Bversity's

MEGA BIOINFORMATICS INTERNSHIP '23

**SUBMITTED
BY
SOUMILI PANDEY**

Day 2:
Gene Annotation
Protein Name: BRCA1
Protein ID –P38398
Find the following Gene function details

Location- Start – 494 and end - 575
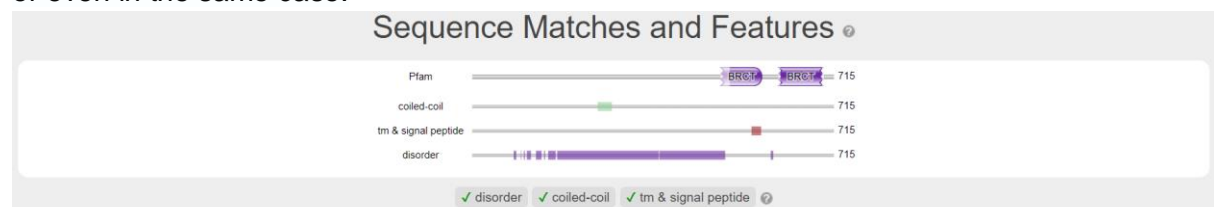   Start – 608 and end - 694
Family - BRCA1-associated
Clan – CL0459
Domain – BRCT Domain, BRCT Domain profile, BRCA1 C terminus (BRCT) Domain,
BRCT_7
Motif - phosphorylated pSXXF
E value – Max: 3.9e-07, Min: 2.0e-11
         Max: 5.5e-10, Min: 2.8e-14
Description of function - A common malignancy originating from breast epithelial tissue.
Breast neoplasms can be distinguished by their histologic pattern. Invasive ductal carcinoma
is by far the most common type. Breast cancer is etiologically and genetically
heterogeneous. Important genetic factors have been indicated by familial occurrence and
bilateral involvement. Mutations at more than one locus can be involved in different families
or even in the same case.

Sequence Matches and Features

Pfam Matches								Advanced
Family		**Clan**	**Description**	**Cross-references**	**Start**	**End**	**Domain E-values**	
Id	**Accession**						**Ind.**	**Cond.**
> BRCT	PF00533.29	CL0459	BRCA1 C Terminus (BRCT) domain		494	575	3.9e-07	2.0e-11
> BRCT	PF00533.29	CL0459	BRCA1 C Terminus (BRCT) domain		608	694	5.5e-10	2.8e-14

Day 3: Phylogenetics

Construct a Phylogenetic tree for components of Corona virus.

You can choose any gene/protein/component associated with Corona virus for atleast 5 different species/variants. Add the screenshot of the tree here.
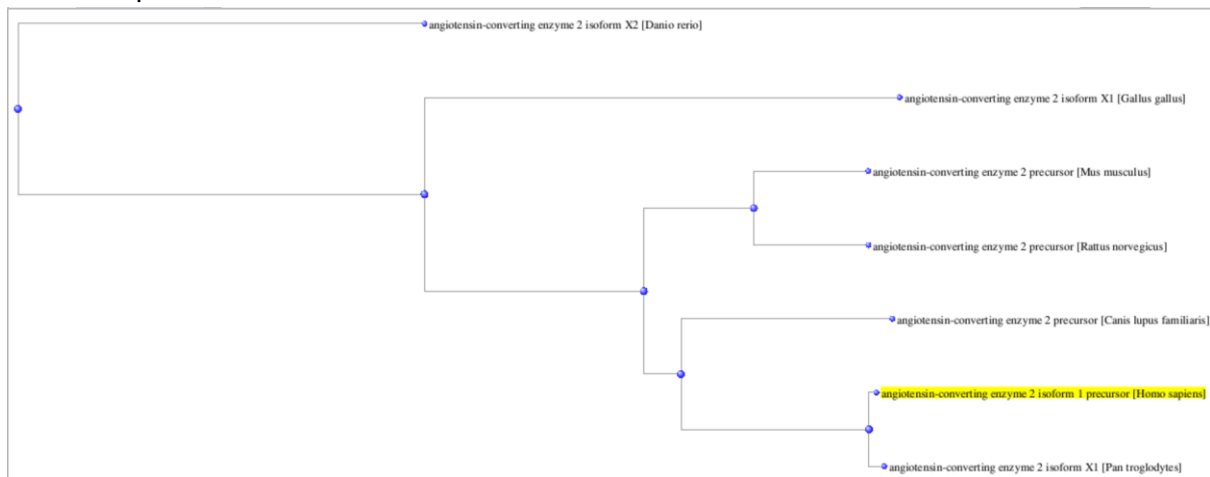
Day 4:

