CAMM535 – Assignment3

Melis Oktayoglu 64388

Q1

a.

The gene I picked was BAX, and it has 12 transcripts. 6 of them are protein coding.

Gene: BAX ENSG00000087088

Description	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
Gene Synonyms	BCL2L4
Location	Chromosome 19: 48,954,815-48,961,798 forward strand. GRCh38:CM000681.2
About this gene	This gene has 12 transcripts (splice variants), 206 orthologues, 8 paralogues and is associated with 51 phenotypes.
Transcripts	Hide transcript table

Show/hide columns (1 hidden) Filter

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000345358.12	BAX-202	795	192aa	Protein coding	CCDS12742	Q07812	NM_138761.4	MANE Select, Ensembl Canonical, GENCODE basic
ENST00000293288.12	BAX-201	1358	218aa	Protein coding	CCDS12744	Q07812-2	-	GENCODE basic, TSL:1
ENST00000415969.6	BAX-205	540	179aa	Protein coding	CCDS12745	Q07812-8	-	GENCODE basic, TSL:1
ENST00000539787.2	BAX-212	458	140aa	Protein coding	CCDS77327	I6LPK7	-	GENCODE basic, TSL:1
ENST00000354470.7	BAX-203	433	143aa	Protein coding	CCDS12743	Q07812-4	-	GENCODE basic, TSL:1
ENST00000506183.5	BAX-208	383	126aa	Protein coding		H0YA56	-	TSL:1, CDS 5' incomplete
ENST00000356483.8	BAX-204	677	164aa	Nonsense mediated decay		Q07812-5	-	TSL:1
ENST00000515540.5	BAX-211	300	41aa	Nonsense mediated decay		Q07812-3	-	TSL:3
ENST00000502487.5	BAX-206	1346	No protein	Retained intron		-	-	TSL:1
ENST00000513217.1	BAX-209	869	No protein	Retained intron		-	-	TSL:2
ENST00000513545.5	BAX-210	775	No protein	Retained intron		-	-	TSL:3
ENST00000503726.2	BAX-207	488	No protein	Retained intron		-	-	TSL:2

The colors in the biotype region indicate the source of the transcript. For protein coding ones red means that its either from Ensembl or Havana, meanwhile gold means that the transcript is same in Ensembl and Havana, and finally blue means that the transcript is non-coding. In the non-coding ones, processed ones don't have a ORF, nonsense-mediated decay biotypes are as the name suggests go through nonsense mediated decay where this process detects nonsense mutations.

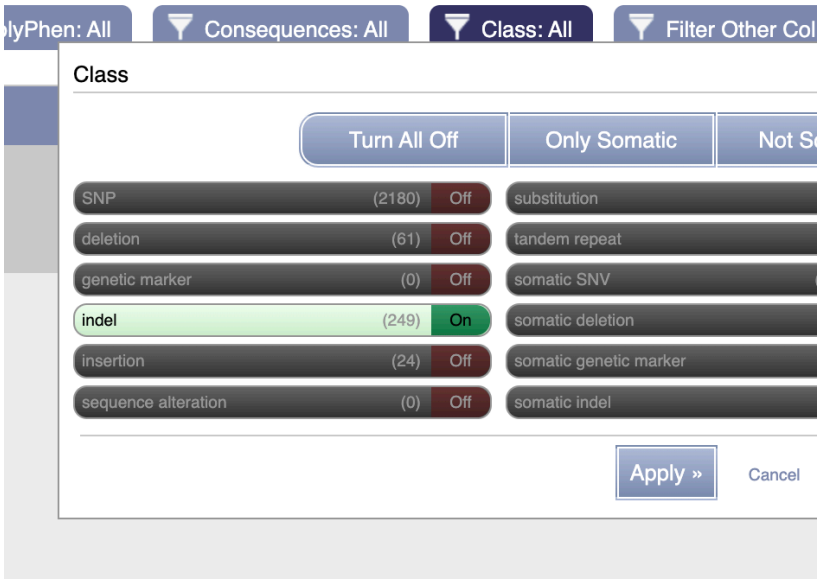
---

b.

The transcript I picked is BAX-202. It has in total 2689 variant alleles. 249 of them are of indel class.

Transcript: ENST00000345358.12 BAX-202

Description	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
Gene Synonyms	BCL2L4
Location	Chromosome 19: 48,954,875-48,961,798 forward strand.
About this transcript	This transcript has 6 exons, is annotated with 70 domains and features, is associated with 2689 variant alleles and maps to 586 oligo probes.
Gene	This transcript is a product of gene ENSG00000087088.21 Hide transcript table



c.

There are 51 phenotypes associated with the gene: (the number can be seen from the screenshot in Q1.a). I picked colorectal cancer for the disease from the source MIM morbid.

