Introduction to Bioinformatic Tools

1

[1]UCSC

Oct 06, 2020

1 Works for me This document is published without a DOI.

UCSC BME 22L

Alyssa Ayala

DOCUMENT CITATION

2020. Introduction to Bioinformatic Tools. **protocols.io**
https://protocols.io/view/introduction-to-bioinformatic-tools-bmfmk3k6

CREATED

Sep 16, 2020

LAST MODIFIED

Oct 06, 2020

DOCUMENT INTEGER ID

42189

DISCLAIMER:

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Introduction to Bioinformatic Tools

Goals

The goal of this lab is to get students well acquainted and familiar with commonly used tools necessary for sequence analysis.

Lesson Plan

Students will learn and perform:

- How to navigate the UCSC Genome Browser
- How to utilize NCBI Blast

Safety

NO PPE IS REQUIRED FOR THIS LAB

For this lab, there will be no need for safety requirements because students will be asked to only use their laptops.

Tips and Hazards

- Highlight important regions on Genome Browser tracks; it helps to better visualize.

Background

In the current age of molecular biology, it is almost essential that people are up to speed and familiar with the bioinformatics tools at their disposal. Although it is good to know how to do actual molecular biology lab techniques, they can almost seem useless without having the bioinformatics skills necessary to interpret and analyze data. This lesson plan has been designed with the intention of introducing some bioinformatics skills and tools so you are capable of analyzing your own data computationally.

With the advancement of computer science and sequencing technologies, comes along the emergence of the field known as bioinformatics. This interdisciplinary field encompasses an array of sciences to ultimately help scientists interpret and analyze biological data. Depending on the field these tools can fit their needs of research and analysis. For instance, a clinical geneticist might use bioinformatic tools to identify commonly known SNPs (single nucleotide polymorphisms) in a patient's genome in order to find diseases associated with the variant.

The ability to analyze nucleic acid and amino acid sequences efficiently is one of the biggest attractions in the field of computational biology. There are several tools bioinformaticians use to get specific and accurate sequence information from databases and resources online. Tools such as BLAST (Basic Local Alignment Search Tool), Geneious Prime, NCBI, and the UCSC Genome Browser provide researchers with their desired genetic information and allow analysis computationally. For this lab, we will be looking at a couple of tools that will be used throughout the remainder of this course.

BLAST Introduction

BLAST (Basic Local Alignment Search Tool) is one of the most widely used tools to gain sequence information. Finding similarity between DNA and protein sequences against a database is one of the first things people do when trying to get immediate information about a sequence of interest. Doing these searches allows scientists to gain knowledge about that particular gene's function. BLAST finds regions of similarity between the input sequence and sequences found in its databases. The program compares nucleotide or protein sequences to sequence databases and then calculates the statistical significance of matches. Doing this search allows scientists to infer functional and evolutionary relationships between sequences and helps identify members of the gene family. BLAST makes use of heuristics to help provide the user with the sequence information quickly. This process occurs through a "speed-read" over similar nucleotides in the respective database. How specific these searches are can be adjusted to the user's desires.

