements, DNA elements and other unclassified repeats, all giving a total interspersed repeat of 16881bp. The catalog file provides a detailed annotation of the repeats identified in each sequence, including the type of repeat, its location, and its score. The catalog file effectively maps all identified repeats within the sequences.

CONCLUSION:

The RepeatMasker tool is widely used in genomics for identifying and masking repetitive elements in DNA sequences. Repeat masking is followed through before proceeding with any downstream analysis of sequences like gene prediction and genome annotation as a thumb rule; because repetitive sequences can complicate sequence alignments and gene identification.

RepeatMasker tool in the Galaxy server, typically generates four output files, each providing different insights into the analyzed sequences.

The analysis concludes with repeats in the contig sequences masked, a list of repeats observed, and statistics on the repeats, useful for both specific studies and general genomic characterization. The diversity of repetitive elements found and mapped in the catalog file within the sequences can reveal predominant repeat families, potential regions of interest due to high repeat content, and evolutionary insights into the dataset used.

REFERENCES:

1. Administrator. (2023, July 8). Zenodo - A universal repository for all your research outcomes. OpenAIRE Graph. <https://www.openaire.eu/zenodo-guide>
 2. Dfam. (n.d.). <https://www.dfam.org/home>
 3. Galaxy Training Network. (2024, January 8). Genome Annotation / Hands-on: Masking repeats with RepeatMasker.
<https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/repeatmasker/tutorial.html>
 4. Liehr, T. (2021). Repetitive elements in humans. International Journal of Molecular Sciences, 22(4), 2072. <https://doi.org/10.3390/ijms22042072>
 5. RepeatMasker home page. (n.d.). <https://www.repeatmasker.org/>
 6. Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, 25(1). <https://doi.org/10.1002/0471250953.bi0410s25>
-

WEBLEM 9
INTRODUCTION TO SEQUENCING METHODS

Sequencing is the process of determining the precise order of nucleotides in a DNA or RNA molecule. It is a basic molecular biology approach that aids in the understanding of genetic information at the molecular level. Researchers can investigate microbial diversity, examine gene expression patterns, and learn more about the genetic makeup of organisms through sequencing. Wide-ranging uses of sequencing can be found in domains including as biotechnology, evolutionary biology, medicine, and agriculture. Sequencing fosters innovation and changes our perception of the complexity of life.

Applications of Sequencing:

Sequencing, a pivotal technique in molecular biology, unveils the genetic makeup of organisms, providing insights into genetic variations and molecular processes. Its applications span diverse fields, driving advancements in personalized medicine, agriculture, evolutionary biology, and biotechnology, which have been explained below:

- Medical Research:** Sequencing aids in personalized medicine, disease diagnosis, and drug development by identifying genetic variations.
- Agricultural Research:** Enhances crop improvement through genomics-assisted breeding, optimizing plant traits for better yields and resilience.
- Evolutionary Biology:** Unravels evolutionary histories and genetic adaptations across species, shedding light on biodiversity and speciation processes.
- Biotechnology:** Enables genetic engineering, synthetic biology applications, and bioprospecting for novel bioactive compounds and enzymes.

Significance of Sequencing:

Sequencing holds immense significance in various fields due to its pivotal role in unraveling genetic information, understanding diseases, and advancing scientific knowledge. Studying sequencing is crucial as it enables:

- Precision Medicine:** By identifying genetic variations through sequencing, tailored treatments can be developed for individuals, enhancing therapeutic outcomes and minimizing adverse effects
- Disease Diagnosis:** Sequencing aids in diagnosing genetic disorders, predicting disease risks, and understanding the molecular basis of illnesses, leading to early interventions and personalized healthcare
- Biological Research:** Sequencing contributes to exploring evolutionary relationships, studying gene expression patterns, and uncovering novel biological mechanisms, driving advancements in diverse scientific disciplines
- Therapeutic Development:** Understanding genetic sequences through sequencing facilitates the development of targeted therapies, precision drugs, and gene editing techniques, revolutionizing medical treatments and interventions

In essence, the study of sequencing is essential for unlocking the mysteries of genetics, advancing medical treatments, and driving scientific discoveries that shape our understanding of life at a molecular level.

Types of Sequencing Methods:

Sequencing methodologies encompass various techniques for determining the order of nucleotides in DNA or RNA molecules. Three main methodologies include:

1. **First-Generation Sequencing (Sanger's Sequencing):** Utilizes chain-termination methods to determine DNA sequences, known for accuracy but limited throughput and high cost.
2. **Next-Generation Sequencing (NGS):** Employs parallel sequencing of DNA fragments, offering high throughput and cost-effectiveness, despite limitations like short reads.
3. **Third-Generation Sequencing:** Features long-read capabilities, enhancing analysis of complex genomic regions and improving genome assembly, demonstrating clinical application value.

First-Generation Sequencing:

First-generation sequencing, exemplified by Sanger's sequencing, involves chain-termination methods to determine DNA sequences. This method, known for its accuracy, was pivotal in completing the "working draft" of the human genome. Despite its precision, first-generation sequencing had limitations such as low throughput and high costs, which restricted its widespread application. Overall, first-generation sequencing laid the foundation for subsequent advancements in sequencing technologies, setting the stage for more efficient and cost-effective methodologies in genetic research.

Methods of First-Generation Sequencing:

1. Sanger's Sequencing:

Sanger's Sequencing, also known as the first-generation sequencing technique, is a method that determines DNA sequences by incorporating dideoxynucleotides during DNA replication, by following sequencing by chain-termination principle.

The Sanger's sequencing involves selectively incorporating chain-terminating dideoxynucleotides (ddNTPs) during DNA replication. This method allows for the generation of a series of DNA fragments with different lengths, which are then separated by gel electrophoresis to determine the sequence. By using these chain-terminating nucleotides, Sanger's sequencing enables the identification of the precise order of nucleotides in a DNA molecule, making it a foundational technique in molecular biology and genetic research.

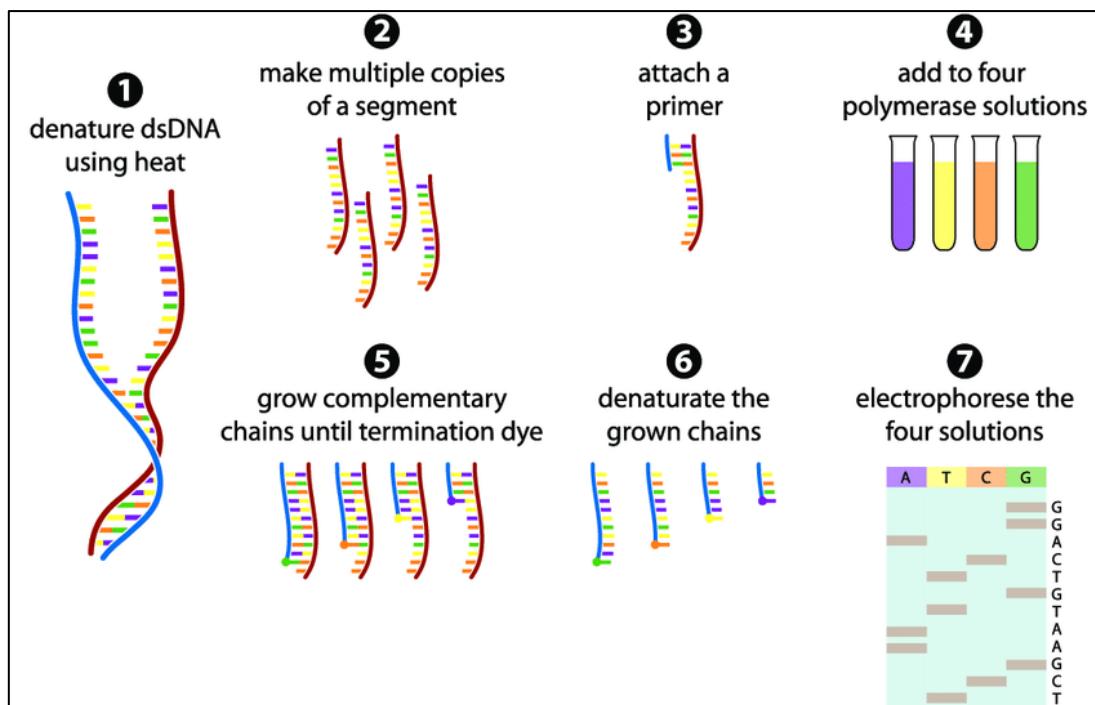


Fig 1: Sanger's Sequencing Method Workflow

2. Maxam-Gilbert's Sequencing:

Maxam-Gilbert sequencing is a method developed in the 1970s for DNA sequencing. It involves chemical cleavage of DNA at specific bases, followed by gel electrophoresis to determine the sequence. This technique, along with Sanger's sequencing, was instrumental in early genetic research and contributed to the completion of the human genome project. The principle behind Maxam-Gilbert sequencing involves chemical cleavage of DNA at specific bases using chemicals like dimethyl sulfate, hydrazine, and piperidine. This method allows for the identification of nucleotide sequences by creating breaks at specific positions in the DNA molecule. By analyzing the fragments produced after cleavage through gel electrophoresis, the sequence of the DNA can be determined.

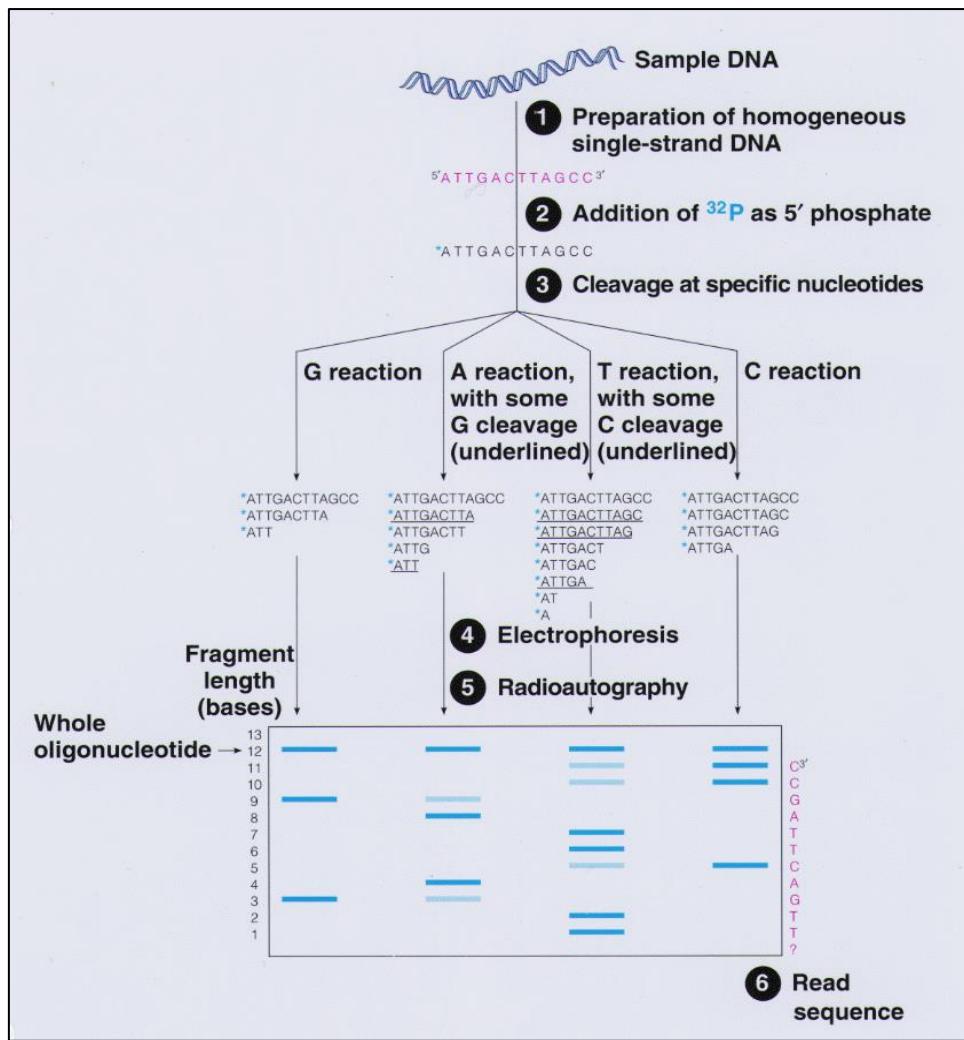


Fig 2: Maxam-Gilbert's Sequencing Method Workflow

Second-Generation Sequencing:

Second-generation sequencing, also known as next-generation sequencing (NGS), revolutionized genomics by enabling high-throughput sequencing of DNA and RNA. This technology, characterized by parallel sequencing of millions of DNA fragments, offers increased speed and cost-effectiveness compared to first-generation Sanger's sequencing. Despite its advantages in high throughput, second-generation sequencing is limited by short read lengths, which can pose challenges in analyzing complex genomic regions and repetitive sequences.

Methods of Second-Generation Sequencing:

1. Roche 454:

The Roche 454 sequencing platform is a technology used for DNA sequencing that was compared with Illumina sequencing technologies in metagenomic studies. Despite differences in read length and sequencing protocols, Roche 454 provided a comparable view of microbial communities sampled. Studies have shown that the Roche 454 platform offers valuable insights into complex microbial communities and genetic diversity, contributing to advancements in genomic research.

The principle behind Roche 454 sequencing involves emulsion PCR and pyrosequencing. In this method, DNA fragments are amplified in water-in-oil emulsions to generate DNA beads, which are then sequenced using a pyrosequencing approach. During pyrosequencing, nucleotides are added one at a time, and the release of pyrophosphate generates a light signal that is detected and used to determine the DNA sequence. This process allows for the generation of longer reads compared to other sequencing technologies, making Roche 454 sequencing valuable for applications requiring detailed genetic analysis and longer sequences.

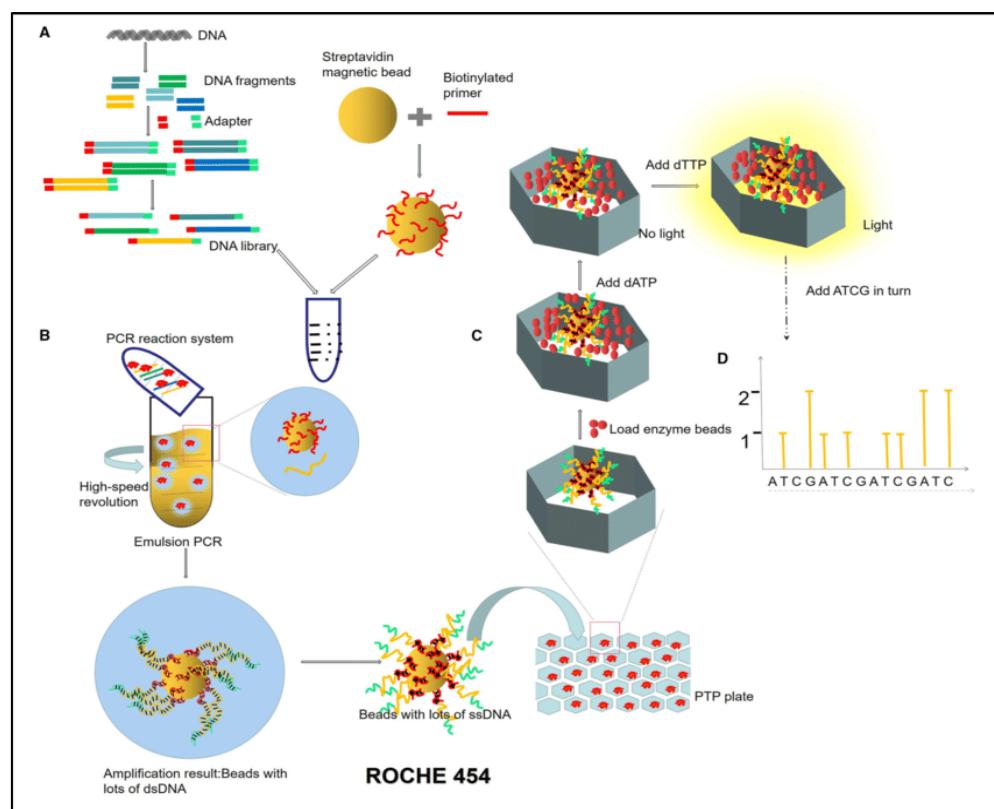


Fig 3 : Roche 454 Sequencing Workflow



Fig 4: Roche 454 Instrument

2. ABI SOLiD Sequencing:

ABI SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing was a technology developed by Applied Biosystems for determining the sequence of DNA.

The ABI SOLiD sequencing method is based on a unique principle involving ligation of fluorescently labeled oligonucleotides to the DNA template. This approach allows for the determination of the DNA sequence by detecting the fluorescent signals emitted during the ligation process. The technology behind ABI SOLiD sequencing enables accurate and high-throughput sequencing, making it a valuable tool in genomic research and analysis.

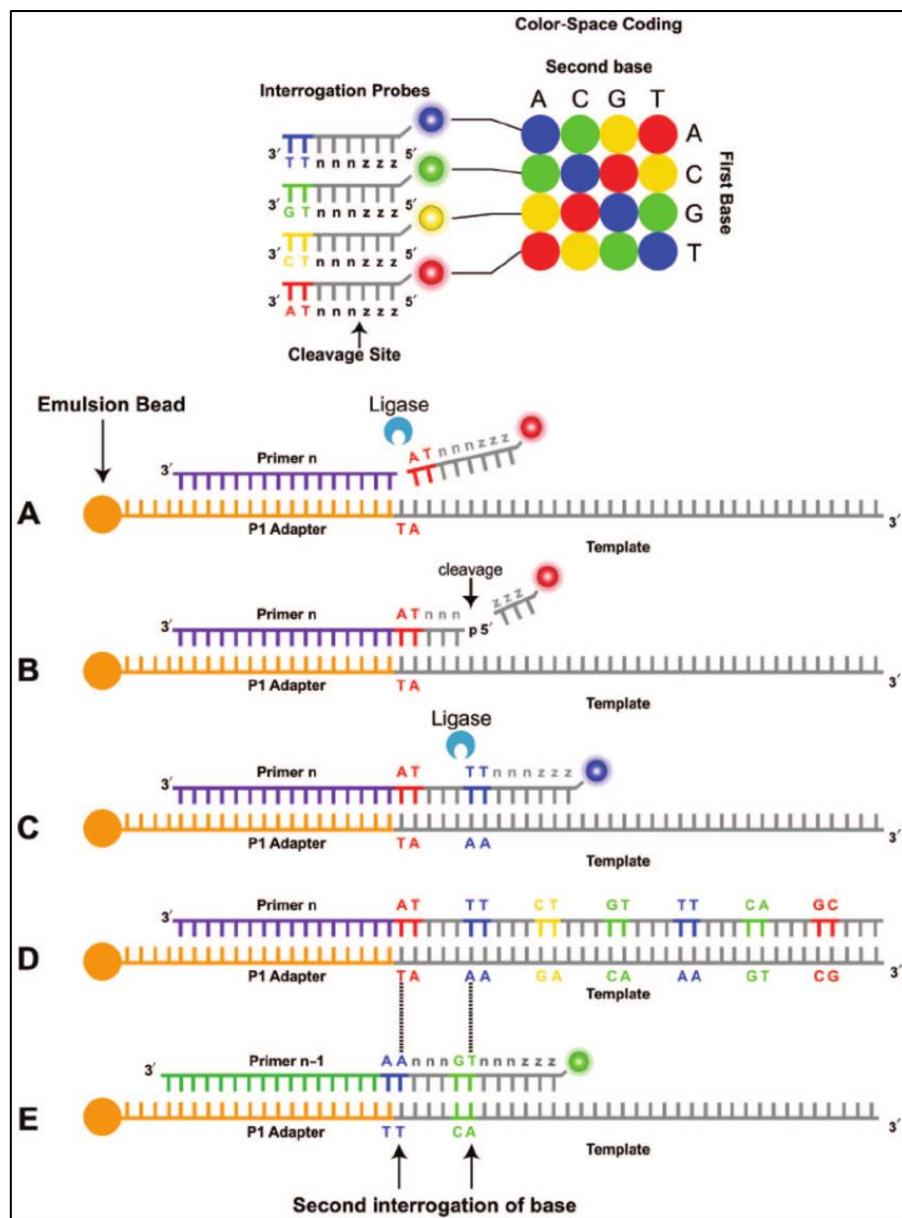


Fig 5: ABI SOLiD Sequencing Workflow



Fig 6: ABI SOLiD Instrument

3. Ion Torrent:

The Ion Torrent sequencing platform, specifically the Ion Torrent Personal Genome Machine (PGM) system, has been validated for detecting gene mutations in biopsy specimens from patients with non-small-cell lung cancer.