Short summary:

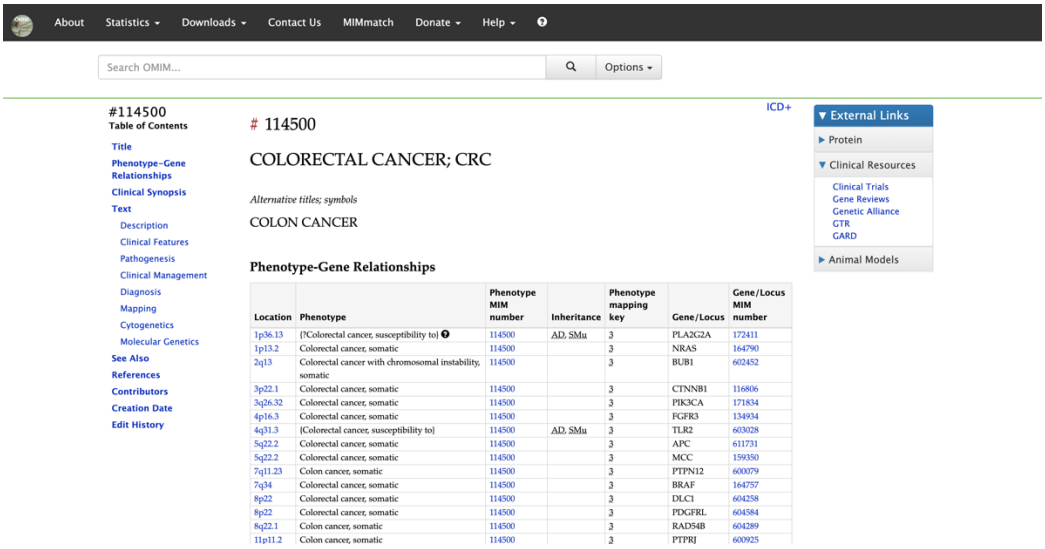
Colorectal cancer (CRC) affects both men and women and is caused by various factors including gene defects and environmental and lifestyle risks. It is characterized by changes in molecular pathways, such as chromosomal instability, CpG island methylator phenotype, and microsatellite instability. Chromosomal instability is the most common alteration in CRC, occurring in about 85% of cases.

Phenotypes

Phenotypes, diseases and traits associated with this gene ENSG00000087088

Show All entries Filter

Phenotype, disease and trait	Source	Study
Acute lymphoblastic leukemia	Cancer Gene Census	PMID:26189108
Acute myeloid leukemia	Cancer Gene Census	-
Anaplastic oligodendroglioma	Cancer Gene Census	PMID:28388591
Atypical Endometrial Hyperplasia	Cancer Gene Census	PMID:33016334
bile duct carcinoma	Cancer Gene Census	-
bladder transitional cell carcinoma	Cancer Gene Census	PMID:33004514
Breast carcinoma	Cancer Gene Census	PMID:27535334
breast ductal adenocarcinoma	Cancer Gene Census	-
cecum adenocarcinoma	Cancer Gene Census	PMID:22810696
cervical squamous cell carcinoma	Cancer Gene Census	-
Chronic lymphocytic leukemia	Cancer Gene Census	-
colon adenocarcinoma	Cancer Gene Census	Study links
colorectal adenocarcinoma	Cancer Gene Census	PMID:24599305, PMID:27149842, PMID:24211491, PMID:24755471
Colorectal Cancer	MIM morbid	-
Cutaneous melanoma	Cancer Gene Census	PMID:22197931, PMID:28467829
diffuse large B-cell lymphoma	Cancer Gene Census	PMID:24531327
Duodenal Adenocarcinoma	Cancer Gene Census	PMID:26804919
Endometrial Endometrioid Adenocarcinoma	Cancer Gene Census	PMID:33016334
endometrial serous adenocarcinoma	Cancer Gene Census	PMID:22923510



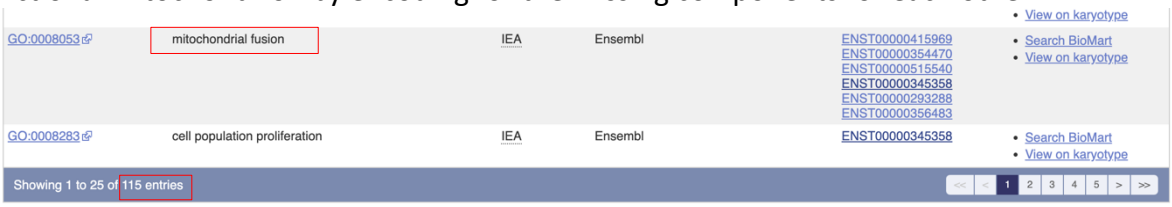
Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
1p36.13	{?Colorectal cancer, susceptibility to}	114500	AD, SMu	3	PLA2G2A	172411
1p13.2	Colorectal cancer, somatic	114500		3	NRAS	164790
2q13	Colorectal cancer with chromosomal instability, somatic	114500		3	BUB1	602452
3p22.1	Colorectal cancer, somatic	114500		3	CTNNB1	116806
3q26.32	Colorectal cancer, somatic	114500		3	PIK3CA	171834
4p16.3	Colorectal cancer, somatic	114500		3	FGFR3	134934
4q31.3	{Colorectal cancer, susceptibility to}	114500	AD, SMu	3	TLR2	603028
5q22.2	Colorectal cancer, somatic	114500		3	APC	611731
5q22.2	Colorectal cancer, somatic	114500		3	MCC	159350
7q11.23	Colon cancer, somatic	114500		3	PTPN12	600079
7q34	Colorectal cancer, somatic	114500		3	BRAF	164757
8p22	Colorectal cancer, somatic	114500		3	DLC1	604258
8p22	Colorectal cancer, somatic	114500		3	PDGFRL	604584
8q22.1	Colon cancer, somatic	114500		3	RAD54B	604289
11p11.2	Colon cancer, somatic	114500		3	PTPRJ	600925

d.