Genome name(the one of your interest): BRCA1

From RAST results:

Mention the desired nucleotide sequence that you choose to perform BLAST on and fill in the following:

1) Nucleotide sequence -
   atgttgacgttggccaggacctttgcaaacaagccaggccaaaaagtttcaatatttacactggctgctttaataagggc
   attgatcttatcctccgtaaaggtcacctcatag
2) Location on the genome - NC_000017.11:c43170327-43044295_417_304
3) Start and end nucleotide atg and tag
4) Function - The encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC). This gene product associates with RNA polymerase II, and through the C-terminal domain, also interacts with histone deacetylase complexes. This protein thus plays a role in transcription, DNA repair of double-stranded breaks, and recombination. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers. Alternative splicing plays a role in modulating the subcellular localization and physiological function of this gene. Many alternatively spliced transcript variants, some of which are disease-associated mutations, have been described for this gene, but the full-length natures of only some of these variants has been described. A related pseudogene, which is also located on chromosome 17, has been identified.

Perform a BLAST on the nucleotide sequence and paste a screenshot of the obtained BLAST results:

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Eukaryotic synthetic construct chromosome 17	eukaryotic synt...	211	211	100%	2e-50	100.00%	88299790	CP034495.1
PREDICTED: Homo sapiens NBR1 autophagy cargo receptor (NBR1), transcript variant X5, mRNA	Homo sapiens	211	211	100%	2e-50	100.00%	8009	XM_054316118.1
PREDICTED: Homo sapiens NBR1 autophagy cargo receptor (NBR1), transcript variant X1, mRNA	Homo sapiens	211	211	100%	2e-50	100.00%	8219	XM_054316114.1
Homo sapiens isolate CHM13 chromosome 17	Homo sapiens	211	211	100%	2e-50	100.00%	84276897	CP068261.2
Homo sapiens DNA, chromosome 17, nearly complete genome	Homo sapiens	211	211	100%	2e-50	100.00%	80688777	AP023477.1
Homo sapiens BRCA1 DNA repair associated (BRCA1), RefSeqGene (LRG_292) on chromosome 17	Homo sapiens	211	211	100%	2e-50	100.00%	193689	NG_005905.2
Homo sapiens BRCA1 pseudogene 1 (BRCA1P1) on chromosome 17	Homo sapiens	211	211	100%	2e-50	100.00%	2085	NG_003183.3
Homo sapiens neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; neighbor of BRCA1 gene 2 (NBR2)...	Homo sapiens	211	211	100%	2e-50	100.00%	150505	DQ478408.1
Homo sapiens clone mck41_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; hypothetical protein...	Homo sapiens	211	211	100%	2e-50	100.00%	147576	DQ190457.1
Homo sapiens clone mck578_U neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical ...	Homo sapiens	211	211	100%	2e-50	100.00%	150669	DQ190456.1
Homo sapiens clone mck554_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical p...	Homo sapiens	211	211	100%	2e-50	100.00%	155470	DQ190455.1
Homo sapiens clone mck43_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical pr...	Homo sapiens	211	211	100%	2e-50	100.00%	150582	DQ190454.1
Homo sapiens clone mck55_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical pr...	Homo sapiens	211	211	100%	2e-50	100.00%	156879	DQ190453.1
Homo sapiens clone mck94_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical pr...	Homo sapiens	211	211	100%	2e-50	100.00%	156121	DQ190452.1
Homo sapiens clone mck47_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical pr...	Homo sapiens	211	211	100%	2e-50	100.00%	161765	DQ190451.1
Homo sapiens clone mck432_A neighbor of BRCA1 gene 1 (NBR1) gene, partial cds; and hypothetical p...	Homo sapiens	211	211	100%	2e-50	100.00%	167910	DQ190450.1
Homo sapiens chromosome 17, clone RP11-242D8, complete sequence	Homo sapiens	211	211	100%	2e-50	100.00%	110669	AC060780.18
Human 1A1.3B gene, promoter region	Homo sapiens	211	211	100%	2e-50	100.00%	3800	U72483.1
PREDICTED: Gorilla gorilla gorilla uncharacterized LOC109026861 (LOC109026861), ncRNA	Gorilla gorilla g...	206	206	100%	9e-49	99.12%	1030	XR_008680118.1
Pan troglodytes BRCA1 pseudogene, NBR1 testis isoform-like, and NBR1 general isoform-like genes, c...	Pan troglodytes	200	200	100%	4e-47	98.25%	1703	AY581856.1