The principle behind the Ion Torrent sequencing platform involves detecting changes in pH that occur when nucleotides are incorporated into a DNA strand during sequencing. This technology is based on the release of hydrogen ions as nucleotides are added to the growing DNA strand, leading to a change in pH that is detected by a sensor. By measuring these pH changes, the Ion Torrent platform can determine the DNA sequence, offering a rapid and cost-effective method for DNA sequencing compared to traditional methods like Sanger's sequencing.

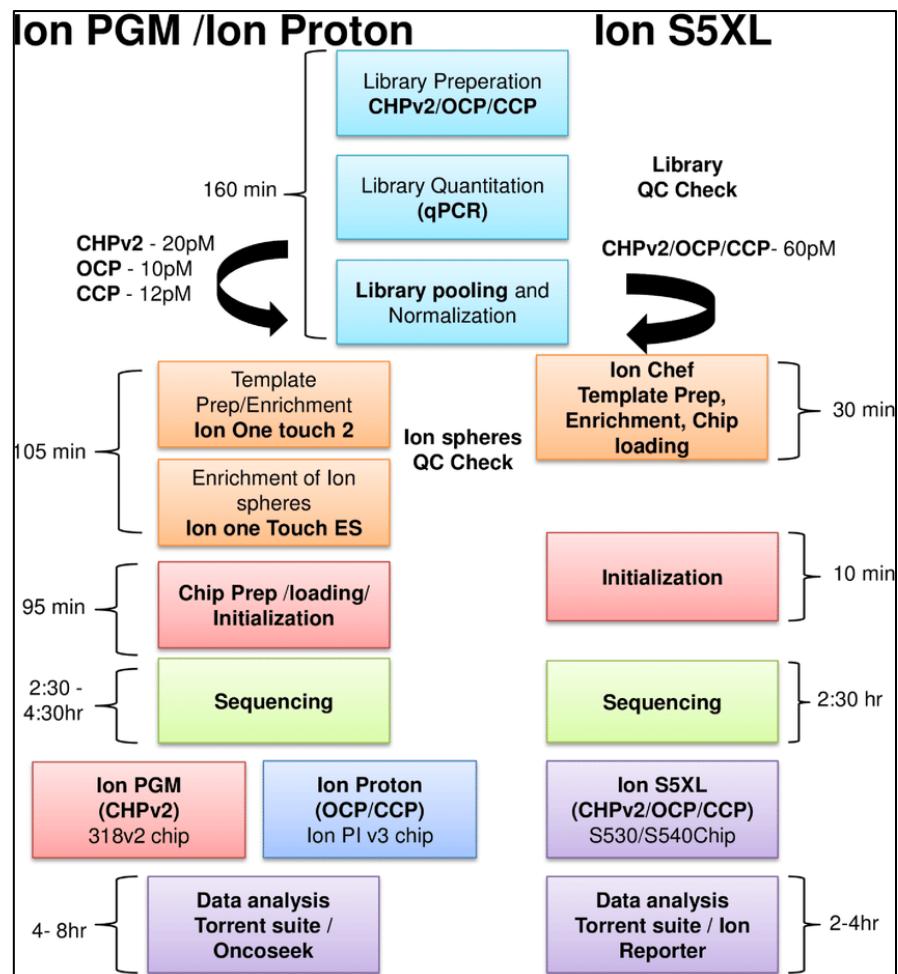


Fig 7: Ion Torrent Sequencing Method Workflow



Fig 8: Ion Torrent Instrument

4. Illumina:

Illumina sequencing is a widely used high-throughput sequencing technology that has revolutionized the field of genomics. It is based on the principle of sequencing-by-synthesis, where a DNA molecule is sequentially read by adding fluorescently labeled nucleotides to the growing DNA strand. The emitted light signals are then captured and used to determine the DNA sequence. Illumina sequencing offers several advantages, such as high-throughput capabilities, cost-effectiveness, and the ability to generate short reads with high accuracy. It is widely used in various applications, including whole-genome sequencing, transcriptome analysis, and metagenomics.

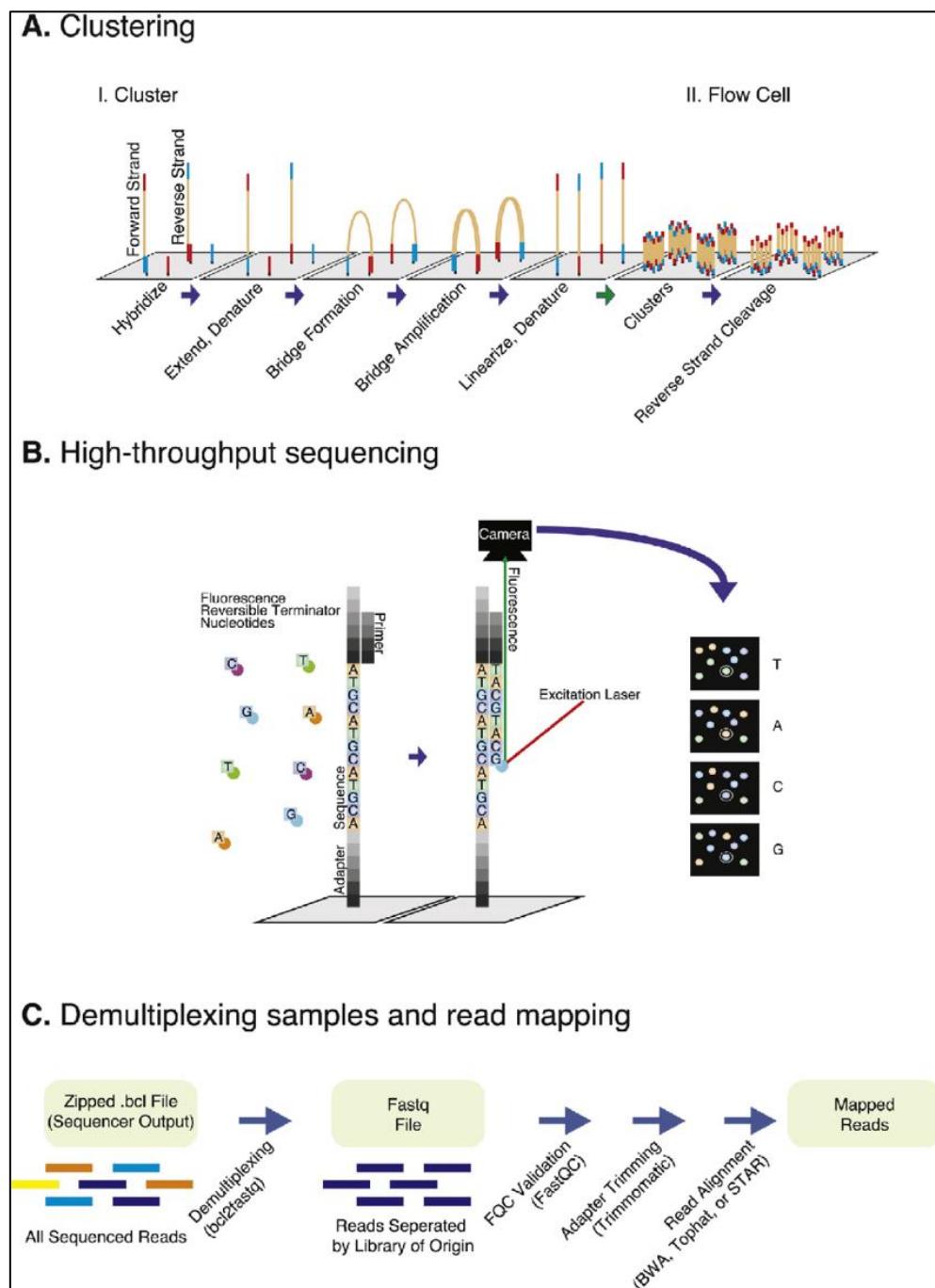


Fig 9: Illumina Sequencing Method Workflow



Fig10: Illumina Sequencing Instrument

Third-Generation Sequencing:

Third-generation sequencing refers to a newer generation of sequencing technologies that offer advantages over previous generations, such as longer read lengths, reduced amplification bias, and the ability to directly sequence native DNA molecules. These technologies, including platforms like PacBio and Oxford Nanopore, have revolutionized genetic research by enabling the sequencing of long DNA fragments without the need for fragmentation or amplification.

Methods of Third-Generation Sequencing:

1. PacBio's Single-Molecule Real-Time (SMRT) Sequencing:

Single-molecule real-time (SMRT) sequencing is a third-generation sequencing technology developed by Pacific Biosciences (PacBio).

The principle behind Single-Molecule Real-Time (SMRT) sequencing, involves the real-time detection of individual nucleotides as they are incorporated into a growing DNA strand. This technology utilizes zero-mode waveguides (ZMWs) to trap and sequence single DNA molecules, enabling long reads and the detection of complex genomic features like repetitive sequences and structural variants. SMRT sequencing offers advantages such as high accuracy, long read lengths, and the ability to directly sequence native DNA molecules without the need for fragmentation or amplification. This innovative approach has significantly advanced genetic research by providing high-throughput and high-sensitivity capabilities that were previously challenging with older sequencing technologies.

This technology has been instrumental in various applications, including genome assembly, transcriptome analysis, and CRISPR-Cas9 gene editing, offering high-throughput and high-sensitivity capabilities that are advantageous over previous sequencing technologies.

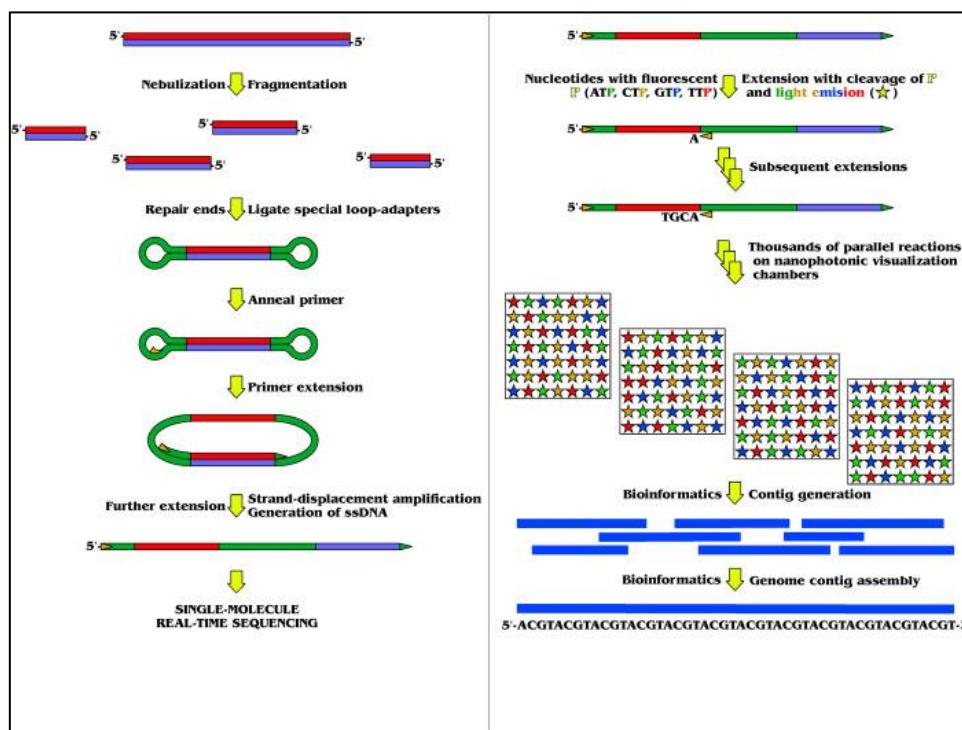


Fig 11: Single-Molecule Real-Time (SMRT) Sequencing Workflow

2. Oxford Nanopore Sequencing:

Oxford Nanopore sequencing is a cutting-edge technology that involves monitoring the progress of DNA molecules through a membrane pore for sequencing. This method utilizes nanopore-based sequencing instruments like the Oxford Nanopore MinION, which can sequence very long DNA molecules in real-time. Despite an accuracy of around 90%, Oxford Nanopore sequencing offers the advantage of generating long reads, sequencing RNA directly, and detecting modified bases.

The principle behind Oxford Nanopore sequencing involves passing DNA molecules through a nanopore embedded in a membrane and measuring changes in electrical current as individual nucleotides pass through the pore. This technology allows for the real-time sequencing of DNA or RNA molecules by detecting the unique electrical signals produced by different nucleotides. Detection of modified bases, makes it a valuable tool for various genomic applications due to its ability to provide real-time data and sequence long molecules.

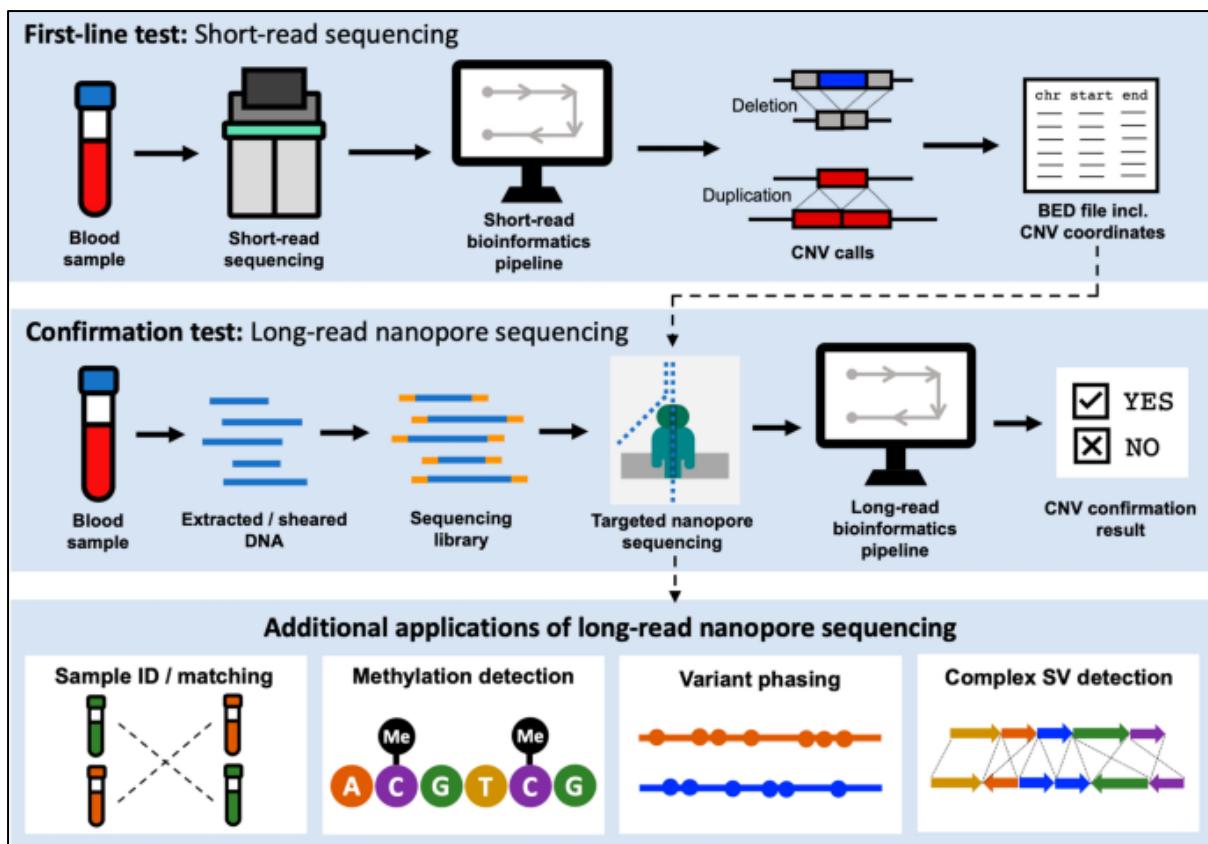


Fig 12: Oxford Nanopore Sequencing Workflow



Fig 13: Oxford Nanopore Instrument

Galaxy Europe Server:

The Galaxy Server, accessible at <https://usegalaxy.eu> and maintained by the Freiburg Galaxy Team as well as a global community of researchers, developers, and infrastructure providers as, is a part of the larger Galaxy Project, which is a robust platform for bioinformatics analysis. It offers a user-friendly web interface, enabling researchers to upload, analyze, and visualize

data without extensive programming skills. Collaboration is encouraged through the sharing of workflows, tools, and datasets. The platform promotes scalability and reproducibility, allowing users to create customized workflows and integrate diverse bioinformatics tools and databases for a wide range of analyses.

The Galaxy Server provides access to:

1. A huge compute and storage resource without any charge
2. More than 2500 different, well-documented and constantly maintained scientific tools
3. 250 GB per user (500 GB for ELIXIR members)
4. Free registration

Data Analysis and Visualization

Galaxy, accessible via web browser, offers powerful data analysis tools without requiring programming knowledge. It automates computation on a cluster and cloud, providing ample storage (250 GB per user) and API access for advanced users. It supports publication-ready visualizations and interactive tools like Trackster and Phinch.

Reproducibility and Transparency

Galaxy ensures reproducibility through its history feature, capturing inputs, parameters, and tool versions, shareable even outside Galaxy. It facilitates workflow creation from histories or scratch, downloadable and shareable, promoting transparency and avoiding vendor lock-in. It hosts thousands of tools with fixed versions managed by Bioconda and BioContainers, with 4 TB of reference data available.