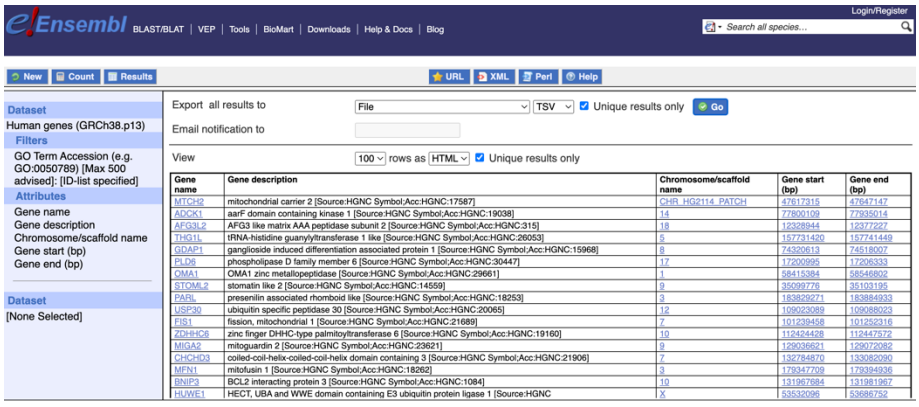
There are 115 biological processes in association with BAX gene. I picked mitochondrial fusion with GO accession code GO:0008053.

Description of mitochondrial fusion:

Mitochondrial fusion is the process by which two mitochondria merge, first through their outer and then their inner membranes. There is also mitochondrial division which is the process by which one mitochondrial body is divided into two. The combination of these two processes leads to the formation of mitochondrial networks that have an intricate and branching structure. These networks allow for the even distribution of important components such as cardiolipin within the mitochondria. The networks also enable genetic complementation, where two mitochondria with different defects can combine to form a functional mitochondrion by encoding for the missing components for each other.



The biomart table with unique results only resulted as follows:



Gene name	Gene description	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)
MTCH2	mitochondrial carrier 2 [Source:HGNC Symbol;Acc:HGNC:17587]	CHR_HG2114_PATCH	47617315	47647147
ADCK1	aarF domain containing kinase 1 [Source:HGNC Symbol;Acc:HGNC:19038]	14	77800109	77935014
AFG3L2	AFG3 like matrix AAA peptidase subunit 2 [Source:HGNC Symbol;Acc:HGNC:315]	18	12328944	12377227
THG1L	tRNA-histidine guanylyltransferase 1 like [Source:HGNC Symbol;Acc:HGNC:26053]	5	157731420	157741449
GDAP1	ganglioside induced differentiation associated protein 1 [Source:HGNC Symbol;Acc:HGNC:15968]	8	74320613	74518007
PLD6	phospholipase D family member 6 [Source:HGNC Symbol;Acc:HGNC:30447]	17	17200995	17206333
OMA1	OMA1 zinc metallopeptidase [Source:HGNC Symbol;Acc:HGNC:29661]	1	58415384	58546802
STOML2	stomatin like 2 [Source:HGNC Symbol;Acc:HGNC:14559]	9	35099776	35103195
PARL	presenilin associated rhomboid like [Source:HGNC Symbol;Acc:HGNC:18253]	3	183829271	183884933
USP30	ubiquitin specific peptidase 30 [Source:HGNC Symbol;Acc:HGNC:20065]	12	109023089	109088023
FIS1	fission, mitochondrial 1 [Source:HGNC Symbol;Acc:HGNC:21689]	7	101239458	101252316
ZDHHC6	zinc finger DHHC-type palmitoyltransferase 6 [Source:HGNC Symbol;Acc:HGNC:19160]	10	112424428	112447572
MIGA2	mitoguardin 2 [Source:HGNC Symbol;Acc:HGNC:23621]	9	129036821	129072082
CHCHD3	coiled-coil-helix-coiled-coil-helix domain containing 3 [Source:HGNC Symbol;Acc:HGNC:21906]	7	132784870	133082090
MFN1	mitofusin 1 [Source:HGNC Symbol;Acc:HGNC:18262]	3	179347709	179394936
BNIP3	BCL2 interacting protein 3 [Source:HGNC Symbol;Acc:HGNC:1084]	10	131967684	131981967
HUWE1	HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1 [Source:HGNC	X	53532096	53686752