Human BRCA1 pseudogene, exons 1a, 1b, and 2	Homo sapiens	195	195	100%	2e-45	97.39%	4068	U77841.2
PREDICTED: Pongo abelii uncharacterized LOC129054949 (LOC129054949), transcript variant X2, nc...	Pongo abelii	187	187	99%	3e-43	96.46%	4659	XR_008519590.1
PREDICTED: Pongo abelii uncharacterized LOC129054949 (LOC129054949), transcript variant X1, nc...	Pongo abelii	187	187	99%	3e-43	96.46%	4537	XR_008519589.1
Hylobates lar NBR1 testis isoform-like and NBR1 general isoform-like genes, complete sequence	Hylobates lar	178	178	100%	2e-40	94.74%	1706	AY581858.1
Homo sapiens ATAC-STARR-seq lymphoblastoid active region 12236 (LOC130060935) on chromosome...	Homo sapiens	167	167	78%	4e-37	100.00%	410	NG_196180.1
PREDICTED: Lemur catta ribosomal protein lateral stalk subunit P1 (RPLP1), mRNA	Lemur catta	152	152	98%	1e-32	91.07%	528	XM_045530739.1
Pan troglodytes chromosome 16 clone CH251-334K6, complete sequence	Pan troglodytes	150	150	98%	4e-32	91.07%	166532	AC275798.1
PREDICTED: Lemur catta 60S acidic ribosomal protein P1-like (LOC123628303), mRNA	Lemur catta	147	147	95%	5e-31	90.83%	384	XM_045537939.1
PREDICTED: Artibeus jamaicensis ribosomal protein lateral stalk subunit P1 (RPLP1), mRNA	Artibeus jamaic...	141	141	98%	3e-29	89.29%	533	XM_037155800.2
PREDICTED: Artibeus jamaicensis 60S acidic ribosomal protein P1-like (LOC128626190), mRNA	Artibeus jamaic...	139	139	97%	9e-29	89.19%	536	XM_053657419.1
PREDICTED: Myodes glareolus 60S acidic ribosomal protein P1-like (LOC125413714), mRNA	Myodes glareolus	137	137	96%	3e-28	89.09%	426	XM_048455345.1
PREDICTED: Artibeus jamaicensis 60S acidic ribosomal protein P1-like (LOC119054551), mRNA	Artibeus jamaic...	134	134	97%	4e-27	88.29%	440	XM_037152469.2
PREDICTED: Phyllostomus hastatus ribosomal protein lateral stalk subunit P1 (RPLP1), mRNA	Phyllostomus h...	130	130	98%	5e-26	87.50%	532	XM_045823885.1
PREDICTED: Acomys russatus 60S acidic ribosomal protein P1-like (LOC127189943), transcript variant...	Acomys russatus	128	128	89%	2e-25	89.22%	339	XM_051147000.1
Acomys russatus genome assembly, chromosome: 5	Acomys russatus	128	128	89%	2e-25	89.22%	79927262	LR877216.1
PREDICTED: Phyllostomus discolor ribosomal protein lateral stalk subunit P1 (RPLP1), mRNA	Phyllostomus di...	124	124	98%	3e-24	86.61%	530	XM_028506807.2
PREDICTED: Phyllostomus discolor 60S acidic ribosomal protein P1-like (LOC114490603), mRNA	Phyllostomus di...	124	124	98%	3e-24	86.61%	458	XM_028504664.2
PREDICTED: Phyllostomus discolor 60S acidic ribosomal protein P1-like (LOC114496558), mRNA	Phyllostomus di...	121	121	85%	3e-23	88.78%	304	XM_036026041.1
Eukaryotic synthetic construct chromosome 15	eukaryotic synt...	117	117	68%	4e-22	93.59%	82521392	CP034493.1
Homo sapiens isolate CHM13 chromosome 15	Homo sapiens	117	117	68%	4e-22	93.59%	99753195	CP068263.2
Homo sapiens DNA, chromosome 15, nearly complete genome	Homo sapiens	117	117	68%	4e-22	93.59%	95537968	AP023475.1
Homo sapiens cDNA, FLJ18498	Homo sapiens	117	117	68%	4e-22	93.59%	907	AK311456.1
Homo sapiens chromosome 15, clone RP11-253M7, complete sequence	Homo sapiens	117	117	68%	4e-22	93.59%	157448	AC027237.8
Homo sapiens RPP1 gene for ribosomal protein P1, complete cds and sequence	Homo sapiens	117	117	68%	4e-22	93.59%	4816	AB061836.1
PREDICTED: Sapajus apella 60S acidic ribosomal protein P1 pseudogene (LOC116564252), misc_RNA	Sapajus apella	111	111	77%	2e-20	89.89%	451	XR_004274813.1
PREDICTED: Saimiri boliviensis boliviensis 60S acidic ribosomal protein P1-like (LOC120363921), misc_	Saimiri bolivien...	111	111	77%	2e-20	89.89%	382	XR_005579329.1
PREDICTED: Cebus imitator 60S acidic ribosomal protein P1-like (LOC108288894), misc_RNA	Cebus imitator	111	111	77%	2e-20	89.89%	488	XR_001818894.2
Felis catus Senzu DNA, chromosome: D4, American Shorthair breed	Felis catus	102	102	98%	1e-17	83.04%	96896805	AP023164.1
PREDICTED: Phyllostomus discolor 60S acidic ribosomal protein P1-like (LOC114505968), mRNA	Phyllostomus di...	102	102	85%	1e-17	85.57%	487	XM_036010130.1
PREDICTED: Aotus nancymaae 60S acidic ribosomal protein P1 pseudogene (LOC105710629), misc_...	Aotus nancymaae	95.3	95.3	77%	2e-15	86.52%	560	XR_001105835.1