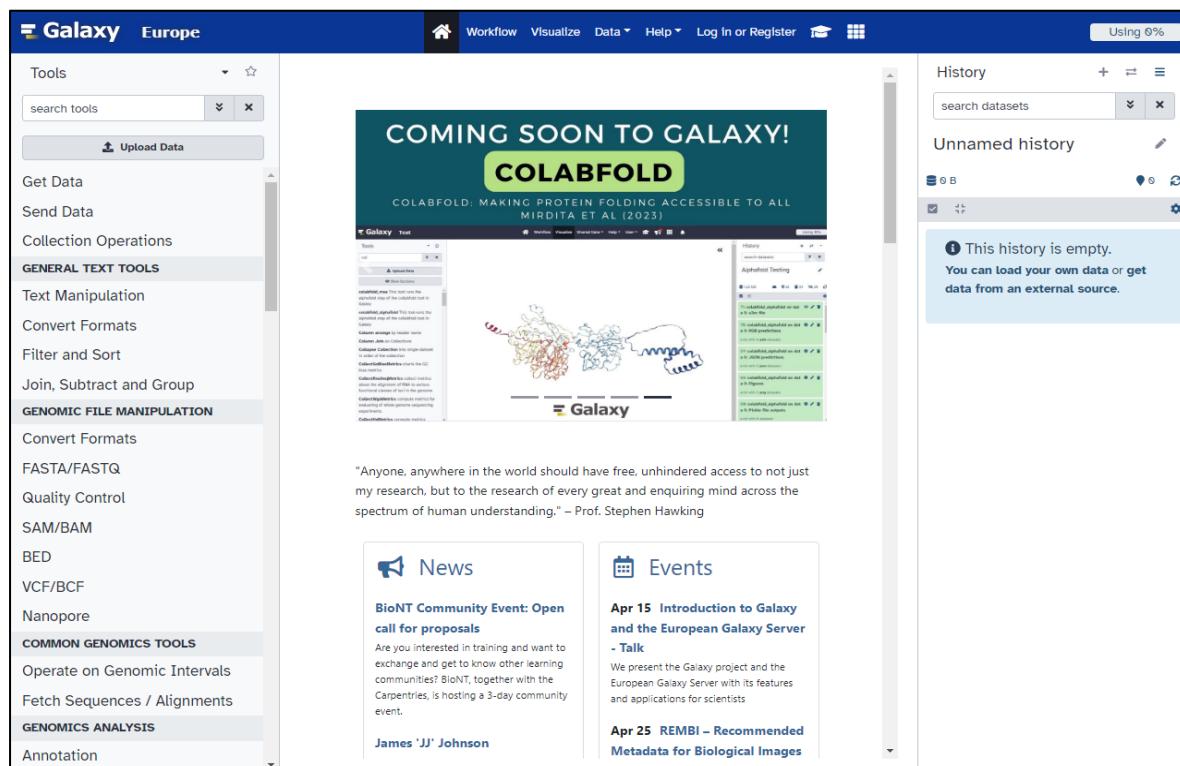


Fig 14: Homepage of Galaxy Server (Europe)

REFERENCES:

1. Dey, P. (2018). Sanger Sequencing and Next-Generation Gene Sequencing: Basic Principles and Applications in Pathology. *Basic and Advanced Laboratory Techniques in Histopathology and Cytology*, 227–231. https://doi.org/10.1007/978-981-10-8252-8_23
2. Tan, D., & Ou, T. (2022, February 25). [Research progress and clinical application of the third-generation sequencing techniques]. *Sheng Wu Gong Cheng Xue Bao (Chinese Journal of Biotechnology)*, 38(9), 3121–3130. <https://doi.org/10.13345/j.cjb.220063>
3. Di Maio, A., & De Castro, O. (2013, January). SSR-patchwork: An optimized protocol to obtain a rapid and inexpensive SSR library using first-generation sequencing technology. *Applications in Plant Sciences*, 1(1). <https://doi.org/10.3732/apps.1200158>
4. De Cabo, S. F., Fernández-Piqueras, J., & Visedo, G. (1994, March). An adaptation of the maxam and gilbert sequencing method to the study of human metaphase chromosomes. *Histochemistry*, 101(3), 205–207. <https://doi.org/10.1007/bf00269545>
5. Janvier, A., Barrington, K., & Lantos, J. (2022, March 29). Next generation sequencing in neonatology: what does it mean for the next generation? *Human Genetics*, 141(5), 1027–1034. <https://doi.org/10.1007/s00439-022-02438-9>
6. Luo, C., Tsementzi, D., Kyripides, N., Read, T., & Konstantinidis, K. T. (2012, February 10). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE*, 7(2), e30087. <https://doi.org/10.1371/journal.pone.0030087>
7. Second-Generation Sequencing. (2021). Encyclopedia of Autism Spectrum Disorders, 4114–4114. https://doi.org/10.1007/978-3-319-91280-6_301420
8. Chen, G., Qiu, Y., Zhuang, Q., Wang, S., Wang, T., Chen, J., & Wang, K. (2018, May 9). Next-generation sequencing library preparation method for identification of RNA viruses on the Ion Torrent Sequencing Platform. *Virus Genes*, 54(4), 536–542. <https://doi.org/10.1007/s11262-018-1568-x>
9. Ling, X., Wang, C., Li, L., Pan, L., Huang, C., Zhang, C., Huang, Y., Qiu, Y., Lin, F., & Huang, Y. (2023, November). Third-generation sequencing for genetic disease. *Clinica Chimica Acta*, 551, 117624. <https://doi.org/10.1016/j.cca.2023.117624>
10. Meslier, V., Quinquis, B., Da Silva, K., Plaza Oñate, F., Pons, N., Roume, H., Podar, M., & Almeida, M. (2022, November 11). Benchmarking second and third-generation sequencing platforms for microbial metagenomics. *Scientific Data*, 9(1). <https://doi.org/10.1038/s41597-022-01762-z>
11. Su, K., Guo, Y., Zhao, Y., Gao, H., Liu, Z., Li, K., Ma, L., & Guo, X. (2019, November 15). Candidate genes for grape white rot resistance based on SMRT and Illumina sequencing. *BMC Plant Biology*, 19(1). <https://doi.org/10.1186/s12870-019-2119-x>
12. Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015, October 7). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11), 1750–1756. <https://doi.org/10.1101/gr.191395.115>
13. McIntyre, A. B. R., Rizzardi, L., Yu, A. M., Alexander, N., Rosen, G. L., Botkin, D. J., Stahl, S. E., John, K. K., Castro-Wallace, S. L., McGrath, K., Burton, A. S., Feinberg, A.

- P., & Mason, C. E. (2016, October 20). Nanopore sequencing in microgravity. *Npj Microgravity*, 2(1). <https://doi.org/10.1038/npjmgrav.2016.35>
14. Jagadeeswaran, P., & Kaul, R. K. (1986, September). Use of reverse-phase chromatography in the Maxam-Gilbert method of DNA sequencing A step toward automation. *Gene Analysis Techniques*, 3(5), 79–85. [https://doi.org/10.1016/0735-0651\(86\)90007-5](https://doi.org/10.1016/0735-0651(86)90007-5)
15. About. (n.d.). Galaxy Europe. <https://usegalaxy-eu.github.io/about>
16. Galaxy. (n.d.). Galaxy Europe. <https://usegalaxy.eu/>
-

DATE: 23/03/2024

WEBLEM 9(A)
NGS FILE FORMATS

- 1. NCBI SRA DATABASE (URL: <https://www.ncbi.nlm.nih.gov/sra>)**
- 2. GALAXY EUROPE TOOL (URL: <https://usegalaxy.org/>)**

AIM:

To study the NGS file format for the sequence read with the Accession ID: ‘SRR26208365’ for the query ‘*Klebsiella pneumoniae*’ using the NCBI Sequence Read Archive (SRA) database and Galaxy Europe platform.

INTRODUCTION:

Sequence Read Archive (SRA) Database

The SRA is NIH's archive of high-throughput sequencing data and is part of the International Nucleotide Sequence Database Collaboration (INSDC) that includes the NCBI Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ). Data submitted to any of the three organizations are shared among them. The SRA is a publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

SRA accepts data from all kinds of sequencing projects including clinically important studies that involve human subjects or their metagenomes, which may contain human sequences. These data often utilize NIH controlled access via dbGaP (the database of Genotypes and Phenotypes). Data submitters need to determine if their data is suitable for public distribution or if it needs controlled access. It is the responsibility of submitting parties to ensure that they have appropriate consent for human sequence data to be distributed publicly without access controls. Following submission, data are subject to automated and manual processing to ensure data integrity and quality and are subsequently made available to the public.

Galaxy Europe Platform

The Galaxy Europe platform is an open, web-based tool designed for accessible, reproducible, and transparent computational biological research. It offers users the ability to run over 2500 scientific tools without the need for coding or command-line interfaces, all through a user-friendly web interface. This open-source platform facilitates the exploration of extensive biological datasets through a comprehensive suite of bioinformatics tools, accessible via a drag-and-drop interface. This platform emphasizes reproducibility by capturing all metadata from analyses, making them completely reproducible. Users can easily share and publish analyses through interactive pages, enhancing them with annotations. Galaxy Europe is scalable, capable of running on various systems from laptops to large clusters or the cloud. It provides free registration, 250 GB per user (500 GB for ELIXIR members), and on-demand training capacity, making it a valuable resource for researchers in the field of bioinformatics.

Klebsiella pneumoniae

Klebsiella pneumoniae is a type of bacteria commonly found in the intestines, typically harmless in healthy individuals but can pose serious risks if it enters other parts of the body, especially in individuals with certain health conditions. This bacterium can lead to severe infections like pneumonia, urinary tract infections, bloodstream infections, and more, particularly affecting individuals in healthcare settings or with compromised immune systems. *Klebsiella pneumoniae* infections are often acquired in hospitals and can be challenging to treat due to increasing antibiotic resistance, with some strains being resistant to common antibiotics. Preventing the spread of *Klebsiella* infections involves strict infection control measures, including hand hygiene and proper cleaning procedures in healthcare facilities. Early diagnosis and appropriate antibiotic treatment are crucial for managing *Klebsiella pneumoniae* infections effectively.

METHODOLOGY:

1. Open the NCBI SRA Homepage (<https://www.ncbi.nlm.nih.gov/sra>).
2. Search for the query ‘*Klebsiella pneumoniae*’ and apply the filter – Source: DNA.
3. Select a particular sequence read from the list of hits obtained and note its Accession ID ('SRR26208365').
4. Open the Galaxy Europe Platform (<https://usegalaxy.org/>).
5. Search for the ‘fasterq’ tool in the Search dialog box.
6. After selecting the first tool ‘**Faster Download and Extract Reads in FASTQ format from NCBI SRA**’, enter the Accession ID of the selected sequence read. ('SRR26208365').
7. Click on ‘Run Tool’.
8. Note the results obtained in the form of fasterq-dump files and a log file.

OBSERVATIONS:

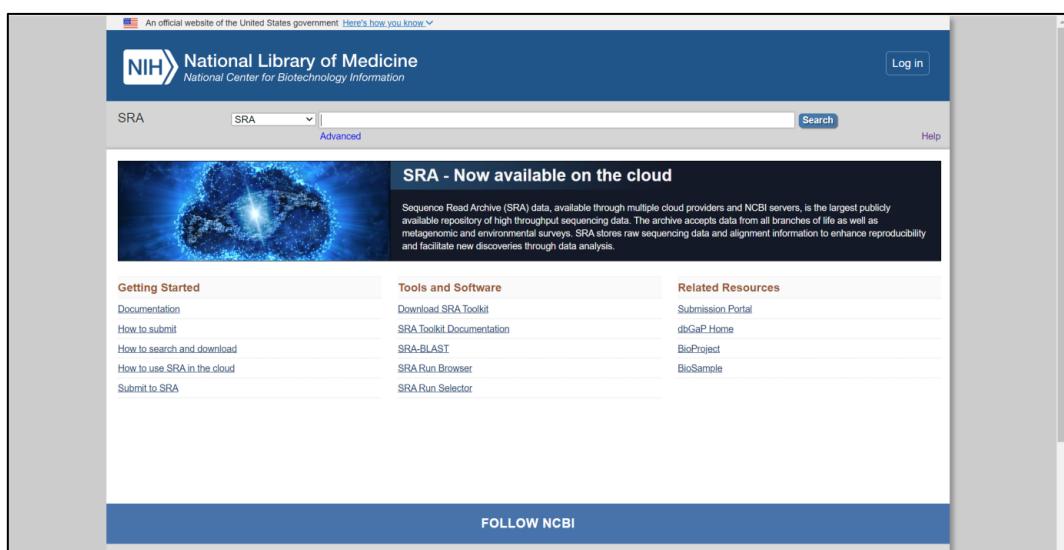


Fig 1: Homepage of NCBI SRA Database

The screenshot shows the National Library of Medicine's SRA search interface. The search term 'klebsiella pneumoniae' is entered in the search bar. A red box highlights the 'Source' dropdown menu, which is set to 'DNA (94,886)'. The search results table shows 20 items per page, with 1 item listed: 'WGS of 2024GN-00081'. The right sidebar includes filters for 'Manage Filters', 'Results by taxon' (listing Klebsiella pneumoniae as the top organism), 'Search in related databases' (showing BioSample, BioProject, dbGaP, and GEO Datasets counts), and 'Find related data'.

Fig 2: Query – ‘*Klebsiella pneumoniae*’ searched and filter applied (Source: DNA) and 94886 hits were obtained

This screenshot displays detailed information for Run 'SRR26208365'. The run is identified as 'WGS of Klebsiella pneumoniae' from an Illumina MiSeq instrument. The 'Run' table is highlighted with a red box, showing the following data:

Run	# of Spots	# of Bases	Size	Published
SRR26208365	2,360,751	1.1G	680Mb	2023-09-28

The right sidebar contains sections for 'Related information' (BioProject, BioSample, Taxonomy) and 'Recent activity' (listing search terms like 'biomol dna', 'Klebsiella pneumoniae', and 'galaxy europe').

Fig 3: Information displayed for Run ‘SRR26208365’ to be used for further analysis

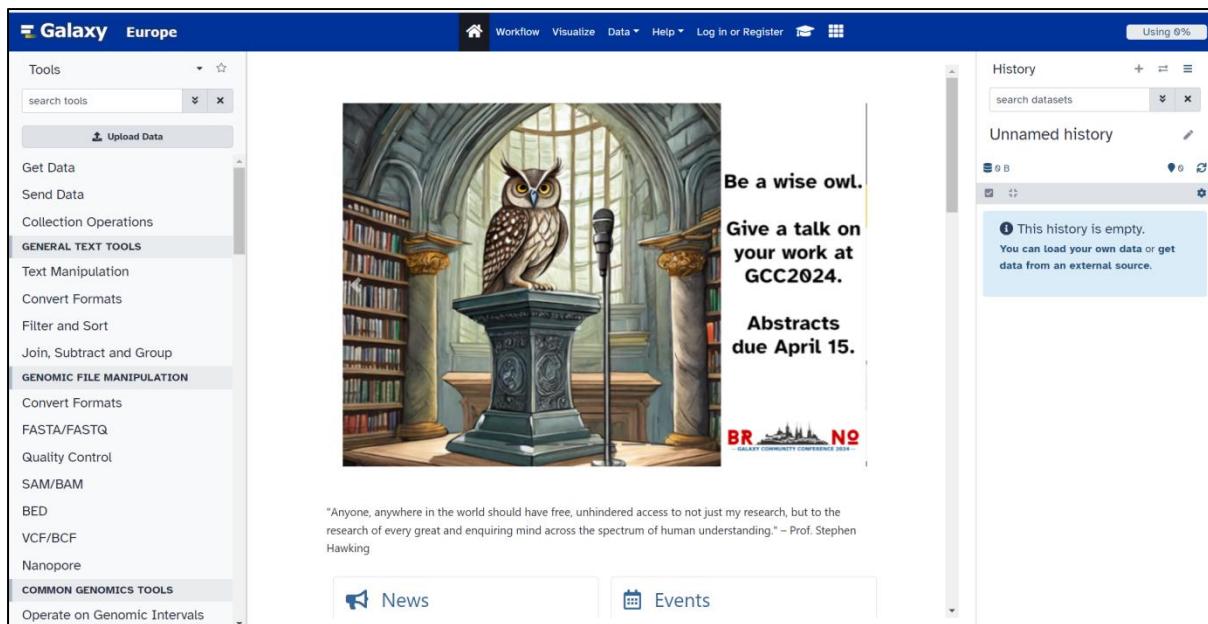


Fig 4: Homepage of Galaxy Europe tool

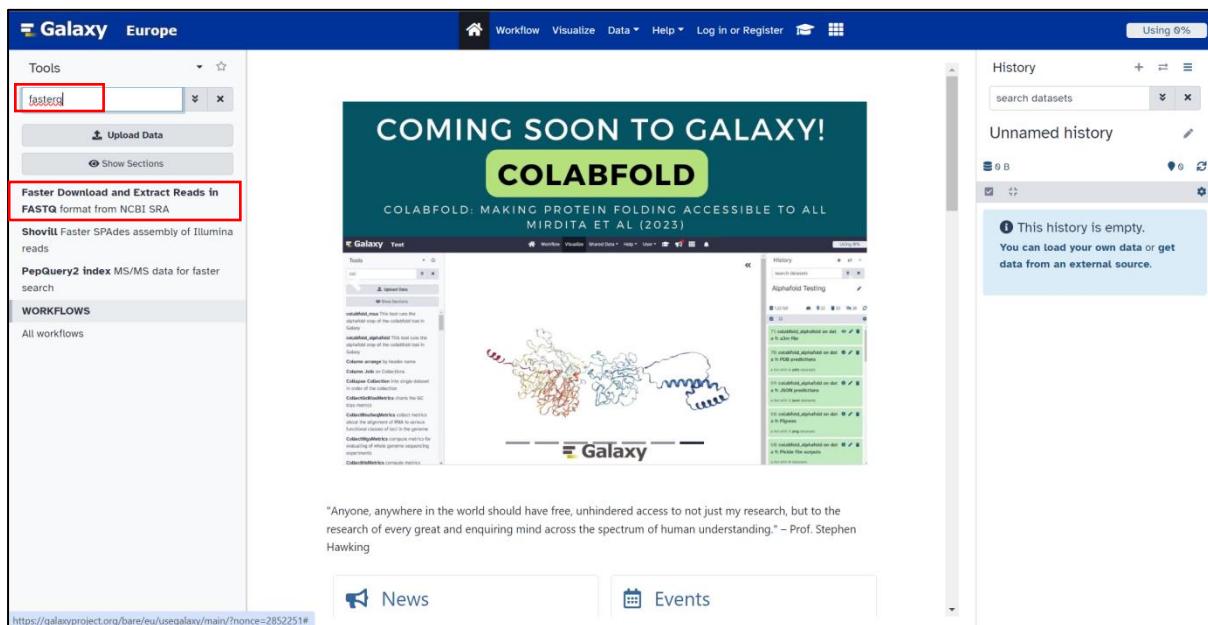


Fig 5: Searching for the 'fasterq' tool

The screenshot shows the Galaxy Europe interface. On the left, there's a sidebar with 'Tools' (fasterq), 'Upload Data', 'Show Sections', and a section for 'Faster Download and Extract Reads in FASTQ format from NCBI SRA'. Below that is a 'WORKFLOWS' section with 'All workflows'. The main area has a title 'Faster Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 3.0.10+galaxy0)'. Under 'Tool Parameters', there's a dropdown for 'select input type' set to 'SRR accession', and an 'Accession' input field containing 'SRR26208365'. A note below says 'Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343869'. There's also an 'Advanced Options' section with a 'Define format specification for sequence' field containing '@\$sn/\$ri' and a note about variables. The 'Minimum read length' field is empty. The 'Select how to split the spots' section has a radio button selected for '--split-3: write properly paired biological reads into different files and single reads in another file'. The right side shows a 'History' panel with an empty history named 'Unnamed history'.

Fig 6: Entering the Accession ID of the selected run ‘SRR26208365’

This screenshot shows the same Galaxy Europe interface as Fig 6, but with the 'Run Tool' button highlighted with a red box. The 'Run Tool' button is located at the bottom of the main form area. The rest of the interface is identical to Fig 6, including the tool parameters and the empty history panel.

Fig 6.1: Display of the Advanced Options

The screenshot shows the Galaxy Europe web interface. On the left, a sidebar lists various tools under categories like FASTQ, FASTA, and SAM/BAM. A message at the top right indicates a successful job submission:

- Started tool **Faster Download and Extract Reads in FASTQ** and successfully added 1 job to the queue.
- It produces this output:
- 14: fasterq-dump log**

Below this, a section titled "Your feedback helps us!" contains a cartoon illustration of laboratory glassware and the text: "Your feedback helps us! Your feedback helps us to improve Open Research and Galaxy. Any feedback is welcome."

Fig 7: Job Running Image for Fastq

The screenshot shows the Galaxy Europe interface with the "Job Information" panel highlighted. The "History" panel on the right displays the results of the job:

- 4: fasterq-dump log** (a list with 1 dataset)
- 3: Other data (fasterq-dump)** (a list with 0 datasets)
- 2: Single-end data (fasterq-dump)** (a list with 0 datasets)
- 1: Pair-end data (fasterq-dump)** (a list with 1 fastqsanger.gz pair)

Fig 8: Viewing the results obtained in the ‘Pair-end data (fasterq-dump)’ file: Job Information

The screenshot shows the Galaxy Europe web interface. On the left, a sidebar lists various tools under categories like GENERAL TEXT TOOLS and GENOMIC FILE MANIPULATION. The main area displays 'Job Parameters' and 'Job Outputs'. The 'Job Parameters' section includes fields for Tool Standard Error (empty), Tool Exit Code (0), Job API ID (11ac94870d0bb33a5f1bc9d30036762a), and a detailed table of input parameters. The 'Job Outputs' section shows a table with columns 'Tool Outputs' and 'Dataset', listing results such as 'fasterq-dump log' and 'Pair-end data (fasterq-dump)'. To the right, the 'History' panel shows a list of datasets, with the '1: Pair-end data (fasterq-dump)' entry highlighted with a red box.

Fig 8.1: Viewing the results obtained in the ‘Pair-end data (fasterq-dump)’ file: Job Parameters

This screenshot is similar to Fig 8.1 but focuses on the 'Job Outputs' section. It shows a detailed list of datasets produced by the job, categorized by tool output type. The 'Tool Outputs' column lists items like 'fasterq-dump log', 'Pair-end data (fasterq-dump)', 'Single-end data (fasterq-dump)', and 'Other data (fasterq-dump)'. The 'Dataset' column provides more detail for each item, such as '4: fasterq-dump log' being a list with 0 datasets. The 'History' panel on the right shows the same list of datasets, with the '1: Pair-end data (fasterq-dump)' entry highlighted with a red box.

Fig 8.2: Viewing the results obtained in the ‘Pair-end data (fasterq-dump)’ file: Job Outputs

The screenshot shows the Galaxy Europe web interface. On the left, the 'Tools' panel is open, displaying various categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. The 'Job Metrics' section is highlighted with a red box. It contains detailed information about a job named 'fasterq-dump log' (ID: 4). The metrics include CPU Time (7 minutes), Failed to allocate memory count (0E-7), Memory limit on cgroup (30.5 GB), and OOM Control enabled (Yes). The 'History' panel on the right shows datasets related to the job, including 'Pair-end data (fasterq-dump)' (ID: 1) which is highlighted with a red box. This dataset is described as a 'list with 1 fastqsanger.gz pair'.