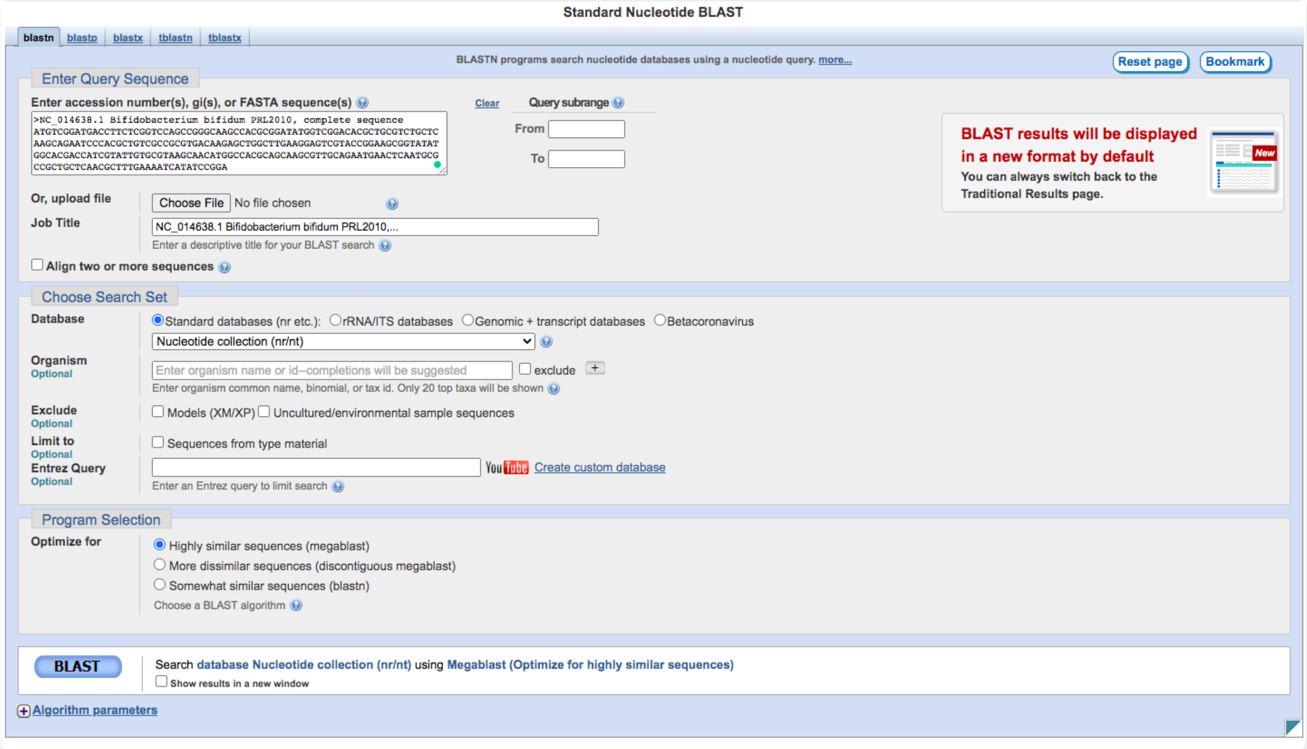
There are different versions of BLAST that can be used for different reasons depending on what sequence you have. Here are the various forms of BLAST and the reasons why each form may be advantageous given the scenario:

Program	Database	Query	Typical Uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses

BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

Overview of How it Works (BLAST)

BLAST makes use of entry sequences called “queries” and compares them to nucleotide and protein sequences called “subject sequences” in a database. Each character in the sequence then gets indexed by their starting position in the sequence. The “wordsize” option is used by the user to configure how long the length of the string they are going to the index will be. The default values for word size for protein BLAST are 3 and the default size for nucleotide BLAST is 11. The query gets accepted as a FASTA and every nucleotide or amino acid is paired to or aligned to a letter or gap of the subject sequence. The overall alignment score is determined by summing up the scores of each nucleotide over the length of the entire sequence. Nucleotide BLAST scores nucleotides by giving +2 for aligned pairs of identical letters and a -3 for every nonidentical aligned pair. For the protein BLAST, scores for every amino acid pair are provided in a substitution matrix. Likely protein pairs are given a positive score whereas unlikely pairs are given a negative score.