mart_export.txt

~/Downloads/mart_export.txt

Gene name	Gene description	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)
MTCH2	mitochondrial carrier 2 [Source:HGNC Symbol;Acc:HGNC:17587]	CHR_HG2114_PATCH	47617315	47647147
ADCK1	aarF domain containing kinase 1 [Source:HGNC Symbol;Acc:HGNC:19038]	14	77800109	77935014
AFG3L2	AFG3 like matrix AAA peptidase subunit 2 [Source:HGNC Symbol;Acc:HGNC:315]	18	12328944	12377227
THG1L	tRNA-histidine guanylyltransferase 1 like [Source:HGNC Symbol;Acc:HGNC:26053]	5	157731420	157741449
GDAP1	ganglioside induced differentiation associated protein 1 [Source:HGNC Symbol;Acc:HGNC:15968]	8	74320613	74518007
PLD6	phospholipase D family member 6 [Source:HGNC Symbol;Acc:HGNC:30447]	17	17200995	17206333
OMA1	OMA1 zinc metallopeptidase [Source:HGNC Symbol;Acc:HGNC:29661]	1	58415384	58546802
STOML2	stomatin like 2 [Source:HGNC Symbol;Acc:HGNC:14559]	9	35099776	35103195
PARL	presenilin associated rhomboid like [Source:HGNC Symbol;Acc:HGNC:18253]	3	183829271	183884933
USP30	ubiquitin specific peptidase 30 [Source:HGNC Symbol;Acc:HGNC:20065]	12	109023089	109088023
FIS1	fission, mitochondrial 1 [Source:HGNC Symbol;Acc:HGNC:21689]	7	101239458	101252316
ZDHHC6	zinc finger DHHC-type palmitoyltransferase 6 [Source:HGNC Symbol;Acc:HGNC:19160]	10	112424428	112447572
MIGA2	mitoguardin 2 [Source:HGNC Symbol;Acc:HGNC:23621]	9	129036621	129072082
CHCHD3	coiled-coil-helix-coiled-coil-helix domain containing 3 [Source:HGNC Symbol;Acc:HGNC:21906]	7	132784870	133082090
MFN1	mitofusin 1 [Source:HGNC Symbol;Acc:HGNC:18262]	3	179347709	179394936
BNIP3	BCL2 interacting protein 3 [Source:HGNC Symbol;Acc:HGNC:1084]	10	131967684	131981967
HUWE1	HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1 [Source:HGNC Symbol;Acc:HGNC:30892]	X	53532096	53686752
BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]	6	33572547	33580293
PRKN	parkin RBR E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:8607]	6	161347417	162727775
MUL1	mitochondrial E3 ubiquitin protein ligase 1 [Source:HGNC Symbol;Acc:HGNC:25762]	1	20499448	20508151
MTCH2	mitochondrial carrier 2 [Source:HGNC Symbol;Acc:HGNC:17587]	11	47617315	47642607
OPA1	OPA1 mitochondrial dynamin like GTPase [Source:HGNC Symbol;Acc:HGNC:8140]	3	193593144	193697811
MFN2	mitofusin 2 [Source:HGNC Symbol;Acc:HGNC:16877]	1	11980181	12015211
BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]	19	48954815	48961798
RCC1L	RCC1 like [Source:HGNC Symbol;Acc:HGNC:14948]	7	75027122	75074228
VAT1	vesicle amine transport 1 [Source:HGNC Symbol;Acc:HGNC:16919]	17	43014607	43025123
TFRC	transferrin receptor [Source:HGNC Symbol;Acc:HGNC:11763]	3	196012511	196082153
MIGA1	mitoguardin 1 [Source:HGNC Symbol;Acc:HGNC:24741]	1	77779624	77879540
MFF	mitochondrial fission factor [Source:HGNC Symbol;Acc:HGNC:24858]	2	227325151	227357836
PID1	phosphotyrosine interaction domain containing 1 [Source:HGNC Symbol;Acc:HGNC:26084]	2	228850526	229271287

The exported gene names and descriptions visualized on BBedit. We can see here that the unique number of genes involved in this biological process is 30.

Q2

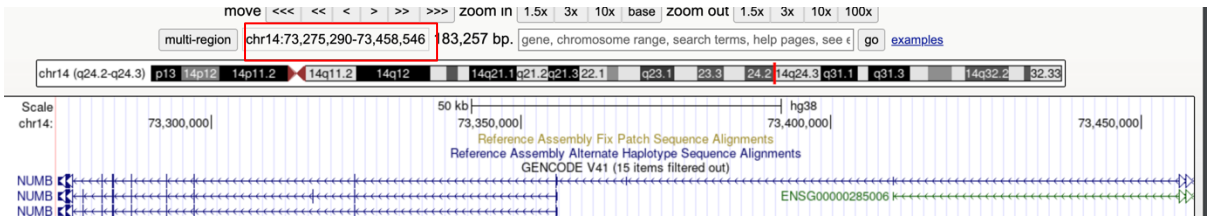
a.

From the USCS browser we got the information about on which chromosome and which along which base positions the gene is located. Also, we retrieved the number of transcript variants and their lengths. There is numerous information we can get from the browser by adding the tracks we need these include protein variants, SNPs, phenotypes, expression levels, repeats along the chromosome region, information about the region's conversation among species through alignment. Therefore, this is a very useful tool to retrieve information tailored to our purpose.

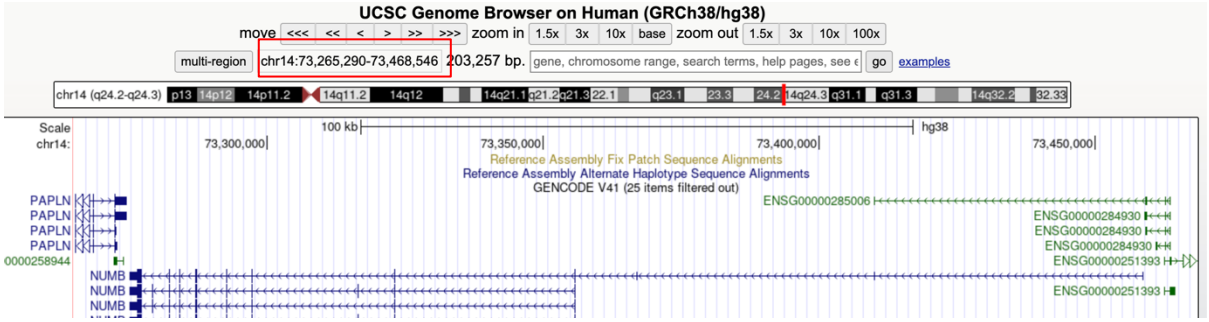
---

b.