Fig 8.3: Viewing the results obtained in the ‘Pair-end data (fasterq-dump)’ file:Job Metrics

This screenshot shows the same Galaxy Europe interface as Fig 8.3. The 'History' panel now displays the contents of the 'Pair-end data (fasterq-dump)' dataset. A specific entry, 'SRR26208365', is highlighted with a red box. It is described as a 'pair with datasets'. Below it, the '1: forward' sequence is shown in a large green box, also highlighted with a red box. The sequence starts with '@M07713:19:000000000-KYD9J:1:1101:17677:2398/1' and continues with several lines of DNA sequence data.

Fig 8.4: Viewing the forward read sequence after clicking on the ‘Pair-end data (fasterq-dump)’ tab

This screenshot shows the Galaxy Europe interface. In the top right, there's a message: "This dataset is large and only the first megabyte is shown below. Show all | Save". The main area shows a large block of sequence data for a dataset named "SRR26208365". The data is organized into two sections: "1: forward" and "2: reverse". A red box highlights the "2: reverse" section. Below the sequence data, there's a note: "format fastqsanger.gz, database ?". On the left sidebar, under "GENERAL TEXT TOOLS", there are several options like Text Manipulation, Convert Formats, Filter and Sort, and Join, Subtract and Group.

Fig 8.5: Viewing the reverse read sequence after clicking on the ‘Pair-end data (fasterq-dump)’ tab

This screenshot shows the Galaxy Europe interface. The left sidebar has "GENERAL TEXT TOOLS" selected. In the center, there's a log entry: "spots read : 2,360,751", "reads read : 4,721,502", and "reads written : 4,721,502". To the right, there's a "History" panel titled "NGS_File format". It shows a log entry for "4: fasterq-dump log" with a red box around it. The log entry includes: "format txt, database ?", "Downloading accession: SRR26208365...", and "spots read : 2,360,751". Below this, there's another entry: "3: Other data (fasterq-dump)".

Fig 9: Viewing the log file of the complete run

RESULTS:

Using the NCBI Sequence Read Archive (SRA) database and Galaxy Europe platform, the fasterq-dump NGS file format was obtained and further studied for the sequence read with the Accession ID: ‘SRR26208365’ for the query ‘*Klebsiella pneumoniae*’. Job ‘information, Job Parameters, Job Outputs and Job Metrics were observed when viewing the ‘Pair-end data (fasterq-dump’ file, along with viewing the sequences for the forward read and the reverse read. When viewing the log file of the run, it was found that 2,360,751 spots were read, 4,721,502 reads were read and 4,721,502 reads were written.

CONCLUSION:

The NGS file format was obtained and further studied for the sequence read with the Accession ID: ‘SRR26208365’ for the query ‘*Klebsiella pneumoniae*’ using the NCBI Sequence Read Archive (SRA) database and Galaxy Europe platform.

REFERENCES:

1. The Sequence Read Archive (SRA). (n.d.). <https://www.ncbi.nlm.nih.gov/sra/docs/>
 2. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1), W537–W544. <https://doi.org/10.1093/nar/gky379>
 3. Galaxy Europe. (n.d.). <https://galaxyproject.eu/about.html>
 4. Nunez, K. (2022, June 2). What You Need to Know About a *Klebsiella pneumoniae* Infection. Healthline. <https://www.healthline.com/health/klebsiella-pneumonia>
 5. Klebsiella pneumoniae in Healthcare Settings | HAI | CDC. (n.d.). <https://www.cdc.gov/hai/organisms/klebsiella/klebsiella.html>
-

DATE: 10/04/2024

WEBLEM 9(B)
TRIMMOMATIC
(URL: <https://usegalaxy.org/>)

AIM:

To improve overall quality of NGS raw data by using trimmomatic tool for the given query.

INTRODUCTION:

Trimming refers to removing unwanted or low-quality sequences from high-throughput sequencing data. Trimming helps to remove the regions of low confidence, sequencing artifacts, adapter sequences, and low-quality bases. This means that these artifacts and errors have to be removed, and this process of removal is known as trimming. By performing trimming data, bioinformaticians can obtain cleaner, more accurate, and more reliable sequencing data. This is further important in obtaining high-quality results in various downstream bioinformatics applications, such as genome assembly, variant calling, gene expression analysis, and other biological investigations.

Trimmomatic is one of the most popular bioinformatics tools for quality control (QC) and next-generation sequencing (NGS) data preprocessing. It is widely used due to its efficiency, flexibility, and ability to work with various sequencing data formats. Trimmomatic's main functionality is to remove low-quality regions and sequencing artifacts from raw NGS reads, ensuring that only high-quality, reliable data is used for the downstream analysis. Trimmomatic is a widely-used tool for this purpose, as it can handle both single-end and paired-end data and supports various types of adapter sequences.

Before trimming data, you may observe sequences with varying lengths, potentially containing adapter sequences, low-quality bases, and regions with poor sequencing quality. The data may also contain artifacts introduced during the sequencing process, such as sequencing errors or biases. After trimming the data, you would typically see cleaner and more uniform sequencing reads. Adapter sequences would have been removed from the reads, ensuring they do not interfere with downstream analysis. Low-quality bases at the ends of reads would have been trimmed, improving the overall quality of the data. Regions of poor sequencing quality would have been identified and trimmed using Trimmomatic's sliding window approach, resulting in more reliable sequencing reads for subsequent analysis.

In Trimmomatic, various parameters can be set to customize the trimming process for NGS data. Some of the key parameters include:

1. **ILLUMINACLIP:** Cut adapter and other illumina-specific sequences from the read.
2. **SLIDINGWINDOW:** Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
3. **MINLEN:** Drop the read if it is below a specified length.
4. **LEADING:** Cut bases off the start of a read, if below a threshold quality.
5. **TRAILING:** Cut bases off the end of a read, if below a threshold quality.
6. **CROP:** Cut the read to a specified length.

7. **HEADCROP:** Cut the specified number of bases from the start of the read.
8. **AVGQUAL:** Drop the read if the average quality is below a specified value.
9. **MAXINFO:** Trim reads adaptively, balancing read length and error rate to maximise the value of each read.

METHODOLOGY:

FastQC was performed but the read quality was not so good, thus this tool is used to improve the quality of the reads.

1. Open homepage of Galaxy Tool.
2. Search Trimomatic in search tool box.
3. Set different parameters of the tool.
4. Then set the different operations.
5. Run the tool and interpret the result which are come in four different sections.

OBSERVATIONS:

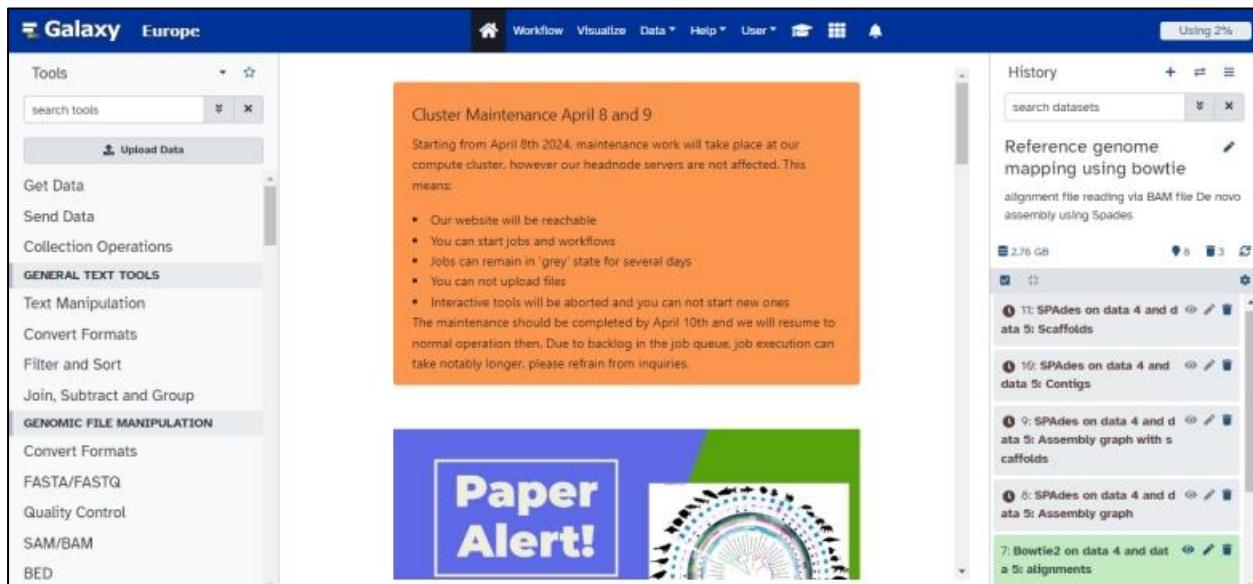


Fig 1: Homepage of Galaxy Europe Tool

The screenshot shows the Galaxy web interface. In the top navigation bar, there are links for Workflow, Visualize, Data, Help, Log in or Register, and a user icon. On the left, a sidebar titled 'Tools' lists several options, including 'Trim leading or trailing characters', 'trimest', 'trimseq', 'Trimmomatic flexible read trimming tool for Illumina NGS data' (which is highlighted with a red border), 'Trim Galore!', 'Trim sequences', 'Trim.flows', and 'Trim.seqs'. The main content area displays details for the selected 'Trimmomatic' tool. It includes a brief description: 'flexible read trimming tool for Illumina NGS data (Galaxy Version 0.38.1)', information about downstream tools, and a note about output being a single FASTQ file. Below this, there's a 'Credits' section mentioning development at the University of Manchester, and links to the Trimmomatic website and documentation. A reference citation is also provided.

Fig 2: Select the Trimmomatic option in search tool box

This screenshot shows the Galaxy Europe interface. The left sidebar lists various tools, including 'Trimmomatic flexible read trimming tool for Illumina NGS data'. The main panel shows the 'Tool Parameters' configuration for this tool. Under 'Single-end or paired-end reads?', it is set to 'Paired-end (two separate input files)'. There are two input fields: 'Input FASTQ file (R1/first of pair)' containing '4: BWA-Sample-2' and 'Input FASTQ file (R2/second of pair)' containing '5: BWA-Sample-1'. Both inputs have dropdown menus for 'accepted formats'. Under 'Perform initial ILLUMINACLIP step?', a dropdown menu shows 'no'. At the bottom, there is a note: 'Cut adapter and other illumina-specific sequences from the read'. The right side of the panel is labeled 'Trimmomatic Operation'.

Fig 3: Set the tool parameters

Galaxy Europe

Tools

- trim
- Upload Data
- Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data
(Galaxy Version 0.39+galaxy2)

Cut adapter and other Illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform: Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across *: 4

Average quality required *: 20

+ Insert Trimmomatic Operation

Quality score encoding - optional: Nothing selected

Fig 4: Set the trimmomatic operations

Galaxy Europe

Tools

- trim
- Upload Data
- Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data
(Galaxy Version 0.39+galaxy2)

2: Trimmomatic Operation

Select Trimmomatic operation to perform: Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across *: 4

Average quality required *: 20

+ Insert Trimmomatic Operation

Quality score encoding - optional: Nothing selected

The phred+64 encoding works the same as the phred+33 encoding, except you add 64 to the phred score to determine the ascii code of the quality character. You will only find phred+64 encoding on older data, which was sequenced several years ago. FASTQC can be used in order to identify the encoding type.

Fig 4.1: Set the trimmomatic operations

The screenshot shows the Galaxy Europe web interface. In the top navigation bar, there are links for Workflow, Visualize, Data, Help, User, and a bell icon. On the left, a sidebar titled 'Tools' has a search bar containing 'trim'. Below it are buttons for 'Upload Data' and 'Show Sections'. A list of trim-related tools is provided, including 'Trim leading or trailing characters', 'Trim sequences', 'Trimmomatic flexible read trimming tool for Illumina NGS data', 'trimest Trim poly-A tails off EST sequences', 'trimseq Trim ambiguous bits off the ends of sequences', 'Trim.flows partition by barcode, trim to length, cull by length and mismatches', 'Trim.seqs Trim sequences - primers, barcodes, quality', 'Trim putative adapter sequence', 'Trim Galore! Quality and adapter trimmer of reads', 'TrimN', and 'seqtk trimfq trim FASTQ using the seqtk command-line tool'. The main content area displays the 'Trimmomatic flexible read trimming tool for Illumina NGS data' tool details. It includes a description of phred+64 encoding, options for 'Output trimlog file?' (set to 'No'), 'Output trimomatic log messages?' (set to 'No'), and 'Additional Options' (set to 'No'). There is also an 'Email notification' section (set to 'No') which sends an email when the job completes. At the bottom are 'Run Tool' and 'Help' buttons.

Fig 5: Run the tool

This screenshot shows the same Galaxy Europe interface as Fig 5, but with several sections highlighted in red boxes. The 'Tool Parameters' section, which includes fields for 'Single-end or paired-end reads?' (set to 'Paired-end (two separate input files)'), 'Input FASTQ file (R1/first of pair)' (set to '4; (unavailable) BWA-Sample-2'), and 'Input FASTQ file (R2/second of pair)' (set to '5; (unavailable) BWA-Sample-1'), is highlighted. The 'Perform initial ILLUMINACLIP step?' section (set to 'no') is also highlighted. To the right, the 'History' panel is shown, displaying a list of jobs: 'Trimming-Trimmomatic' (using 2%), '6: Trimmomatic on trimming seq (R2 unpaired)', '5: Trimmomatic on trimming seq (R1 unpaired)', '4: Trimmomatic on trimming seq (R2 paired)', '3: Trimmomatic on trimming seq (R1 paired)', '2: trimming seq', and '1: trimming seq'. The last four items are highlighted with a red border.

Fig 6: Results appears in 4 different sections

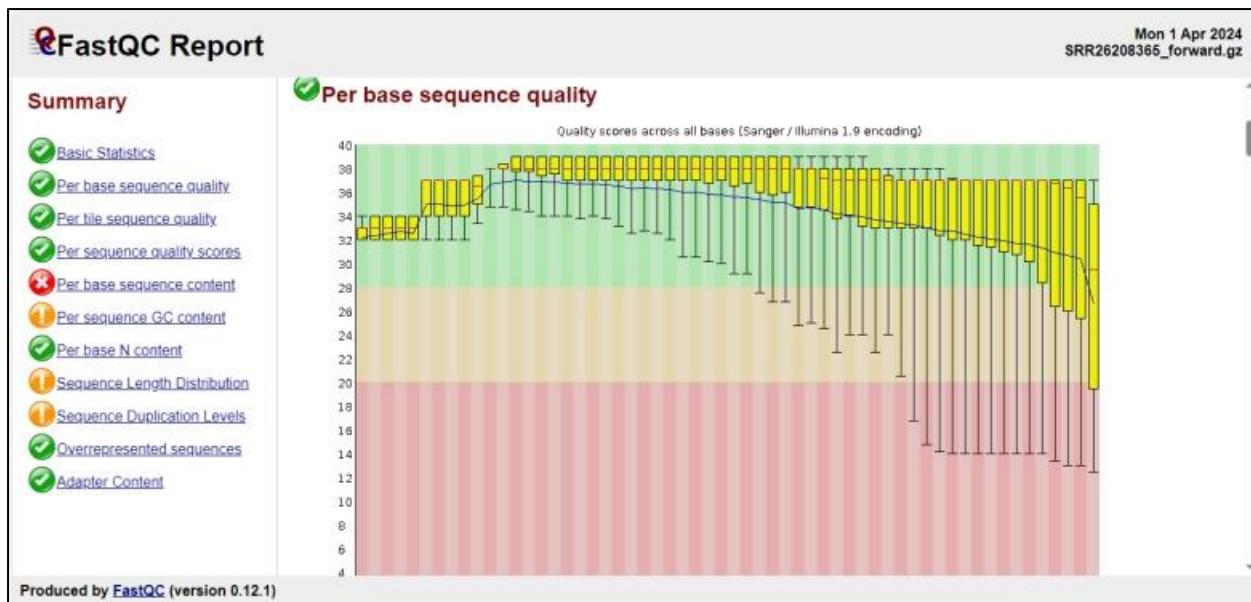


Fig 7: Pre-Trimming data

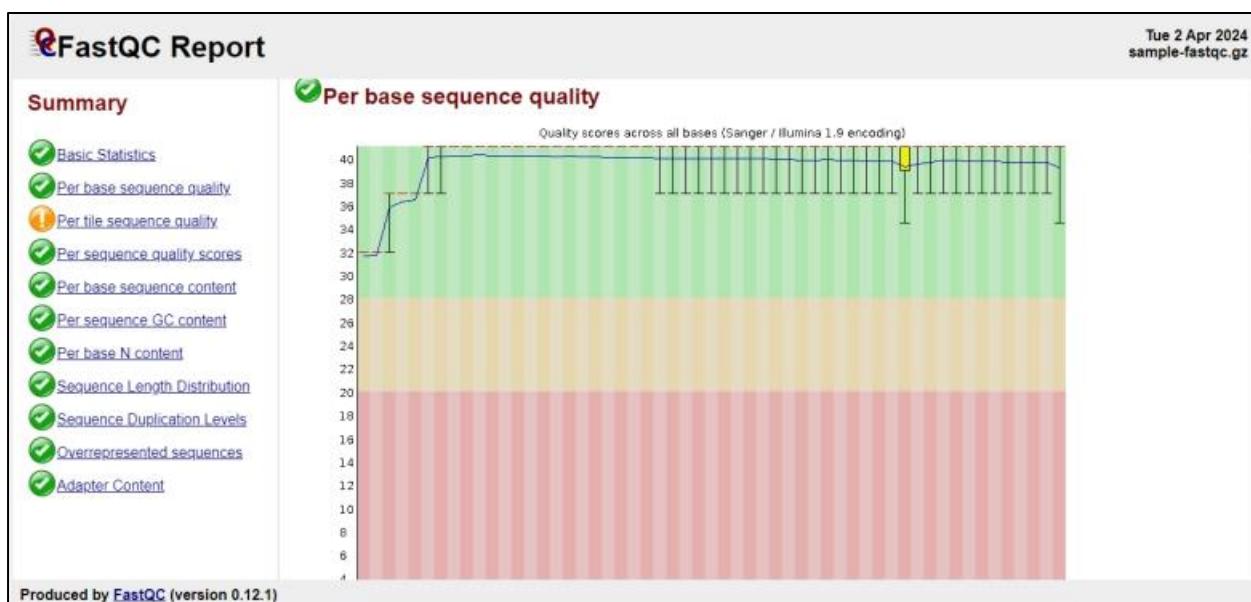


Fig 7.1: Post-Trimming data

RESULTS:

The trimmomatic produces trimmed FASTQ files containing sequences that pass the specified quality thresholds by generates quality control metrics such as the number of reads processed, the number of reads kept after trimming, the average quality scores before and after trimming, and the percentage of reads trimmed. Thus it provide insights into the quality of the sequencing data before and after trimming. Higher retention of reads and improvement in quality scores indicate effective trimming. The generated in form of .bam will be further used for assembly studies.

CONCLUSION:

Trimmomatic tool has effectively removed low-quality regions and adapters for the query sequence.

REFERENCES:

1. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* (Oxford, England), 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
 2. https://www.researchgate.net/publication/261327470_Trimmomatic_A_Flexible_Trimmer_for_Illumina_Sequence_Data
 3. Hamzić, E. (2024, January 3). All You Need To Know About Trimmomatic & NGS Data Trimming. BioComputiX. <https://www.biocomputix.com/post/trimming/ngs-data-trimmomatic>
-

DATE: 04/04/2024

WEBLEM 9(C)
NGS- FASTQC
(URL: <https://usegalaxy.eu/>)

AIM:

To assess the quality of raw sequencing data and identify potential issues or biases using FASTQC for quality control.

INTRODUCTION:

Quality control (QC) is critical in next-generation sequencing (NGS) to ensure the reliability and integrity of the vast amounts of sequencing data generated. QC measures are essential to identify and address any anomalies, errors, and biases that may arise during library preparation, sequencing, and data analysis. Not only do QC measures enhance the accuracy of biological interpretations, but they also contribute to the cost-effectiveness, reproducibility, and comparability of results. Furthermore, QC enables the detection and quantification of biases and artifacts, optimization of experimental parameters, and ultimately the generation of high-quality data essential for advancing biological research, clinical diagnostics, and various applications of NGS technology.

FastQC is a highly versatile quality control tool designed specifically for analyzing sequencing data, particularly FastQ files, generated by modern high-throughput sequencers. Its primary function is to identify potential issues or biases in the raw data before downstream analysis. Unlike many sequencer-generated QC reports that focus solely on issues originating from the sequencer itself, FastQC aims to detect problems originating from both the sequencer and the initial library material. It offers two operational modes: an interactive standalone application for analyzing small numbers of files and a non-interactive mode suitable for processing large numbers of files within larger analysis pipelines.