hover to see the title   click to show alignments

Alignment Scores   ■ < 40   ■ 40 - 50   ■ 50 - 80   ■ 80 - 200   ■ >= 200   ?

51 sequences selected ?

Close

Distribution of the top 51 Blast Hits on 51 subject sequences

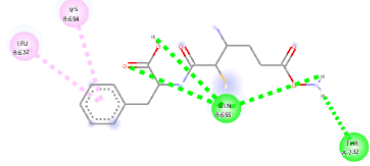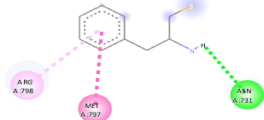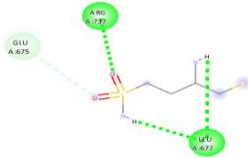Query

1   20   40   60   80   100
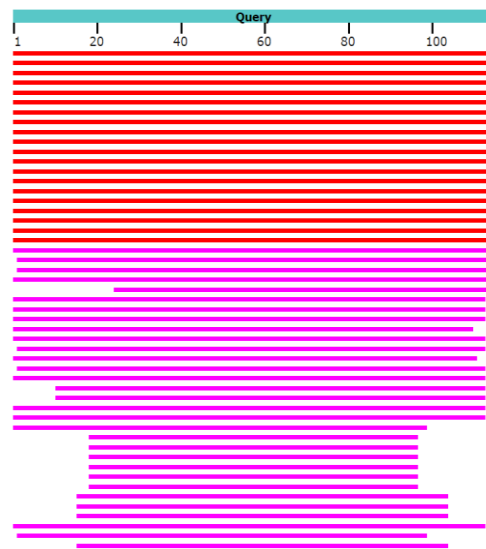
Day 5 & 6:

Molecular Docking
Protein Name: Tetanus neurotoxin translocation domain - C467S
Protein ID – 7BXX