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>NC_014638.1 Bifidobacterium bifidum PRL2010, complete sequence
ATGTCGGATGACCTTCTCGGTCCAGCCGGGCAAGCCACGCGGATATGGTCGGACACGCTGCGTCTGCTC
AAGCAGAATCCCACGCTGTCGCCGCGTGACAAGAGCTGGCTTGAAGGAGTCGTACCGGAAGCGGTATAT
GGCACGACCATCGTATTGTGCGTAAGCAACATGGCCACGCAGCAAGCGTTGCAGAATGAACTCAATGCG
CCGCTGCTCAACGCTTTGAAAATCATATCCGGA
```

Clear

Query subrange

From

To

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Or, upload file: Choose File No file chosen

Job Title: NC_014638.1 Bifidobacterium bifidum PRL2010,...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested □ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional: Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for: ● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

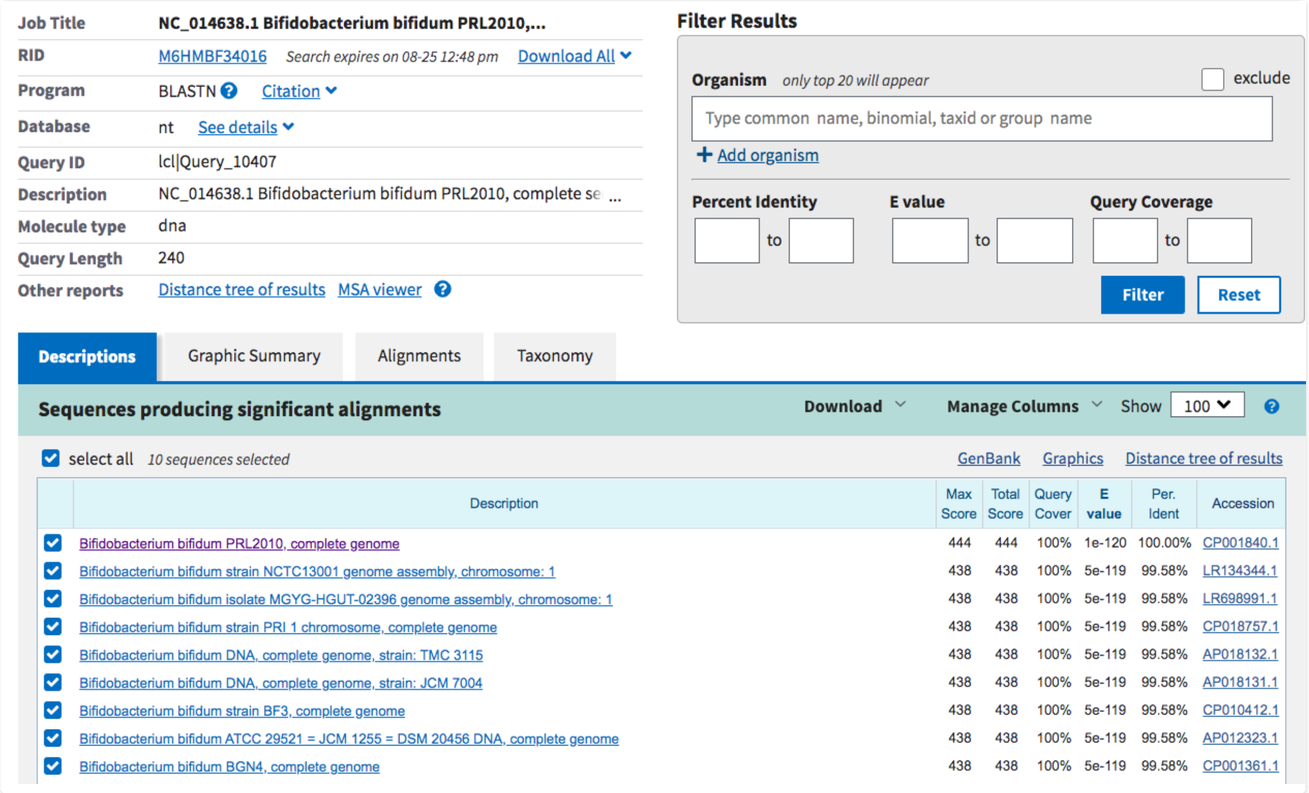
BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Algorithm parameters

Search results give a list of hits; where the most similar result appears at the start of the list. These hits can also be known as alignments. Each alignment is assigned a statistical value known as an “e-value”. The e-value is the number of times that alignment as good or better than the one found on BLAST would be expected to occur given the size of the database that was searched. The smaller the e-value the better the match. The user can set the threshold for the e-value and this determines which alignments will appear. A higher “Expect Value” threshold is less stringent and the BLAST default of “10” is designed to ensure that no biologically significant alignment is missed. However, “Expect Values” in the range of 0.001 to 0.0000001 are commonly used to restrict the alignments shown to those of high quality.



Job Title	NC_014638.1 Bifidobacterium bifidum PRL2010,...
RID	M6HMBF34016 Search expires on 08-25 12:48 pm Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_10407
Description	NC_014638.1 Bifidobacterium bifidum PRL2010, complete se ...
Molecule type	dna
Query Length	240
Other reports	Distance tree of results MSA viewer

Filter Results

Organism only top 20 will appear □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