One of FastQC's fundamental operations is opening and analyzing sequence files. Users can open files interactively, and newly opened files appear as tabs, allowing for easy navigation and analysis. FastQC supports various file formats, including FastQ, Casava FastQ, ColorSpace FastQ, SAM, BAM, and GZip compressed FastQ. It also provides options for users to manually specify the file format if needed. Upon analysis, FastQC generates a summary report indicating whether the results are normal, slightly abnormal, or very unusual, enabling users to quickly identify potential issues in their data.

METHODOLOGY:

1. Open the Galaxy Server homepage and search for FastQC tool.
2. Use the result file of trimmomatic as the input for FastQC analysis.
3. Set the tool parameters and submit the file.
4. 4 result files are generated (for paired end sequence) out of which 2 are of raw data and 2 of webpage.

- Open the webpage in which FastQC report is generated.
- Interpret the results accordingly.

OBSERVATIONS:

The screenshot shows the Galaxy Europe homepage. On the left, there's a sidebar with a search bar and a 'Tools' dropdown menu. Under 'Tools', sections include 'GENERAL TEXT TOOLS' (Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group) and 'GENOMIC FILE MANIPULATION' (Convert Formats, FASTA/FASTQ, Quality Control, SAM/BAM, BED). A central orange box contains a 'Cluster Maintenance April 8 and 9' notice. Below it is a 'Project Highlight' section for 'The Vertebrate Genomes Project in Galaxy' featuring a 'VERTEBRATE' logo. The right side shows a 'History' panel with several dataset entries, including 'Quality check-FastQC' and 'FastQC' runs.

Fig 1: Galaxy Server Homepage

This screenshot shows the Galaxy Europe homepage with the search bar highlighted and the term 'FastQC' typed into it. The rest of the interface is identical to Fig 1, including the cluster maintenance notice and the history panel on the right.

Fig 2: Searching for FastQC tool on Galaxy server

3a

3b

Fig 3a & 3b: Select Tool Parameters in the FastQC and submit

Fig 4: Sequence result for the forward strand

Fig 5: Sequence result of reverse strand

The screenshot shows the Galaxy Europe interface with the following details:

- Tools:** FASTQC
- Workflow:** Workflow tab selected.
- Data:** Data tab selected.
- Help:** Help tab selected.
- User:** User tab selected.
- History:** Shows a list of datasets:
 - 6: FastQC on data 2: RawData (a)
 - 5: FastQC on data 2: Webpag (e)
 - 4: FastQC on data 1: RawData (d)
 - 3: FastQC on data 1: Webpag (e)
 - 2: sample-fastqc (d)
 - 1: sample-fastqc (d)

FastQC Read Quality reports

ab1 to FASTQ converter

Make.fastq Convert fasta and quality to fastq

Tabular to FASTQ converter

Tabular to FASTQ converter

WORKFLOWS

All workflows

FASTQC Read Quality reports

ab1 to FASTQ converter

Make.fastq Convert fasta and quality to fastq

Tabular to FASTQ converter

Tabular to FASTQ converter

WORKFLOWS

All workflows

FASTQC 0.12.1

>>Basic Statistics pass

#Measure Value

Filename sample-fastqc.gz

File type Conventional base calls

Encoding Sanger / Illumina 1.9

Total Sequences 10602766

Total Bases 1 Gbp

Sequences flagged as poor quality 0

Sequence length 101

%GC 49

>>END_MODULE

>>Per base sequence quality pass

#Base	Mean	Median	Lower Quartile	Upper Quartile	10th Percentile	90th Percentile
1	31.6954261744641597	32.0	32.0	32.0	32.0	32.0
2	31.71053078413688	32.0	32.0	32.0	32.0	32.0
3	35.83380478263879	37.0	37.0	37.0	32.0	37.0
4	36.326352199039384	37.0	37.0	37.0	37.0	37.0
5	36.47943819565574	37.0	37.0	37.0	37.0	37.0
6	40.1280265923062	41.0	41.0	41.0	37.0	41.0
7	40.194028408907235	41.0	41.0	41.0	37.0	41.0
8	40.2894277172579	41.0	41.0	41.0	41.0	41.0
9	40.263644977169164	41.0	41.0	41.0	41.0	41.0
10-11	40.30620439043925	41.0	41.0	41.0	41.0	41.0
12-13	40.279906064134586	41.0	41.0	41.0	41.0	41.0
14-15	40.27795968082291	41.0	41.0	41.0	41.0	41.0
16-17	40.28560410557019	41.0	41.0	41.0	41.0	41.0
18-19	40.260026015852844	41.0	41.0	41.0	41.0	41.0
20-21	40.2313253356106	41.0	41.0	41.0	41.0	41.0
22-23	40.205913107956924	41.0	41.0	41.0	41.0	41.0
24-25	40.22087307217758	41.0	41.0	41.0	41.0	41.0
26-27	40.175660766256655	41.0	41.0	41.0	41.0	41.0
28-29	40.1692140752675236	41.0	41.0	41.0	41.0	41.0
30-31	40.1692056516196	41.0	41.0	41.0	41.0	41.0
32-33	40.12707564614743	41.0	41.0	41.0	41.0	41.0
34-35	40.09569304339888	41.0	41.0	41.0	41.0	41.0
36-37	40.10155281557211	41.0	41.0	41.0	41.0	41.0

Fig 6: Raw data file of forward strand

The screenshot shows the Galaxy Europe interface with the following details:

- Tools:** FASTQC
- Workflow:** Workflow tab selected.
- Data:** Data tab selected.
- Help:** Help tab selected.
- User:** User tab selected.
- History:** Shows a list of datasets:
 - 6: FastQC on data 2: RawData (a)
 - 5: FastQC on data 2: Webpag (e)
 - 4: FastQC on data 1: RawData (d)
 - 3: FastQC on data 1: Webpag (e)
 - 2: sample-fastqc (d)
 - 1: sample-fastqc (d)

FastQC Read Quality reports

ab1 to FASTQ converter

Make.fastq Convert fasta and quality to fastq

Tabular to FASTQ converter

Tabular to FASTQ converter

WORKFLOWS

All workflows

FASTQC 0.12.1

>>Basic Statistics pass

#Measure Value

Filename sample-fastqc.gz

File type Conventional base calls

Encoding Sanger / Illumina 1.9

Total Sequences 10602766

Total Bases 1 Gbp

Sequences flagged as poor quality 0

Sequence length 101

%GC 49

>>END_MODULE

>>Per base sequence quality pass

#Base	Mean	Median	Lower Quartile	Upper Quartile	10th Percentile	90th Percentile
1	31.166752336135684	32.0	32.0	32.0	32.0	32.0
2	31.11097396604565	32.0	32.0	32.0	32.0	32.0
3	34.63101534071392	37.0	32.0	37.0	32.0	37.0
4	35.248595252605731	37.0	37.0	37.0	32.0	37.0
5	34.746840305633455	37.0	37.0	37.0	27.0	37.0
6	38.837276423906744	41.0	37.0	41.0	37.0	41.0
7	37.852960821732744	41.0	37.0	41.0	32.0	41.0
8	38.80500503359218	41.0	41.0	41.0	37.0	41.0
9	39.368580047885615	41.0	41.0	41.0	37.0	41.0
10-11	39.59799066924291	41.0	41.0	41.0	37.0	41.0
12-13	39.6517968047206	41.0	41.0	41.0	37.0	41.0
14-15	39.72906654735189	41.0	41.0	41.0	37.0	41.0
16-17	39.6765132528945	41.0	41.0	41.0	37.0	41.0
18-19	39.6869379423364	41.0	41.0	41.0	37.0	41.0
20-21	39.71846437133976	41.0	41.0	41.0	37.0	41.0
22-23	39.7237847652207	41.0	41.0	41.0	37.0	41.0
24-25	39.70545641827248	41.0	41.0	41.0	37.0	41.0
26-27	39.698604555453279	41.0	41.0	41.0	37.0	41.0
28-29	39.694385219856784	41.0	41.0	41.0	37.0	41.0
30-31	39.634118493230915	41.0	41.0	41.0	37.0	41.0
32-33	39.68006023145281	41.0	41.0	41.0	37.0	41.0
34-35	39.655025113258176	41.0	41.0	41.0	37.0	41.0
36-37	39.62610320710756	41.0	41.0	41.0	37.0	41.0

Fig 7: Raw data file of reverse strand

The results analysis of FastQC report (forward and reverse strand) is based on 11 analysis modules :

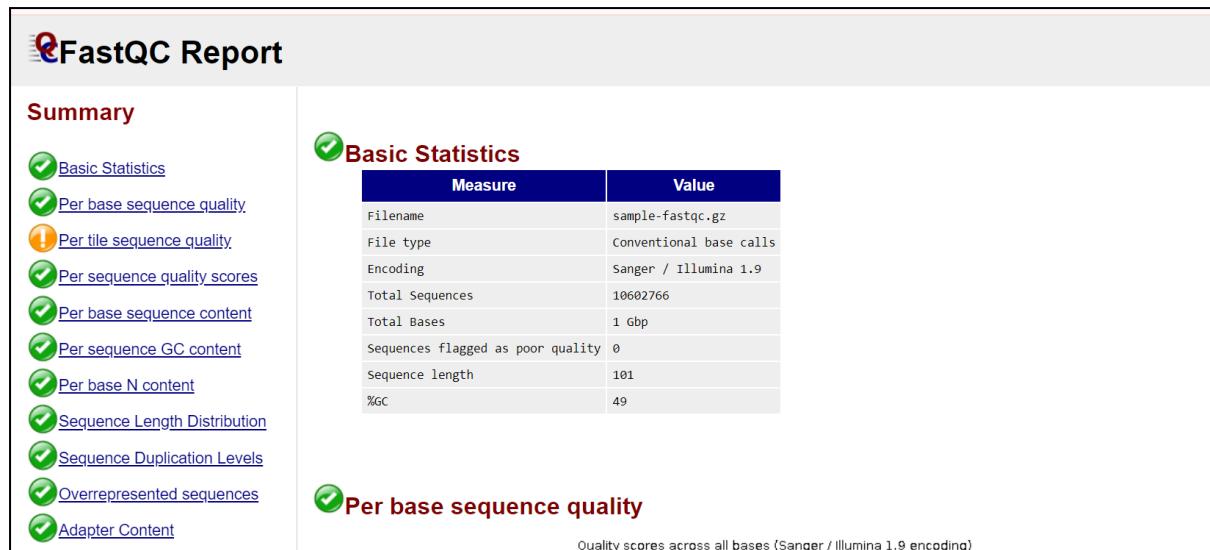


Fig 8: Basic statistics results

i) **Basic Statistics:** Provides composition statistics for the analyzed file, including the filename, file type, encoding, total sequences, filtered sequences, sequence length, and % GC. No warnings or failures are raised in this module.

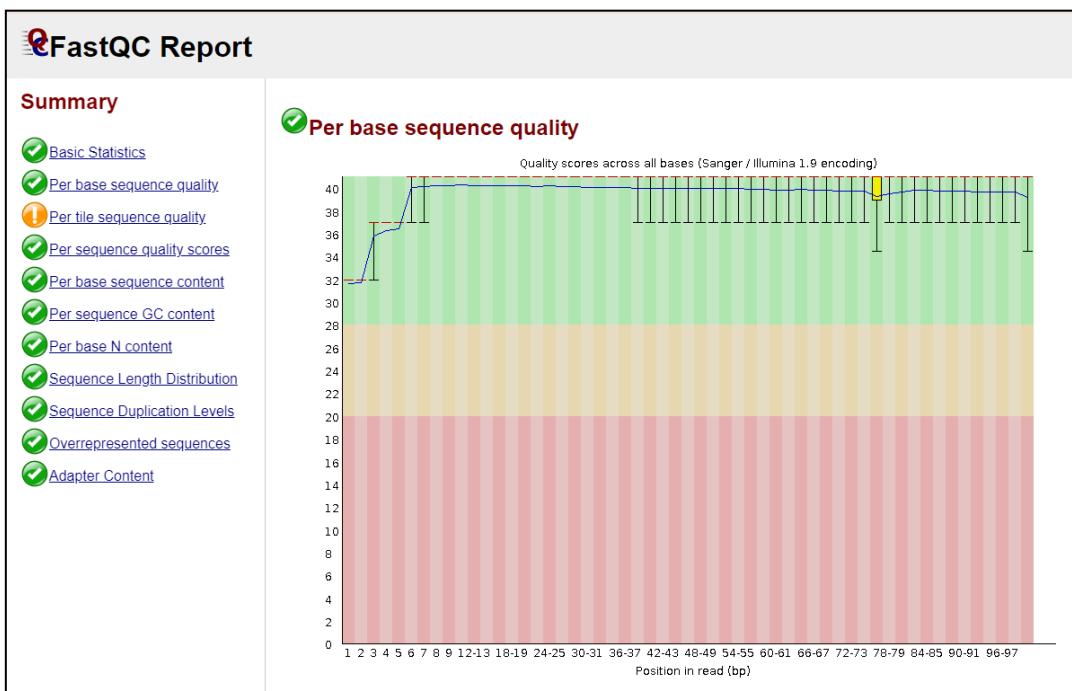


Fig 9: Per base sequence quality

ii) **Per Base Sequence Quality:** Presents a Box Whisker plot for the range of quality values across all bases at each position in the FastQ file. The initial position are of better quality which seems to be decreasing at the end of the read but still lies in the green zone.

Warning: Lower quartile - less than 10 or median for any base is less than 25.

Failure: Lower quartile - less than 5 or median for any base is less than 20.

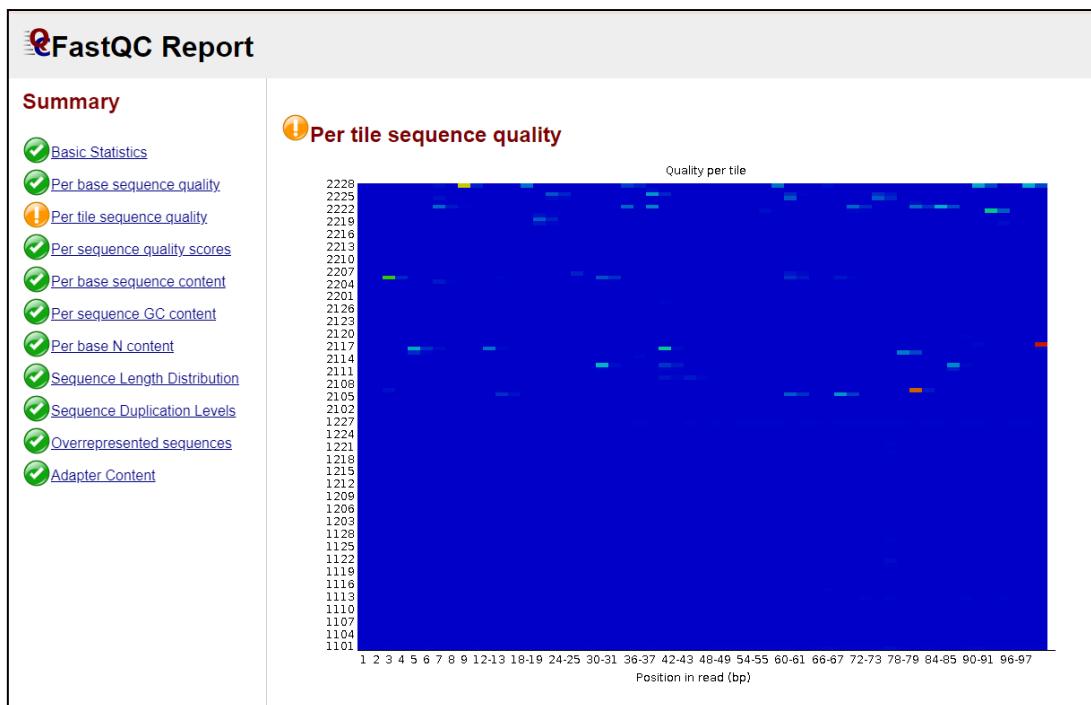


Fig 10: Per tile sequence quality

iii) Per Tile Sequence Quality: Evaluates sequencing quality across individual tiles, issuing a warning for excessive variability and passes if quality remains consistent, without specific failure criteria. Here marked tiles shows poor quality of reads or deviation from average quality of the reads.

Warning: Mean Phred score > 2 and less than the mean (base across all tiles).

Failure: Mean Phred score > 5 and less than the mean (base across all tiles).

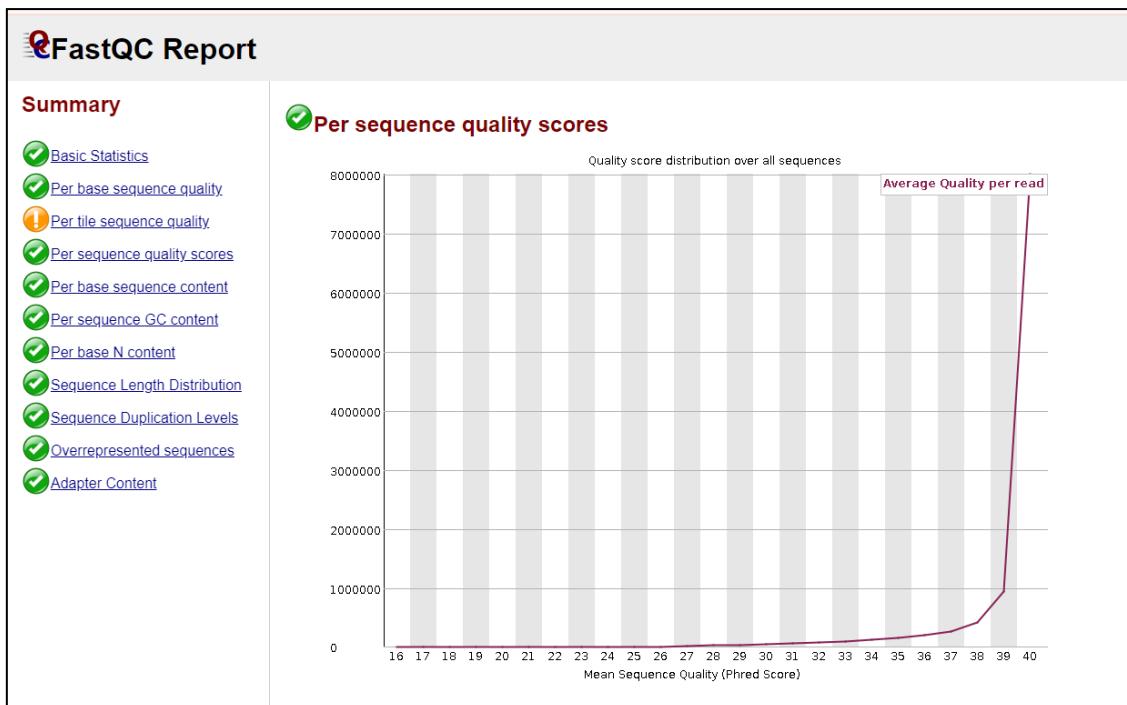


Fig 11: Per sequence quality score

iv) Per Sequence Quality Scores: Allows the assessment of universally low-quality values across a subset of sequences. Here the graph shows stable quality score distribution for overall sequences.

Warning: Mean quality < 27 - this equates to a 0.2% error rate.

Failure: Mean quality < 20 - this equates to a 1% error rate.