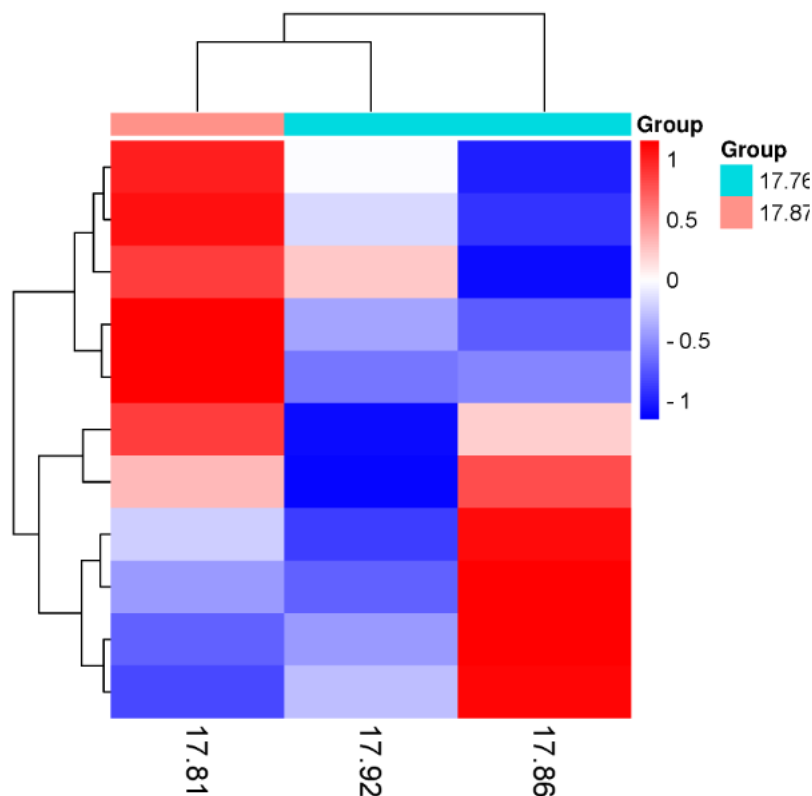| Ligand Name	Ligand ID	Follows Lipinski Rule?	Energy value	Dock Image
(2R,3R,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol Synonym - Glucoside	64689	Yes	-5.6	
CHEMBL414393	15159715	Yes	-4.7	
CHEMBL11611297	44365833	Yes	-4.3	

Day 7: Heatmap analysis
Objective : To plot a heat map and understand the differential expression based on numbered data.

Input data excel sheet link -
https://docs.google.com/spreadsheets/d/1golLLGFwt0qXHUlmcqVERfNMXt5VCJmw/edit#gid=429090649

Output heatmap:

Discussion points:
The sample data majorly provides three types of data. The genes upregulation has been denoted by red and downregulation shown in blue, other colours are denoting the regulation at middle point to consider.

Five interpretation points understood:
- 17.86 values have the highest level of upregulation.
- 17.92 can be seen going towards downregulation.
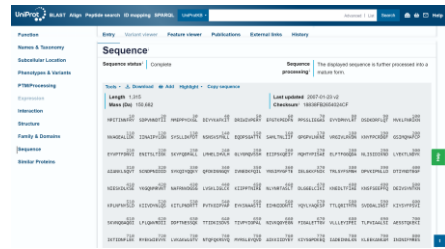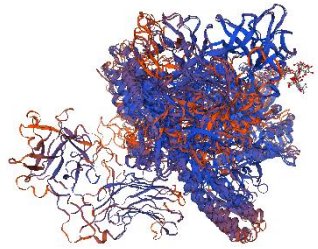- 17.81 can be seen downregulation of the gene.

Day 8 & 9: Homology Modelling:

You can choose any protein which is involved in SARS CoV-2 Pathogenesis (Eg: ACE2 receptor, any envelope protein) and can take at least 2 homologous sequences with sequence similarity >30%. Try to develop an hypothesis around it (Like Why you want to use Homology modelling for your protein of interest, Purpose and outcome of it) and more importantly how it is going to add value to your hypothesis.

Protein: Tetanus toxin
Gene: tetX
PDB: P04958

Target	Sequence	Result

Day 10:

Please paste your GitHub account link: https://github.com/Soumili0815