Sequences producing significant alignments

☑ select all 10 sequences selected

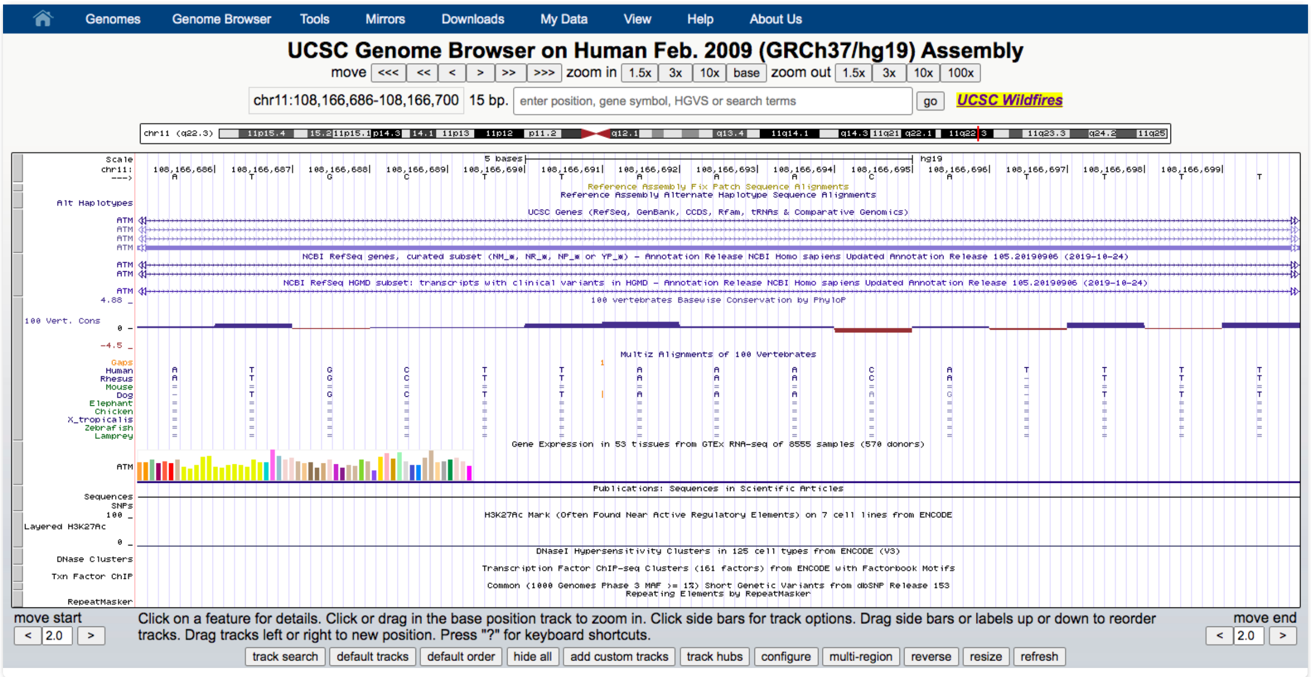
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Bifidobacterium bifidum PRL2010, complete genome	444	444	100%	1e-120	100.00%	CP001840.1
☑	Bifidobacterium bifidum strain NCTC13001 genome assembly, chromosome: 1	438	438	100%	5e-119	99.58%	LR134344.1
☑	Bifidobacterium bifidum isolate MGYG-HGUT-02396 genome assembly, chromosome: 1	438	438	100%	5e-119	99.58%	LR698991.1
☑	Bifidobacterium bifidum strain PRI 1 chromosome, complete genome	438	438	100%	5e-119	99.58%	CP018757.1
☑	Bifidobacterium bifidum DNA, complete genome, strain: TMC 3115	438	438	100%	5e-119	99.58%	AP018132.1
☑	Bifidobacterium bifidum DNA, complete genome, strain: JCM 7004	438	438	100%	5e-119	99.58%	AP018131.1
☑	Bifidobacterium bifidum strain BF3, complete genome	438	438	100%	5e-119	99.58%	CP010412.1
☑	Bifidobacterium bifidum ATCC 29521 = JCM 1255 = DSM 20456 DNA, complete genome	438	438	100%	5e-119	99.58%	AP012323.1
☑	Bifidobacterium bifidum BGN4, complete genome	438	438	100%	5e-119	99.58%	CP001361.1

UCSC Genome Browser Introduction

The UCSC Genome Browser is an online application that establishes the reference genomes for many species, including humans. Scientists use the genome browser as a reference tool in many different disciplinary fields. It can be used in bioinformatics, clinical genetics, genomic research, pharmaceutical development, and many others. Scientists can navigate the entire human genome, as well as other species, base pair by base pair. The genome browser application provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks. Tracks can be added to the display of the genome browser and serve as an additional tool for more information on specific parts of the genome. The website itself has multiple reference species outside of the human genome, including SARS Covid-19, and are considered model organisms. A Model organism is a non-human species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the model organism will provide insight into the workings of other organisms [wiki].