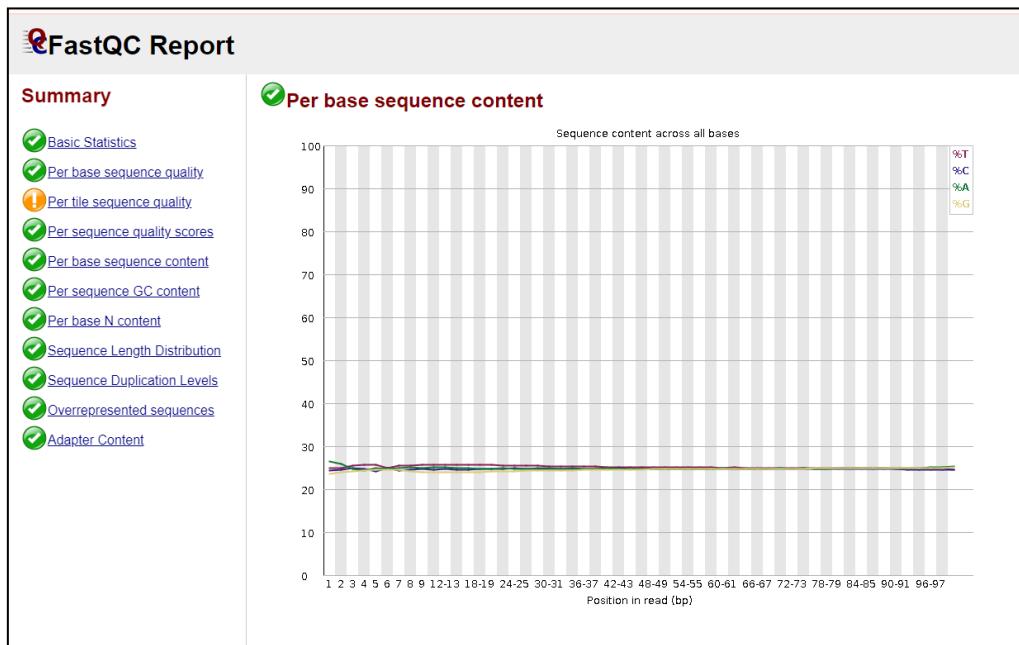


Fig 12: Per base sequence content

v) **Per Base Sequence Content:** Plots the proportion of each base position in a file for which each of the four normal DNA bases has been called. Here in the start (1-12) some unparallel lines are seen indicating difference in the proportion. Further the proportion of each base position is stable and parallel lines are seen indicating no difference in the base proportion.

Warning: Difference between A and T, or G and C is greater than 10% in any position.

Failure: Difference between A and T, or G and C is greater than 20% in any position.

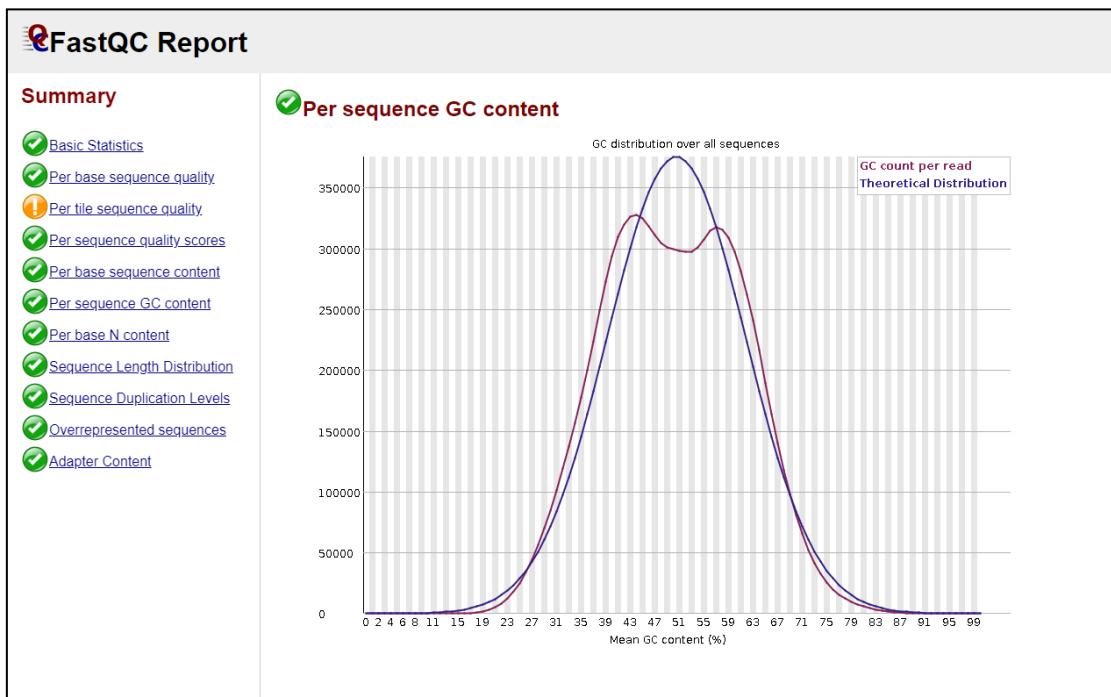


Fig 13: Per sequence GC content

vi) **Per Sequence GC Content:** Measures the GC content across the whole length of each sequence and compares it to a modeled normal distribution. Here the GC content across the whole length of sequence is upto the set theoretical value with little difference at the peak.

Warning: Sum of the deviations from the normal distribution > 15% of the reads.

Failure: Sum of the deviations from the normal distribution > 30% of the reads.

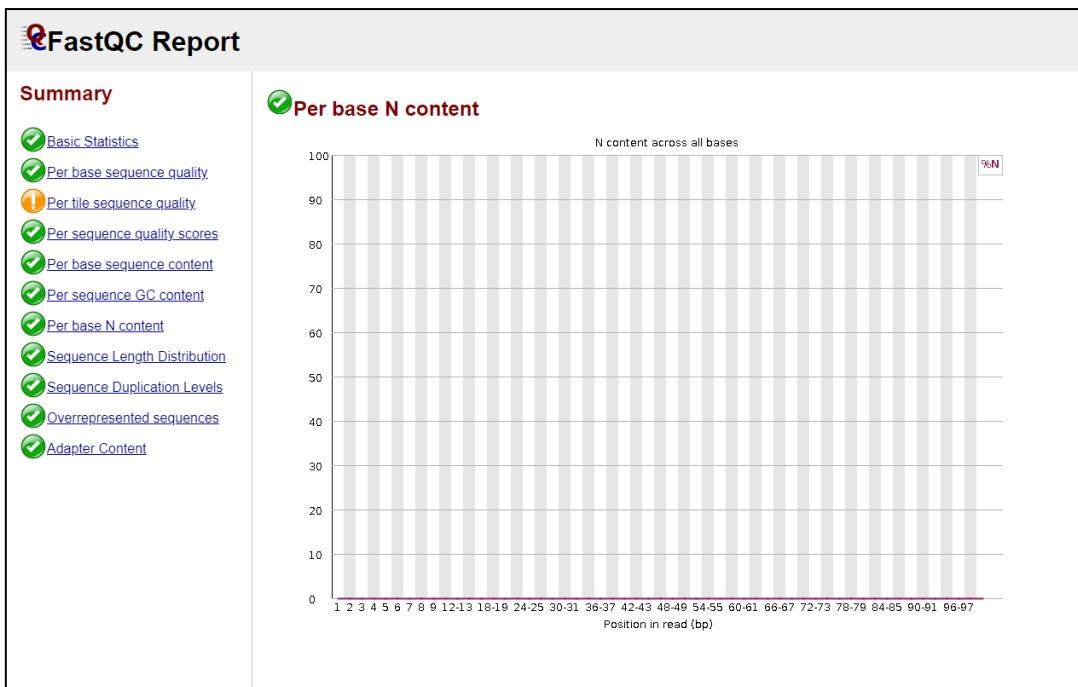


Fig 14: Per base N content

vii) Per Base N Content: Plots the percentage of base calls at each position for which an N was called. Here no presence of N content anywhere in the sequence indicating the base were properly called with confidence during sequencing

Warning: Position shows an N content of >5%.

Failure: Position shows an N content of >20%.

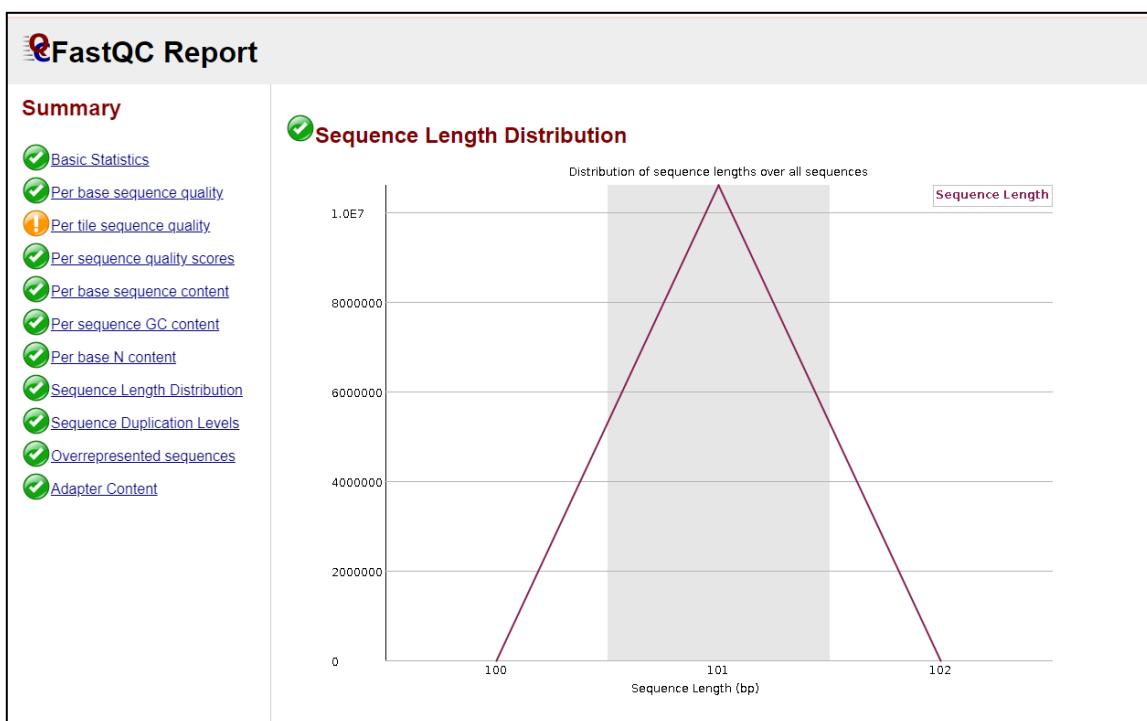


Fig 15: Sequence length distribution

viii) Sequence Length Distribution: Generates a graph showing the distribution of fragment sizes in the analyzed file. Graph shows constant sequence length (100bp-102bp) for overall sequences. Indicating proper fragmentation of the sequence.

Warning: if all sequences are not the same length.

Failure: if any of the sequences have zero length.

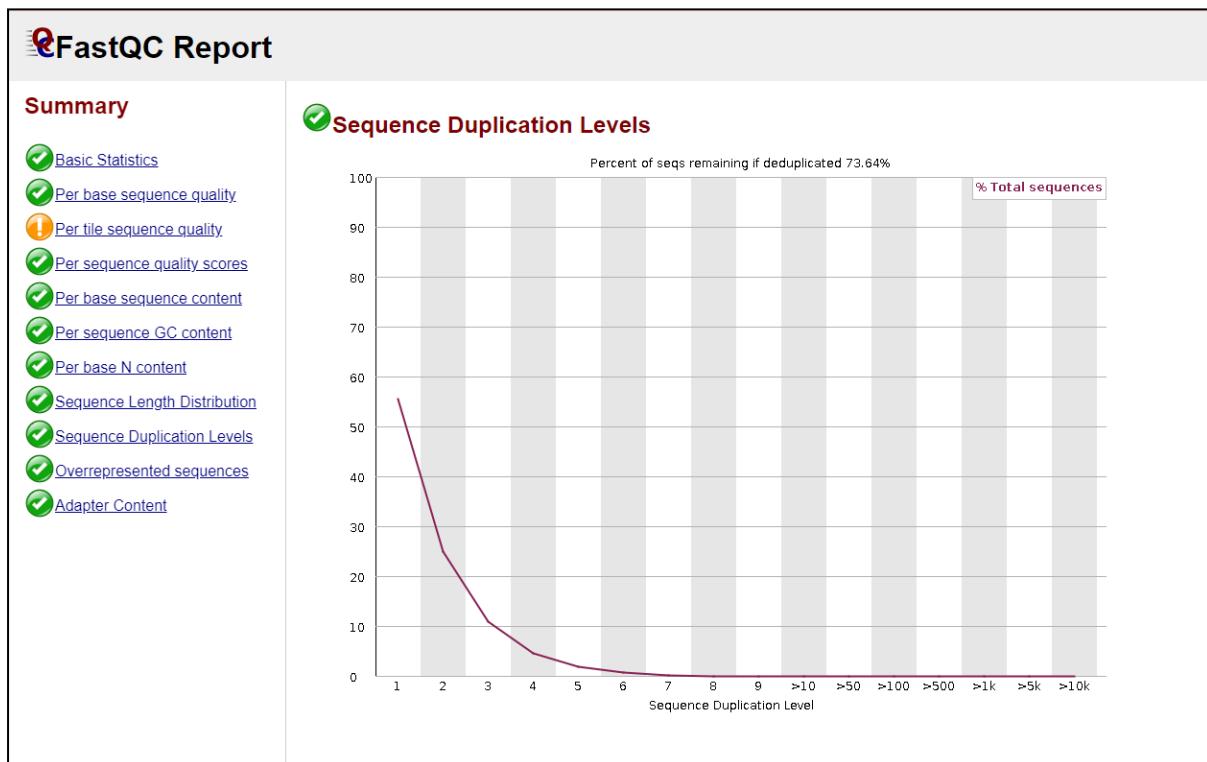


Fig 16: Sequence duplication levels

ix) Duplicate Sequences: Counts the degree of duplication for every sequence in the set. Shows some duplicate sequences but the duplication level is less than 10 which is acceptable.

Warning: if non-unique sequences make up more than 20% of the total.

Failure: if non-unique sequences make up more than 50% of the total.

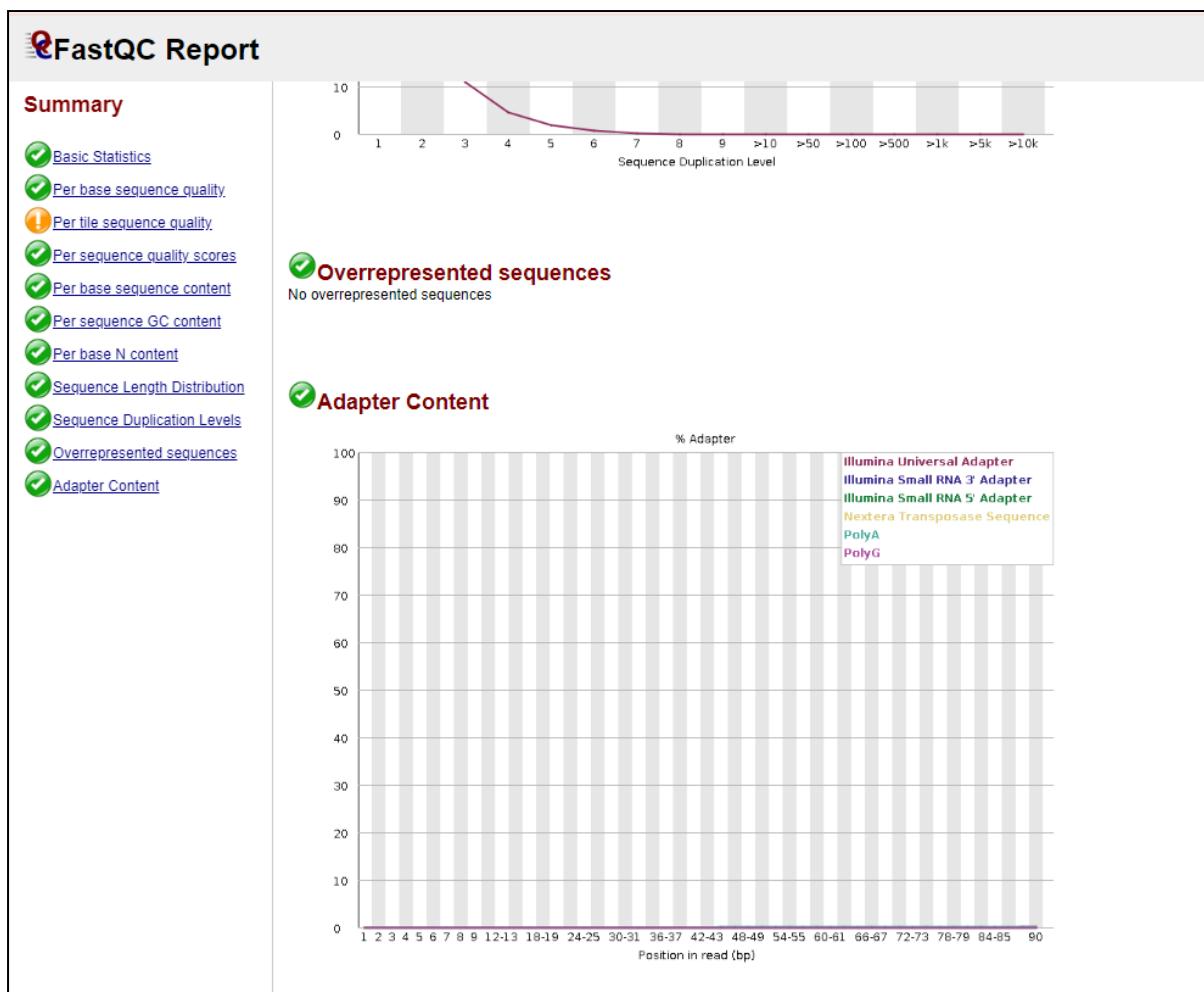


Fig 17: Overrepresented sequences and adapter content

x) Overrepresented Sequences: Lists all sequences representing more than 0.1% of the total. No overrepresented sequences were present.

Warning: sequence is found to represent more than 0.1% of the total.

Failure: sequence is found to represent more than 1% of the total.

xi) Adapter Content: A cumulative plot that shows the fraction of reads that contain a sequence library adapter at a given base position. Here there is no presence of adapter content in the sequence.

Warning: Any sequence is present in more than 5% of all reads.

Failure: Any sequence is present in more than 10% of all reads.

RESULTS:

The analysis in FastQC is performed by a series of analysis modules. The left hand side of the main interactive display or the top of the HTML report show a summary of the modules which were run, and a quick evaluation of whether the results of the module seem entirely normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross). The specific module details like the graph, warning and failures have been discussed above in detail.

CONCLUSION:

The quality of raw sequencing data was checked using FastQC tool. All the potential issues were looked into. FastQC has been done to get accurate biological interpretations and check whether our sequence is of good quality or not to do further process.

REFERENCES:

1. FastQC. (n.d.). FastQC. https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf
 2. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
 3. Index of /projects/fastqc/Help. (n.d.).
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>
 4. Sanger, J. M. a. V. C. (n.d.). The FASTQ format and quality control.
<https://slideplayer.com/slide/13772534/>
-

DATE: 10/04/2024

WEBLEM 9(D)
Bowtie Tool
(URL: <https://usegalaxy.org/>)

AIM:

To map reads against reference genome using bowtie 2 tool for the given query.

INTRODUCTION:

Bowtie package is used to align short sequencing reads, such as those output by second-generation sequencing instruments. It also includes protocols for building a genome index and calling consensus sequences from Bowtie alignments using SAMtools. The Bowtie package enables ultrafast and memory-efficient alignment of large sets of sequencing reads to a reference sequence, such as the human genome. The package contains tools for building indexes of reference genomes and for aligning short reads using the index as a guide. This is the first step of many comparative genomics workflows, including variant detection and digital gene expression. In what follows, the term read refers to a short DNA sequence, typically as output by a sequencing instrument. A read may be accompanied by a corresponding string of quality values, where each value estimates the probability that the corresponding base was miscalled by the instrument software. The term subject sequence refers to the true sequence of the sample(s) from which the reads were drawn. The term reference sequence or reference genome refers to the sequence to which the subject is to be compared.

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index (based on the Burrows-Wheeler Transform or BWT) to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 gigabytes of RAM. Bowtie 2 supports gapped, local, and paired-end alignment modes. Multiple processors can be used simultaneously to achieve greater alignment speed. Bowtie 2 outputs alignments in SAM format, enabling interoperation with a large number of other tools (e.g. SAMtools, GATK) that use SAM. Bowtie 2 is distributed under the GPLv3 license, and it runs on the command line under Windows, Mac OS X and Linux and BSD. Bowtie 2 is often the first step in pipelines for comparative genomics, including for variation calling, ChIP-seq, RNA-seq, BS-seq. Bowtie 2 and Bowtie (also called "Bowtie 1" here) are also tightly integrated into many other tools.

METHODOLOGY:

1. Open homepage of Galaxy Tool.
2. Search Bowtie in search tool box.
3. Set different parameters of the tool.
4. Set the different operations.
5. Run the tool and interpret the result which are come in four different sections.

OBSERVATIONS:

The screenshot shows the Galaxy Europe interface. On the left, a sidebar lists various tools under categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. A central orange box contains a maintenance notice for April 8 and 9, stating that the website will be reachable, jobs can remain in 'grey' state, and no new files can be uploaded. Below this is a 'Paper Alert!' graphic. To the right is a 'History' panel showing a list of recent datasets, including SPAdes assembly steps and Bowtie2 alignments.