The initial range:



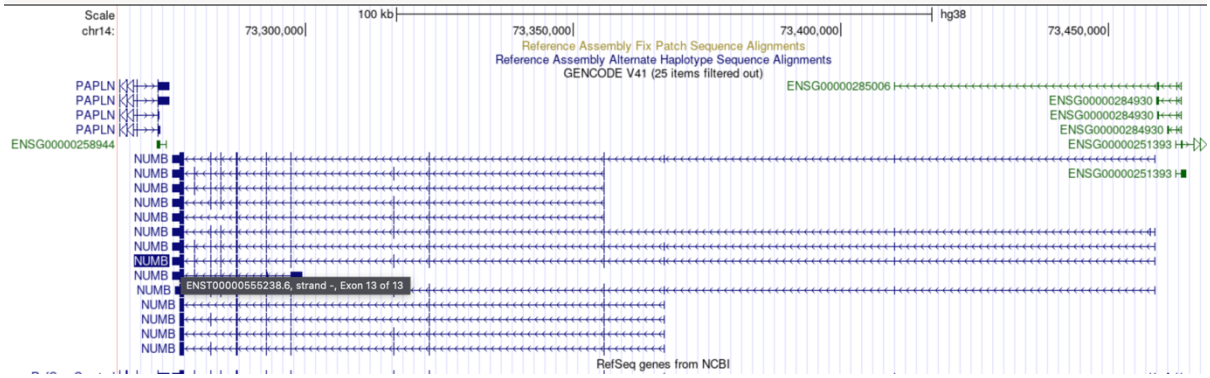
Range after extension of 1000 bases start and end:



---

c.

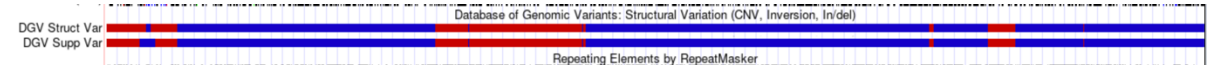
Transcript with code ENST00000555238.6 is highlighted:



Coding Region
Position: hg38 chr14:73,276,578-73,355,751 **Size:** 79,174 **Coding Exon Count:** 10

---

d.

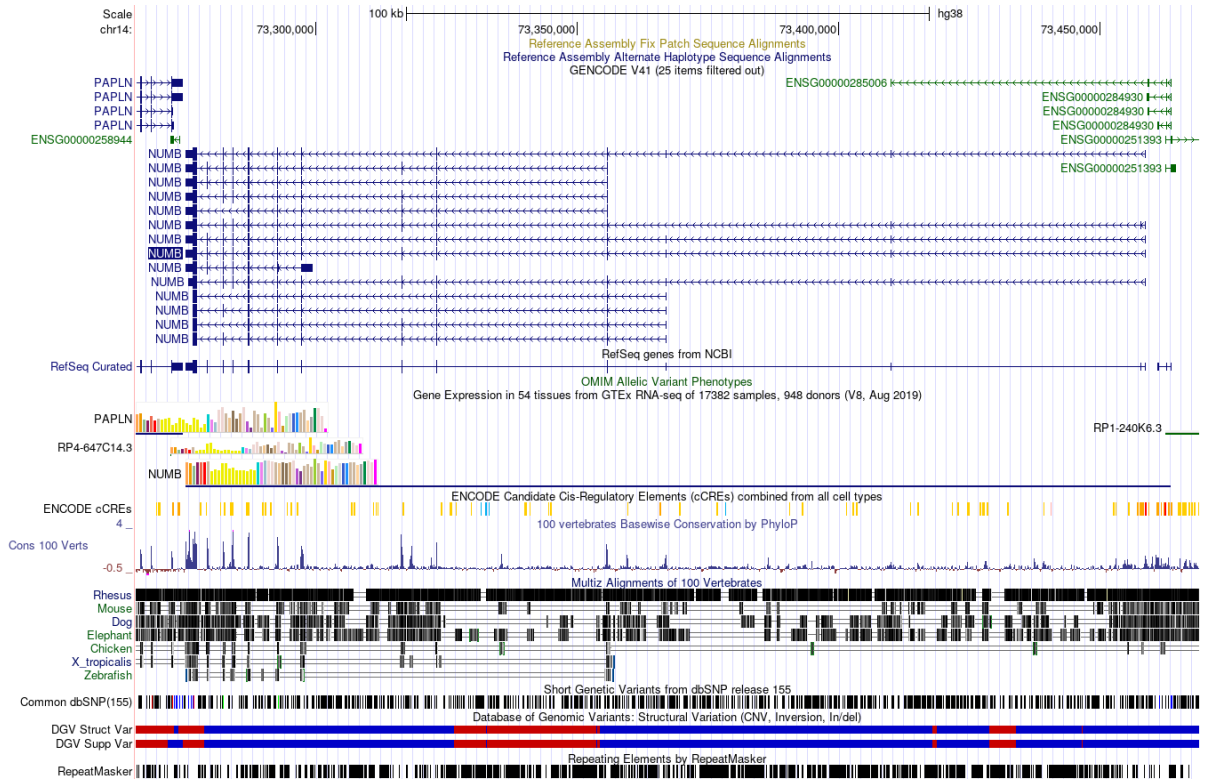


The colors indicate gain in size with blue and loss in size in red. In the extended region there are 16 variants in total.