Overview of How it Works (UCSC Genome Browser)

To open a track, there must be a specific species genome to look at. For the purpose of this course, we will look at the GRCh37/hg19 version which is a version of the human genome assembled in 2009. Once the version is selected, input a specific region to look at. An input region can be any chromosomal position (ex. chr11:108,093,559-108,239,826) or specific gene/transcription (ex. ATM). The default display shows the region of interest with associated nucleotide sequences, genes, and other tracks.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

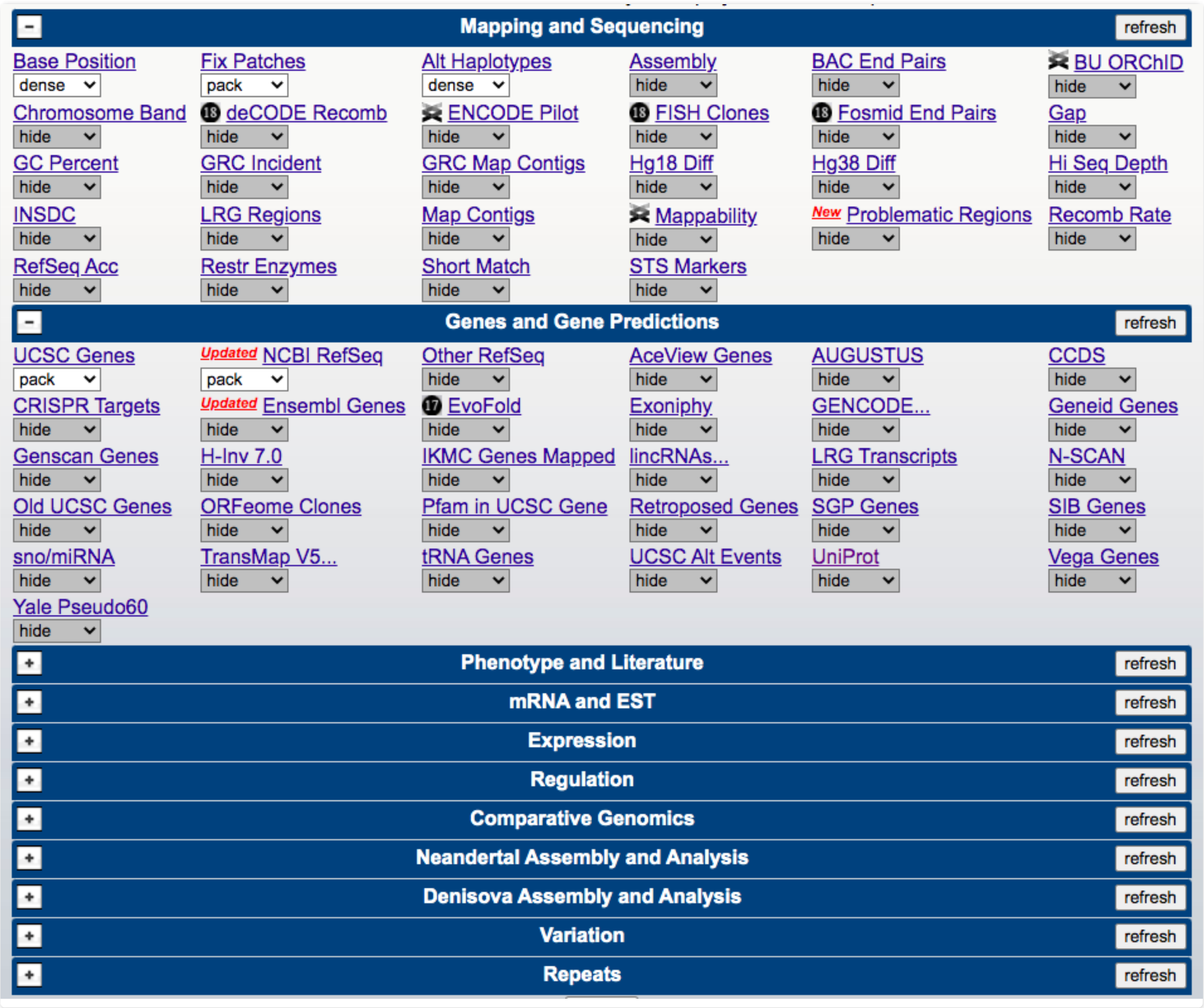
chr11:108,166,686-108,166,700 15 bp. enter position, gene symbol, HGVS or search terms

Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 105.20190906 (2019-10-24)
NCBI RefSeq HGMD subset: transcripts with clinical variants in HGMD - Annotation Release NCBI Homo sapiens Updated Annotation Release 105.20190906 (2019-10-24)
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
Publications: Sequences in Scientific Articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Common (1000 Genomes Phase 3 MAF >= 1%) Short Genetic Variants from dbSNP Release 153
Repeating Elements by RepeatMasker

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

The regions of interest can be altered directly on the display screen using the zoom in or out buttons or with the move buttons. The default display depicts the reference nucleotides in the leading strand and can be indicated by the arrow on the first track, left side of the screen. However, the display can be switched to depict the lagging strand by clicking on the arrow.

These tracks are annotated tools that serve a specific purpose such as displaying common SNPs (single nucleotide polymorphism) or protein domains (Uniprot). These tracks can be moved on the display by dragging and dropping the grey bars on the left-hand column. These tracks can also be added or removed from the display. All possible tracks are displayed below the tracks and are given in multiple categories; such as Mapping and Sequencing, Genes and Gene Predictions, and others. Add tracks by changing the status from 'hide' to any other option; preferred for this course would be 'pack'. Descriptions on tracks are given if the name of the track is clicked.