Fig 1: Homepage of Galaxy Tool

This screenshot shows the Galaxy Europe search interface. The search bar at the top has 'bowt' typed into it. The results list includes 'Bowtie2 - map reads against reference genome' and other related tools like 'Map with Bowtie for SOLiD'. The 'Tool Parameters' section for Bowtie2 is expanded, showing options for single or paired libraries, interleaved FASTQ files (with samples 5: BWA-Sample-1 and 4: BWA-Sample-2 selected), and various output write options. The right side of the screen shows the same history panel as Fig 1.

Fig 2: Select the Bowtie option in search tool box

Tool Parameters

Is this single or paired library
Paired-end data from single interleaved dataset

Interleaved FASTQ file *
5: BWA-Sample-1 x 4: BWA-Sample-2 x

This is a batch mode input field. Individual jobs will be triggered for each dataset.
Must be of datatype "fastqsanger" or "fasta". --interleaved

Write unaligned reads (in fastq format) to separate file(s)
No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)
No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Do you want to set paired-end options?
No

History

- Reference genome mapping using bowtie
- alignment file reading via BAM file De novo assembly using Spades
- 2.76 GB
- 11: SPAdes on data 4 and data 5: Scaffolds
- 10: SPAdes on data 4 and data 5: Contigs
- 9: SPAdes on data 4 and data 5: Assembly graph with scaffolds
- 8: SPAdes on data 4 and data 5: Assembly graph
- 7: Bowtie2 on data 4 and data 5: alignments

Fig 3: Set the tool parameters

Tool Parameters

See "Alignment Options" section of Help below for information

Will you select a reference genome from your history or use a built-in index?
Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below

Select reference genome *
A. mellifera genome (apiMe3, Baylor HGSC Amel_3.0)

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?
Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

History

- Reference genome mapping using bowtie
- alignment file reading via BAM file De novo assembly using Spades
- 2.76 GB
- 11: SPAdes on data 4 and data 5: Scaffolds
- 10: SPAdes on data 4 and data 5: Contigs
- 9: SPAdes on data 4 and data 5: Assembly graph with scaffolds
- 8: SPAdes on data 4 and data 5: Assembly graph
- 7: Bowtie2 on data 4 and data 5: alignments

Fig 3.1: Set the tool parameters

Bowtie2 - map reads against reference genome (Galaxy Version 2.5.3+galaxy9)

Select analysis mode

1: Default setting only

Do you want to use presets? *

- No, just use defaults
- Very fast end-to-end (--very-fast)
- Fast end-to-end (--fast)
- Sensitive end-to-end (--sensitive)
- Very sensitive end-to-end (--very-sensitive)
- Very fast local (--very-fast-local)
- Fast local (--fast-local)
- Sensitive local (--sensitive-local)
- Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

Do you want to tweak SAM/BAM Options?

No

See "Output Options" section of Help below for Information

Save the bowtie2 mapping statistics to the history

Yes

History

Reference genome mapping using bowtie

alignment file reading via BAM file De novo assembly using Spades

2.76 GB 10 3

11: SPAdes on data 4 and data 5: Scaffolds

10: SPAdes on data 4 and data 5: Contigs

9: SPAdes on data 4 and data 5: Assembly graph with scaffolds

8: SPAdes on data 4 and data 5: Assembly graph

7: Bowtie2 on data 4 and data 5: alignments

Fig 4: Set the analysis mode

Bowtie2 - map reads against reference genome (Galaxy Version 2.5.3+galaxy9)

Save the bowtie2 mapping statistics to the history

No

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

Help

Bowtie2 Overview

Bowtie2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 supports gapped, local, and paired-end alignment modes. Galaxy wrapper for Bowtie 2 outputs alignments in BAM format, enabling interoperation with a large number of other tools available at this site. Majority of information in this page is derived from an excellent [Bowtie2 manual](#) written by Ben Langmead.

Selecting reference genomes for Bowtie2

Galaxy wrapper for Bowtie2 allows you select between precomputed and user-defined indices for reference genomes using **Will you select a reference genome from your history or use a built-in index?** flag. This flag has two options:

History

Reference genome mapping using bowtie

alignment file reading via BAM file De novo assembly using Spades

2.76 GB 10 3

11: SPAdes on data 4 and data 5: Scaffolds

10: SPAdes on data 4 and data 5: Contigs

9: SPAdes on data 4 and data 5: Assembly graph with scaffolds

8: SPAdes on data 4 and data 5: Assembly graph

7: Bowtie2 on data 4 and data 5: alignments

Fig 5: Run the Tool

Started tool **Bowtie2** and successfully added 2 jobs to the queue.

It produces 2 outputs:

- 14: Bowtie2 on data 4: alignments
- 15: Bowtie2 on data 5: alignments

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

PHD Comics

Random

8:00AM 9:00AM 10:00AM

Reference genome mapping using bowtie

alignment file reading via BAM file De novo assembly using Spades

2.76 GB 12 3

ata 5: Assembly graph wrt scaffolds

8: SPAdes on data 4 and d 5: Assembly graph

7: Bowtie2 on data 4 and data 5: alignments

6: Map with BWA-MEM on data 5 (mapped reads in BAM format)

5: BWA-Sample-1

4: BWA-Sample-2

Fig 6: Results shown in four different options

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT
@HD VN:1.5 SO:coordinate											
@SQ SN:chr1 LN:248956422											
@SQ SN:chr1_KI270706v1_random LN:175055											
@SQ SN:chr1_KI270707v1_random LN:32032											
@SQ SN:chr1_KI270708v1_random LN:127682											
@SQ SN:chr1_KI270709v1_random LN:66860											
@SQ SN:chr1_KI270710v1_random LN:40176											
@SQ SN:chr1_KI270711v1_random LN:42210											
@SQ SN:chr1_KI270712v1_random LN:176043											
@SQ SN:chr1_KI270713v1_random LN:40745											
@SQ SN:chr1_KI270714v1_random LN:41717											
@SQ SN:chr1_GL383518v1_alt LN:182439											
@SQ SN:chr1_GL383519v1_alt LN:110268											
@SQ SN:chr1_GL383520v2_alt LN:366560											
@SQ SN:chr1_KI270759v1_alt LN:425601											
@SQ SN:chr1_KI270760v1_alt LN:109528											
@SQ SN:chr1_KI270761v1_alt LN:165834											

https://usegalaxy.eu/datasets/4838ba20a6d867655d026247ea5fc3b1/preview

Reference genome mapping using bowtie

alignment file reading via BAM file De novo assembly using Spades

2.76 GB 12 3

ata 5: Assembly graph wrt scaffolds

8: SPAdes on data 4 and data 5: Assembly graph

7: Bowtie2 on data 4 and data 5: alignments

6: Map with BWA-MEM on data 5 (mapped reads in BAM format)

5: BWA-Sample-1

4: BWA-Sample-2

Fig 7: Result Page

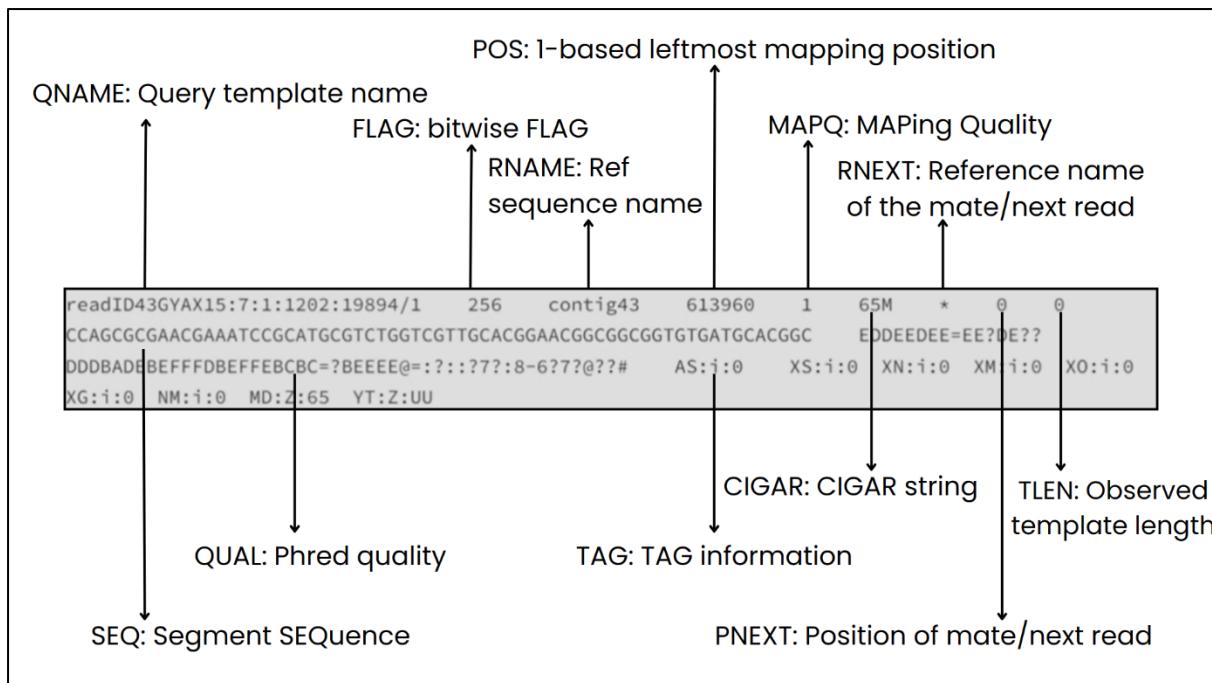


Fig 7.1 Detail result of the alignment

RESULTS:

Bowtie generates an alignment report summarizing the results of the alignment process. This report includes information such as the number of reads processed, the number of reads aligned to the reference genome, the alignment rate (percentage of reads aligned), and any potential issues encountered during the alignment process. In addition to the basic alignment report, Bowtie can generate more detailed statistics about the alignment process, such as the distribution of alignment quality scores, the distribution of alignment positions across the reference genome, and the frequency of mismatches or indels between the reads and the reference genome.

CONCLUSION:

For the given sequence, mapping was done with reference using Bowtie.

REFERENCES:

1. Trapnell C, et al. Nat. Biotechnol. 2010;28:511–515. [PMC free article] [PubMed] [Google Scholar]
2. Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). <https://doi.org/10.1038/nmeth.1923>
3. Bowtie 2: Manual. (n.d.). <https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>

DATE: 13/04/2024

WEBLEM 9(E)
VELVETH
(URL:<https://usegalaxy.org/>)

AIM:

To perform de novo assembly of sequencing reads into contigs using Velveth for genome assembly.

INTRODUCTION:

Velvet is software to perform DNA assembly from short reads by manipulating de Bruijn graphs. It is capable of forming long contigs (n50 of in excess of 150kb) from paired end short reads. It has several input parameters for controlling the structure of the de Bruijn graph and these must be set optimally to get the best assembly possible. Velvet can read Fasta, FastQ, sam or bam files. However, it ignores any quality scores and simply relies on sequencing depth to resolve errors. The Velvet Optimiser software performs many Velvet assemblies with various parameter sets and searches for the optimal assembly automatically. It is designed to handle short read data with high coverage and is suitable for assembling genomes of organisms that do not have a reference genome available. The algorithm implemented in Velvet is capable of manipulating de Bruijn graphs to remove sequencing errors and resolve repeats, allowing for the reconstruction of the original genome sequence.

In the context of the Galaxy platform, Velvet is available as a tool in the Galaxy platform, where it can be used as part of a larger bioinformatics workflow. The tool is particularly useful for assembling short DNA sequences and can be used to resolve repeats with the addition of long reads.

Galaxy is an open-source web-based platform for data analysis that provides a user-friendly interface for running a wide range of bioinformatics tools, including Velveth. By using Velveth in Galaxy, researchers can easily perform de novo genome assembly and integrate it with other analysis steps in their workflow.

To use Velveth in Galaxy, users can navigate to the tool's page in the Galaxy interface and provide the necessary input parameters, such as the location of the input reads and the desired k-mer size. Once the tool is executed, the resulting assembly can be further analyzed using other tools available in Galaxy.

Functions:

1. Velveth breaks down each read into k-mers of a specified length, where a k-mer is a subsequence of the read.
2. These k-mers and their reverse complements are then added to a hash table for categorization, with each k-mer stored once along with the number of times it appears.
3. This hashing process is crucial for organizing the reads and preparing them for the construction of the de Bruijn graph, which is essential for the subsequent steps in the assembly process.

Input formats:

1. Velveth can handle sequence files in formats such as fasta, fastq, raw, sam, and bam.
2. For paired-end reads, the assumption is that each read is paired with its mate read, facilitating the assembly process.

Usage in velveth assembly:

1. Velveth plays a fundamental role in the Velvet assembly process by preparing the input reads for further analysis and assembly.
2. It is a critical step in the de novo assembly of short read sequences, enabling the creation of high-quality unique contigs that form the basis of the reconstructed genome sequence.

METHODOLOGY:

1. Open homepage of Galaxy server.
2. In the search panel search VELVETH tool.
3. Select the required parameters and upload the dataset
 - a. “Hash Length”: 21
 - b. “Input Files”
 - i. Click on param-repeat “Input Files”
 - ii. “Choose the input type”: interleaved paired end
 - iii. “read type”: shortPaired reads
 - iv. param-files “Dataset”: Upload sample ‘Velveth sample’
4. The tool takes our reads and break them into k-mers.
5. Click on ‘Run Tool’.
6. Three output files will be generated.

OBSERVATIONS:

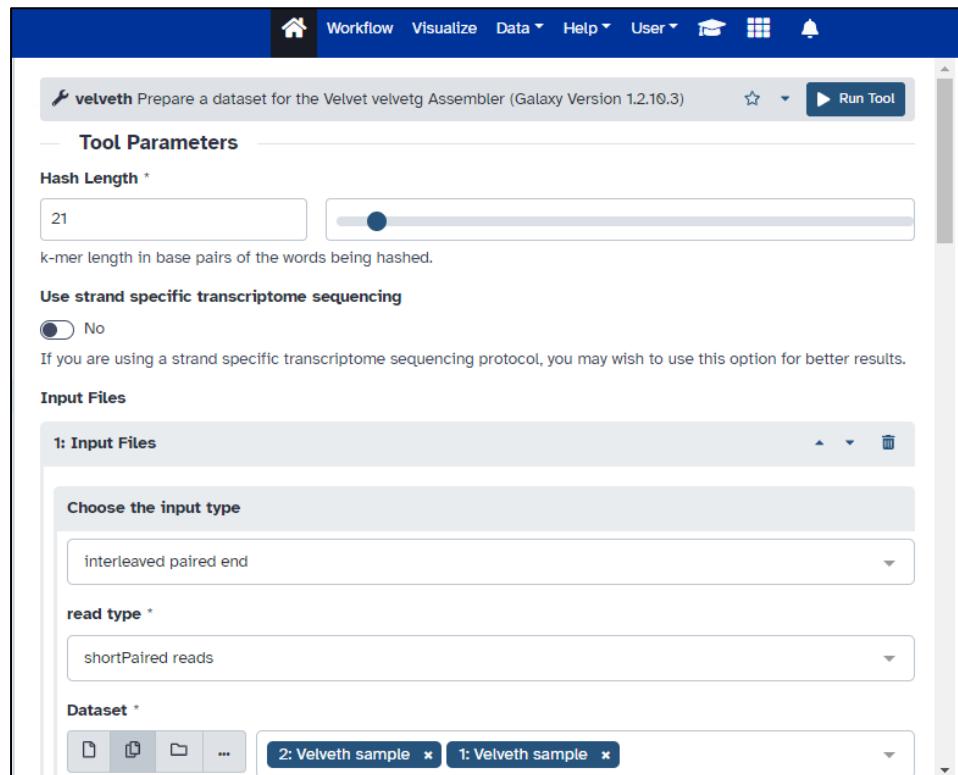


Fig 1.1: Select the parameters and upload the Dataset

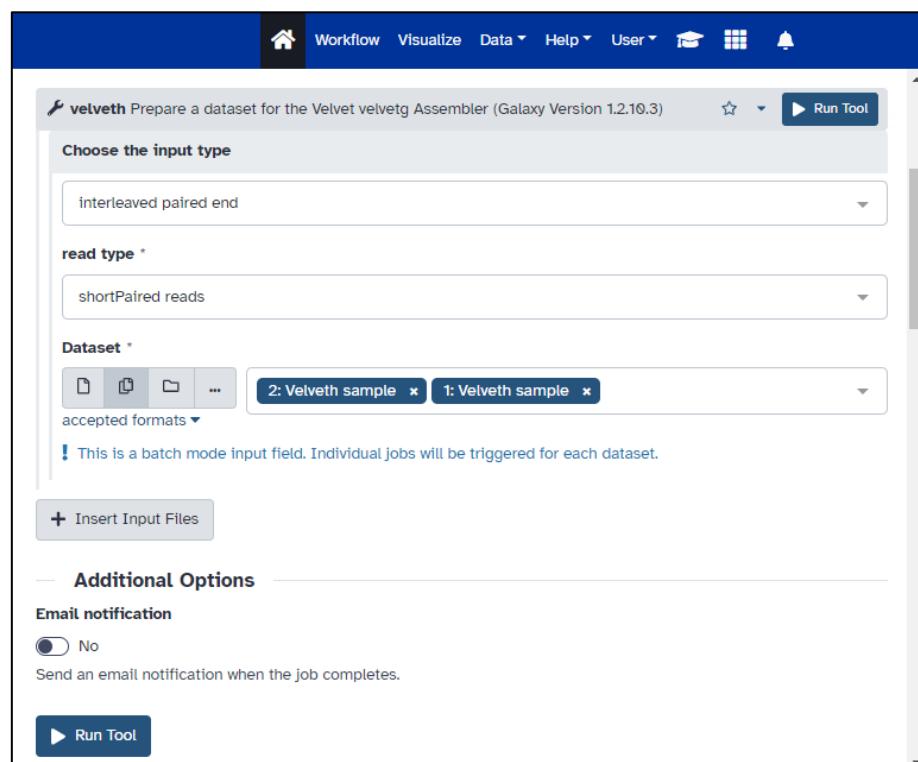


Fig 1.2: Click on ‘Run Tool’

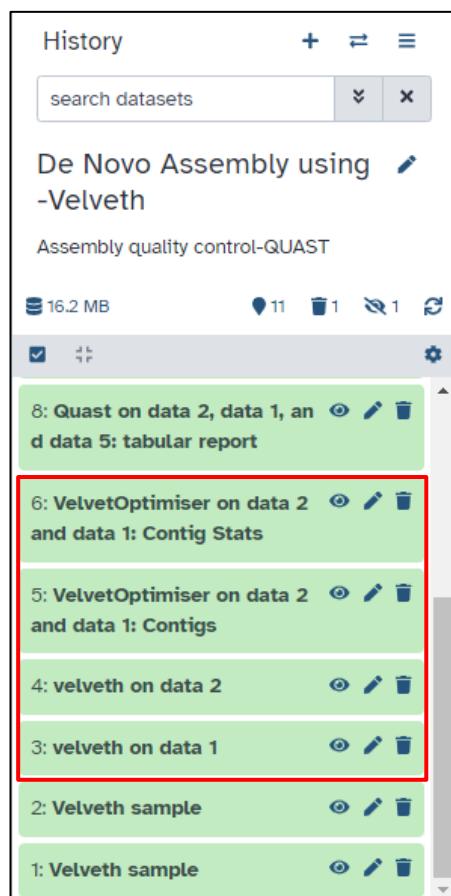


Fig 2: Output files (Sequence File, Contigs File, Contig Statistics File)

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11
1	16	2	0	0.000000	21.812500	21.812500	0.000000	0.000000	0.000000	0.000000
2	203	3	2	0.000000	54.620690	49.226601	0.000000	0.000000	0.000000	0.000000
3	3	2	1	0.000000	21.666667	21.666667	0.000000	0.000000	0.000000	0.000000
4	38	1	1	0.000000	15.921053	15.921053	0.000000	0.000000	0.000000	0.000000
5	14	1	2	0.000000	15.428571	15.428571	0.000000	0.000000	0.000000	0.000000
6	3	0	2	0.000000	14.333333	14.333333	0.000000	0.000000	0.000000	0.000000
7	121	0	0	0.000000	1.033058	1.033058	0.000000	0.000000	0.000000	0.000000
8	132236	1	1	0.000000	10.730338	4.428960	0.000000	0.000000	0.000000	0.000000
9	7	1	1	0.000000	64.714286	64.714286	0.000000	0.000000	0.000000	0.000000
10	5	1	2	0.000000	65.600000	65.600000	0.000000	0.000000	0.000000	0.000000
11	9	1	2	0.000000	66.666667	65.888889	0.000000	0.000000	0.000000	0.000000
12	5	1	2	0.000000	62.800000	61.000000	0.000000	0.000000	0.000000	0.000000
13	8	1	2	0.000000	63.625000	61.625000	0.000000	0.000000	0.000000	0.000000
14	21	1	2	0.000000	65.285714	62.904762	0.000000	0.000000	0.000000	0.000000
15	423	2	2	0.000000	55.411348	51.957447	0.000000	0.000000	0.000000	0.000000

Fig 3: Tabular file giving for each contig the k-mer lengths, k-mer coverages and other measures

Dataset Information

Number	6
Name	VelvetOptimiser on data 2 and data 1: Contig Stats
Created	Saturday Apr 13th 5:25:46 2024 GMT+5:30
Filesize	6.6 KB
Dbkey	?
Format	tabular
File contents	contents
History Content API ID	4838ba20a6d86765dc2aad7a500b7a72
History API ID	078cbff2ce6c6446
UUID	67853b15-f215-4d35-82c2-e05da8bf1961
Full Path	/data/dnb09/galaxy_db/files/6/7/8/dataset_67853b15-f215-4d35-82c2-e05da8bf1961.dat
Originally Created From a File	stats.txt
Named	

Tool Parameters

Input Parameter	Value
Start k-mer size	31
End k-mer size	191
K-mer search step size	2
Input file type	Fastq
Single or paired end reads	paired

Select first set of reads: **1: Velvet sample**

History

De Novo Assembly using **-Velveth**

Assembly quality control-QUAST

16.2 MB 11 1 1 16.2 MB

and data 1: Contig Stats

Add Tags

65 lines format tabular, database ?