Color is used in both subtracks to indicate the type of variation:

- **Inversions** and **inversion breakpoints** are purple.
- CNVs and InDels are blue if there is a **gain in size** relative to the reference.
- CNVs and InDels are red if there is a **loss in size** relative to the reference.
- CNVs and InDels are brown if there are reports of **both a loss and a gain in size** relative to the reference.

e.

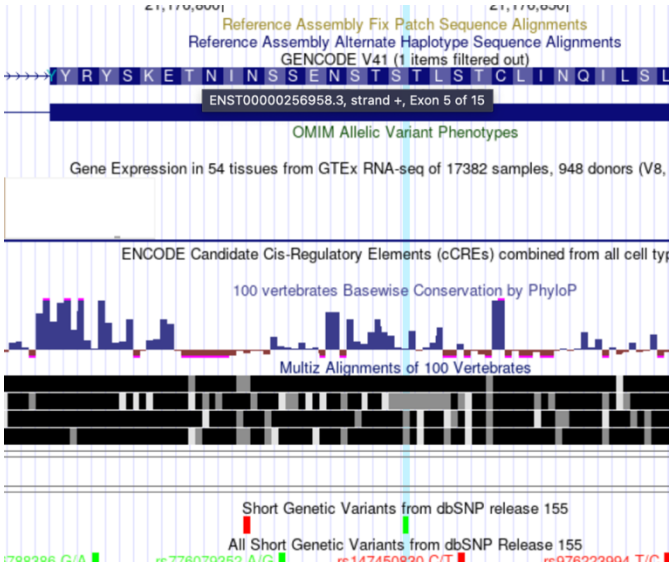


Q3

a.

The SNP rs11045818 is located on the exon5 of the gene SLCO1B1, then I fit the chromosome range to only fix the exon5:

Image link: https://genome.ucsc.edu/trash/hgt/hgt_genome_133a4_6f8790.png



Then I switched to the table browser and picked the position, to limit the search to exon 5:

Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on da
DNA sequence covered by a track. More...

Select dataset

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)
group: Variation **track:** All SNPs(151)
table: snp151 describe table schema

Note: Most dbSNP tables are huge. Trying to download them through the Table Browser usually leads to a timeout.
Please see our Data Access FAQ on how to download dbSNP data.

Define region of interest

region: genome ENCODE Pilot regions position chr12:21,329,710-21,329,831 lookup define regions
identifiers (names/accessions): paste list upload list

Then I used the filters for synonymous and non-synonymous variants:

func	does	include	*	unknown	coding-synon	intron	near-gene-3
			near-gene-5	ncRNA	nonsense	missense	stop-loss
			frameshift	cds-indel	untranslated-3	untranslated-5	splice-3
			splice-5				

This resulted in **39** synonymous and non-synonymous variants:

All SNPs(151) (snp151) Summary Statistics

item count	39
item bases	39 (31.97%)
item total	39 (31.97%)
smallest item	1
average item	1
biggest item	1

---

b. It corresponds to nucleotide substitution G->A, and amino acid substitution S->S:

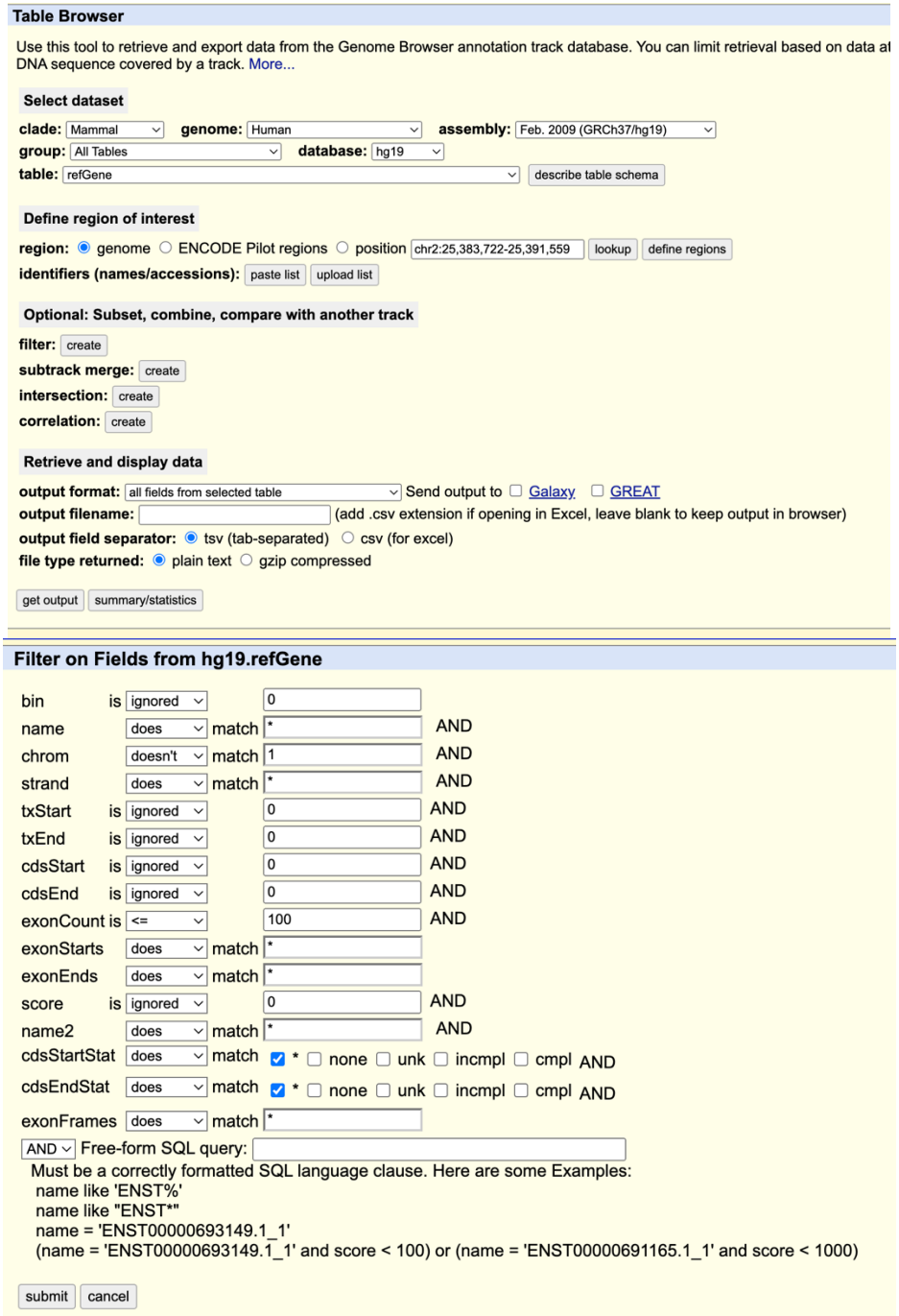
UCSC's predicted function relative to selected gene tracks:
GENCODE V41 SLCO1B1 (ENST00000256958.3) synonymous_variant S (TC**G**) --> S (TC**A**)
GENCODE V41 ENSG00000257062 (ENST00000543498.5) 3_prime_UTR_variant