Mapping and Sequencing refresh

Base Position dense	Fix Patches pack	Alt Haplotypes dense	Assembly hide	BAC End Pairs hide	BU ORChID hide
Chromosome Band hide	⑱ deCODE Recomb hide	ENCODE Pilot hide	⑱ FISH Clones hide	⑱ Fosmid End Pairs hide	Gap hide
GC Percent hide	GRC Incident hide	GRC Map Contigs hide	Hg18 Diff hide	Hg38 Diff hide	Hi Seq Depth hide
INSDC hide	LRG Regions hide	Map Contigs hide	Mappability hide	New Problematic Regions hide	Recomb Rate hide
RefSeq Acc hide	Restr Enzymes hide	Short Match hide	STS Markers hide		

Genes and Gene Predictions refresh

UCSC Genes pack	Updated NCBI RefSeq pack	Other RefSeq hide	AceView Genes hide	AUGUSTUS hide	CCDS hide
CRISPR Targets hide	Updated Ensembl Genes hide	⑰ EvoFold hide	Exoniphy hide	GENCODE... hide	Geneid Genes hide
Genscan Genes hide	H-Inv 7.0 hide	IKMC Genes Mapped hide	lincRNAs... hide	LRG Transcripts hide	N-SCAN hide
Old UCSC Genes hide	ORFeome Clones hide	Pfam in UCSC Gene hide	Retroposed Genes hide	SGP Genes hide	SIB Genes hide
sno/miRNA hide	TransMap V5... hide	tRNA Genes hide	UCSC Alt Events hide	UniProt hide	Vega Genes hide
Yale Pseudo60 hide					

Phenotype and Literature refresh

mRNA and EST refresh

Expression refresh

Regulation refresh

Comparative Genomics refresh

Neandertal Assembly and Analysis refresh

Denisova Assembly and Analysis refresh

Variation refresh

Repeats refresh

Resources

Here are some resources that can be of use when first getting started with using these bioinformatics tools or working with Unix:

- Linux Beginner Cheat Sheet
- BLAST NCBI Handbook
- Getting Started Genome Browser
- Introduction to Unix, Sean Davis Tutorial

Disclaimer:

The information provided on this document is intended for the educational purposes of the BME 22L laboratory course. It is worth noting that the information listed on this document is subject to change and is not finalized. Therefore, the information on this document should not be used outside of this course.