Logfile name: 13-04-2024-01-56-14_Logfile.txt

1 2 3 4 5 6 7

ID	Length	Coverage	Min cov	Max cov	Avg cov	SD cov
1	16	2	0	0.00000	21.812500	21.812500
2	283	3	2	0.00000	54.620600	49.226601
3	3	2	1	0.00000	21.666667	21.666667
4	38	1	1	0.00000	15.921053	15.921053

Fig 4: Dataset Information of the Contig Statistic File

Job Information

Galaxy Tool ID	toolshed.g2.bx.psu.edu/repos/simon-gladman/velvetoptimiser/velvetoptimiser/2.2.6
Job State	ok
Command Line	export OMP_NUM_THREADS=2 && export OMP_THREAD_LIMIT=2 && VelvetOptimiser.pl -t ..
Tool Standard Output	Logfile name: 13-04-2024-01-56-14_Logfile.txt
Tool Standard Error	***** VelvetOptimiser.pl Version ...
Tool Exit Code	0
Job API ID	11ac94870d0bb33ab62dc513be56a80d

Dataset Storage

This dataset is stored in a Galaxy sharable storage location with id **files24**.

8.8 GB of 250 GB used
3% of disk quota used

Inheritance Chain

Fig 5: Click on ‘Total Standard Error’

```
Velvet hash value: 55
Roadmap file size: 1237819
Total number of contigs: 36
n50: 132290
length of longest contig: 132290
Total bases in contigs: 182468
Number of contigs > 1k: 4
Total bases in contigs > 1k: 177614
```

Fig 6: Size, Length and Number of Contigs built from the Sample

RESULTS:

There is total 36 contigs built. The length of the longest contig is 132290. Total bases in contigs are 182468.

CONCLUSION:

De Novo Assembly was performed on query using Velveth Tool.

REFERENCES:

1. Introduction to de novo assembly with Velvet - Bioinformatics Documentation. (n.d.).
<https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly-background/>
 2. Introduction to de novo genome assembly for Illumina reads - Bioinformatics Documentation. (n.d.).
<https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly-protocol/>
-

DATE:13/04/2024

WEBLEM 9(F)
Quality Assessment Tool-QUAST
(URL: <https://usegalaxy.eu/>)

AIM:

To assess the quality of the sequence assemblies using QUAST tool

INTRODUCTION:

Modern DNA sequencing technologies cannot produce the complete sequence of a chromosome. Instead, they generate large numbers of reads, ranging from dozens to thousands of consecutive bases, sampled from different parts of the genome. Genome assembly software combines the reads into larger regions called contigs. However, current sequencing technologies and software face many complications that impede reconstruction of full chromosomes, including errors in reads and large repeats in the genome.

Different assembly programs use different heuristic approaches to tackle these challenges, resulting in many differences in the contigs they output. This leads to the questions of how to assess the quality of an assembly and how to compare different assemblies.

QUAST—a quality assessment tool for evaluating and comparing genome assemblies. This tool improves on leading assembly comparison software with new ideas and quality metrics. QUAST can evaluate assemblies both with a reference genome, as well as without a reference. QUAST produces many reports, summary tables and plots to help scientists in their research and in their publications. It is rather fast, and its most time-consuming steps are parallelized; therefore, it can be effectively run on multi-core processors

QUAST aggregates methods and quality metrics from existing software, such as Plantagora, GAGE, GeneMark.hmm and GlimmerHMM, and it extends these with new metrics. For example, the well-known N50 statistic can be artificially increased by concatenating contigs, at the expense of increasing the number of misassemblies. QUAST also computes metrics that are useful for assessing assemblies of previously unsequenced species, whereas most other assembly assessment software require a reference genome.

Many assembly algorithms have been developed for the challenging problem of genome assembly from short reads. QUAST will help scientists to assess different assembly software to choose the best pipeline for their research, and it will help developers of genome assemblers to improve their software and algorithms.

METHODOLOGY:

1. Open home page of Galaxy server.
2. In the search panel search QUAST tool
3. Upload the files of BAM and SAM data (Alignment mapping step) in Assembly mode
4. Select Co-assembly from the two given options
5. Set the parameters accordingly
6. Click on “RUN TOOL”.

OBSERVATIONS:

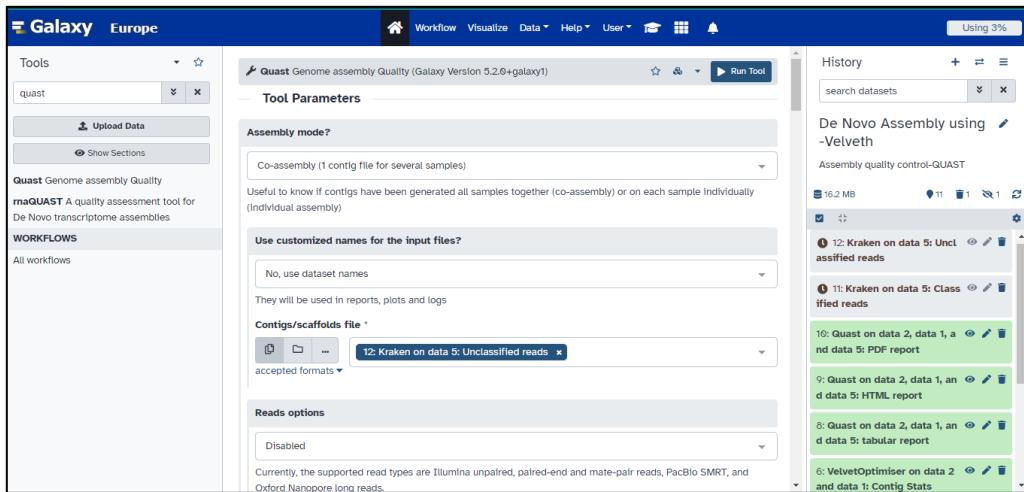


Fig 1: Home page of QUAST Tool

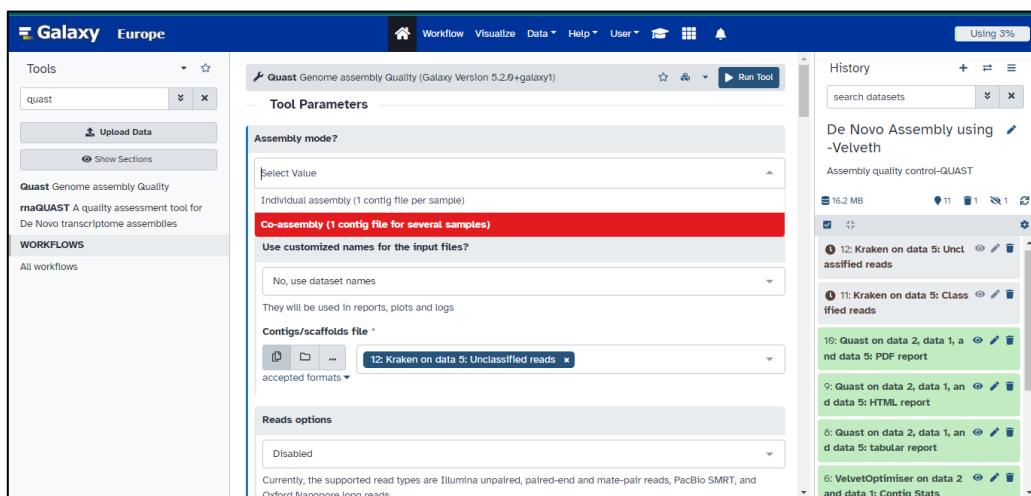


Fig 2.1: Upload the BAM and SAM files obtained from Alignment mapping

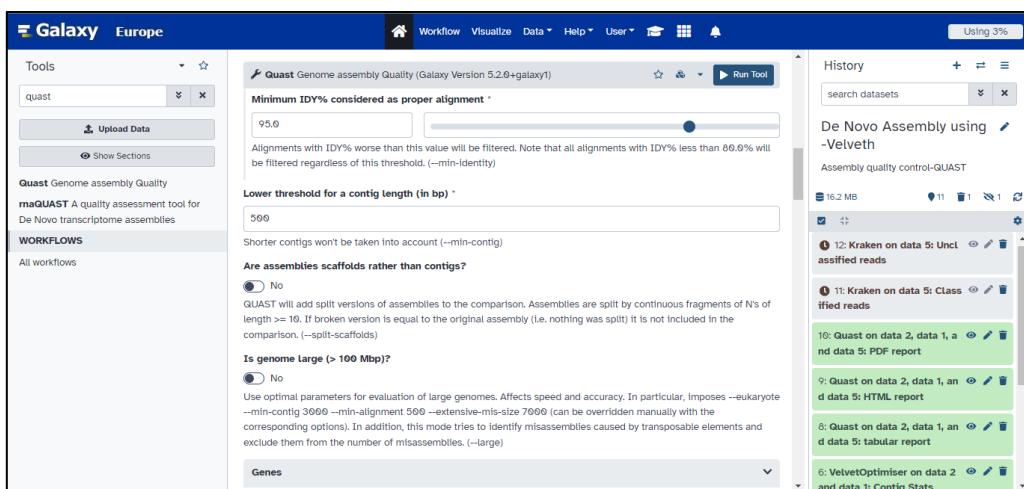


Fig 2.2: Set the parameters accordingly

Fig 2.3: Select the output files i.e. HTML Reports, PDF Reports, Tabular Reports

Fig 3: The PDF report, HTML Report and Tabular Report can be visualized

VelvetOptimiser_on_data_2_and_data_1_Contigs	
# contigs (>= 0 bp)	36
# contigs (>= 1000 bp)	4
Total length (>= 0 bp)	182468
Total length (>= 1000 bp)	177614
# contigs	5
Largest contig	132290
Total length	178145
GC (%)	33.60
N50	132290
N90	34980
auN	105411.6
L50	1
L90	2
# N's per 100 kbp	67.36

Fig 4: Tabular data showing sequence information

1. Contigs: is the total number of contigs in the assembly.
2. Total length: is the total number of bases in the assembly.
3. GC (%): is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.
4. N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly.
5. L50 (L_x , LG_{50} , LG_x) is the number of contigs equal to or longer than N50 (N_x , NG_{50} , NG_x)
6. In other words, L_{50} , for example, is the minimal number of contigs that cover half the assembly.

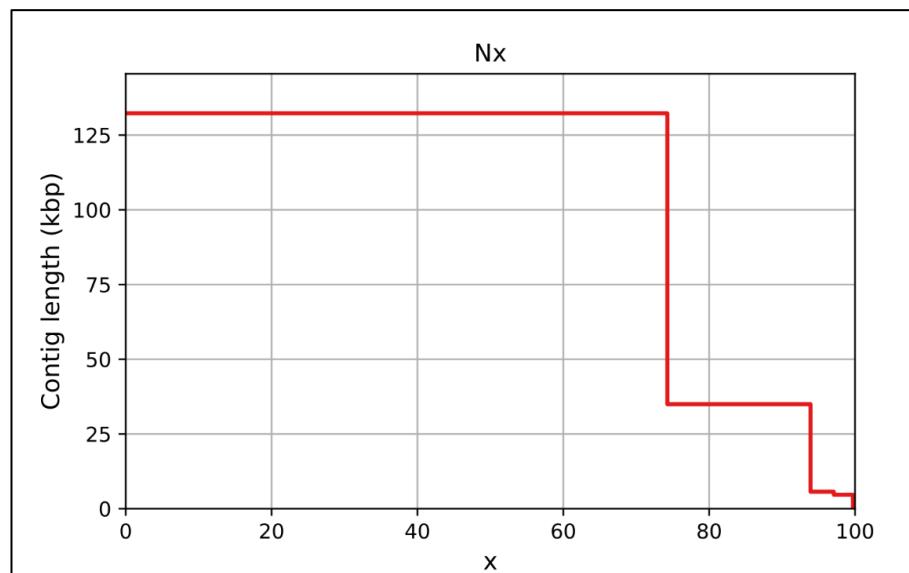


Fig 5: Nx-like plot

1. Nx plot shows Nx values as x varies from 0 to 100 %.
2. Plot of contig number vs. contig length.
3. NGx plot shows NGx values as x varies from 0 to 100 %.

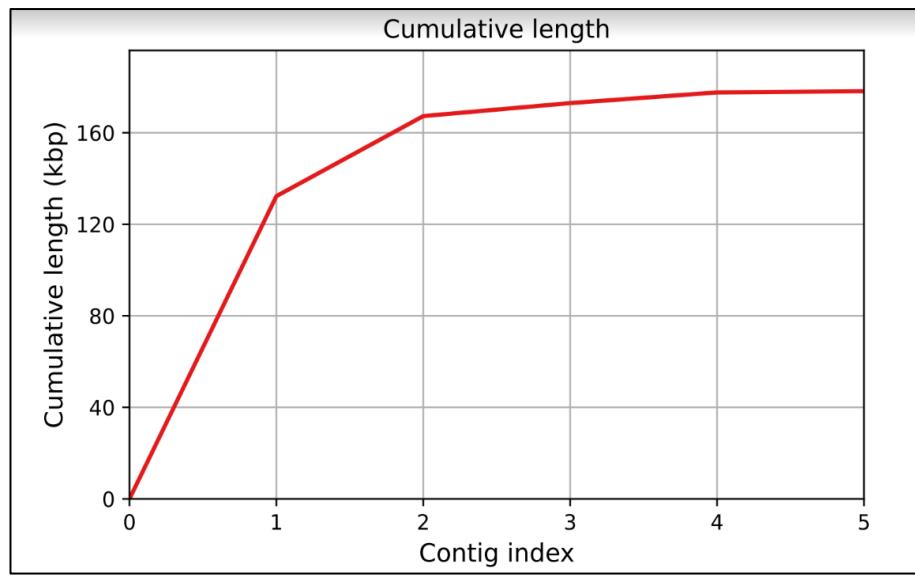


Fig 6: Graph of cumulative length

1. Cumulative length plot shows the growth of contig lengths. On the x-axis, contigs are ordered from the largest to smallest.
2. Graph of contig index vs cumulative length.
3. The y-axis gives the size of the x largest contigs in the assembly.

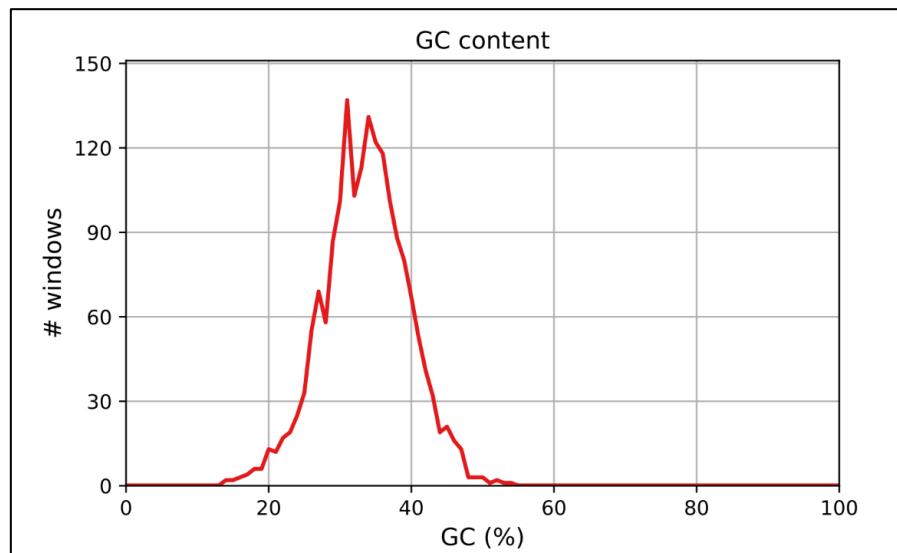


Fig 7: Graph of GC Content

1. GC content plot shows the distribution of GC content in the contigs.
2. Graph of GC% vs #hash windows
3. The x value is the GC percentage (0 to 100 %).
4. The y value is the number of non-overlapping 100 bp windows which GC content equals x %.

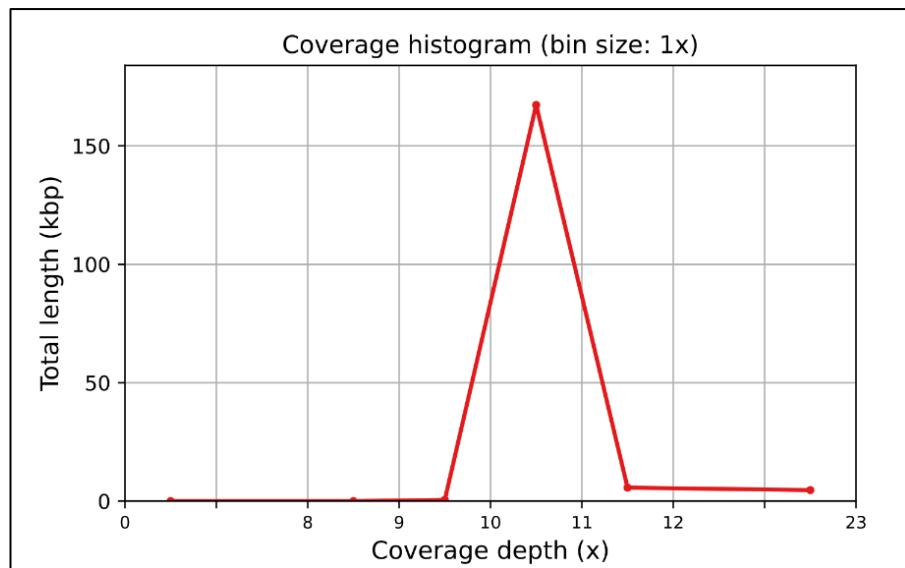


Fig 8: Graph of Coverage histogram

1. Coverage histogram shows distribution of total contig lengths (y-axis) at different read coverage depths (x-axis, grouped in bins).
2. Coverage bin size is automatically selected based on the number of contigs and coverage deviation.

Note: These histograms are only available for assemblies with SPAdes/ Velvet-like contig naming style (...length_X_cov_Y...).

RESULTS:

The QUAST tool was explored and studied to evaluate the quality of genome assemblies with the help of metrics like N50, L50, and GC content was studied to assess the completeness and accuracy of the assemblies.

CONCLUSION:

The completeness and accuracy of the assemblies were assessed by exploring and studying the QUAST tool.

REFERENCES:

1. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072-5.
doi: 10.1093/bioinformatics/btt086. Epub 2013 Feb 19. PMID: 23422339; PMCID: PMC3624806.
2. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler Author NotesBioinformatics, Volume 29, Issue 8, April 2013, Pages 1072–1075,
<https://doi.org/10.1093/bioinformatics/btt086>
3. QUAST 5.2.0 manual. (n.d.). <https://quast.sourceforge.net/docs/manual.html>