View table schema

Q4

a.

This SQL query searches for the genes with exon count less than 100 and are not located on the first chromosome.



b.

The query returned **81,375** items:

refGene (refGene) Summary Statistics

item count	81,375
item bases	1,385,593,812 (46.31%)
item total	5,526,163,587 (184.72%)
smallest item	20
average item	67,910
biggest item	2,320,933
block count	832,617
block bases	90,525,373 (3.03%)
block total	271,202,292 (9.07%)
smallest block	2
average block	326
biggest block	91,671

Region and Timing Statistics

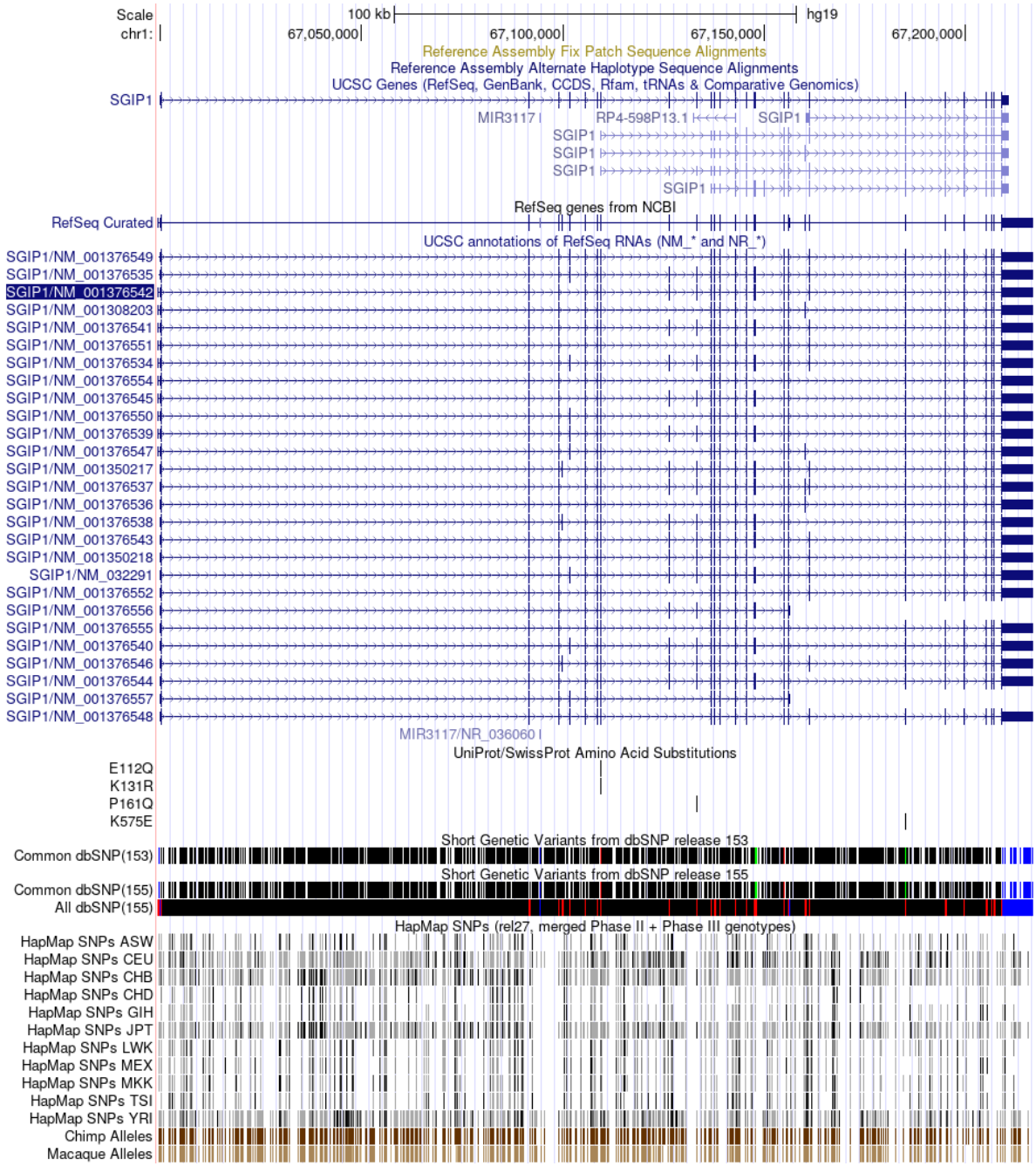
region	genome
bases in region	3,234,851,260
bases in gaps	243,140,514
load time	1.76
calculation time	1.56
free memory time	0.00
filter	on
intersection	off

A screenshot of the results:

```
#filter: refGene.exonCount <= 100 and not (refGene.chrom = '1')
#bin    name    chrom   strand  txStart txEnd   cdsStart        cdsEnd  exonCount       exonStarts      exonEnds        score   name2
0       NM_001376542    chr1    +       66999275        67216822        67000041        67208778        25
66999275,66999928,67091529,67098752,67105459,67108492,67109226,67126195,67133212,67136677,67137626,67138963,67142686,67145360,6714755
205017,67206340,67206954,67208755,
66999620,67000051,67091593,67098777,67105516,67108547,67109402,67126207,67133224,67136702,67137678,67139049,67142779,67145435,6714805
205220,67206405,67207119,67216822,      0       SGIP1   cmpl    cmpl    -1,0,1,2,0,0,1,0,0,0,1,2,1,1,1,1,0,1,1,2,2,0,2,1,1,
0       NM_001308203    chr1    +       66999275        67216822        67000041        67208778        22
66999275,66999928,67091529,67098752,67105459,67108492,67109226,67136677,67137626,67138963,67142686,67145360,67154830,67155872,6716012
208755,
66999355,67000051,67091593,67098777,67105516,67108547,67109402,67136702,67137678,67139049,67142779,67145435,67154958,67155999,6716018
216822, 0       SGIP1   cmpl    cmpl    -1,0,1,2,0,0,1,0,1,2,1,1,1,1,0,1,1,2,2,0,2,1,1,
0       NM_001376541    chr1    +       66999275        67216822        67000041        67208778        25
66999275,66999928,67091529,67098752,67105459,67108492,67109226,67126195,67133212,67136677,67137626,67138963,67142686,67145360,6714755
205017,67206340,67206954,67208755,
66999355,67000051,67091593,67098777,67105516,67108547,67109402,67126207,67133224,67136702,67137678,67139049,67142779,67145435,6714805
205220,67206405,67207119,67216822,      0       SGIP1   cmpl    cmpl    -1,0,1,2,0,0,1,0,0,0,1,2,1,1,1,1,0,1,1,2,2,0,2,1,1,
0       NM_001376549    chr1    +       66999043        67216822        67000041        67208778        23
66999043,66999928,67091529,67098752,67101626,67105459,67108492,67109226,67136677,67137626,67138963,67142686,67145360,67154830,6715587
```

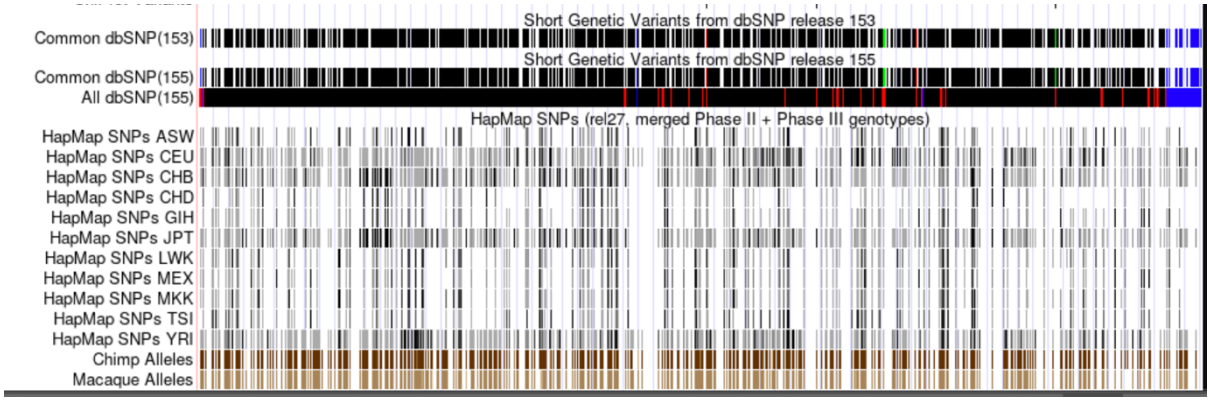
c.

I picked NM_001376542.



https://genome.ucsc.edu/trash/hgt/hgt_genome_27bec_6f42f0.png

d.



---

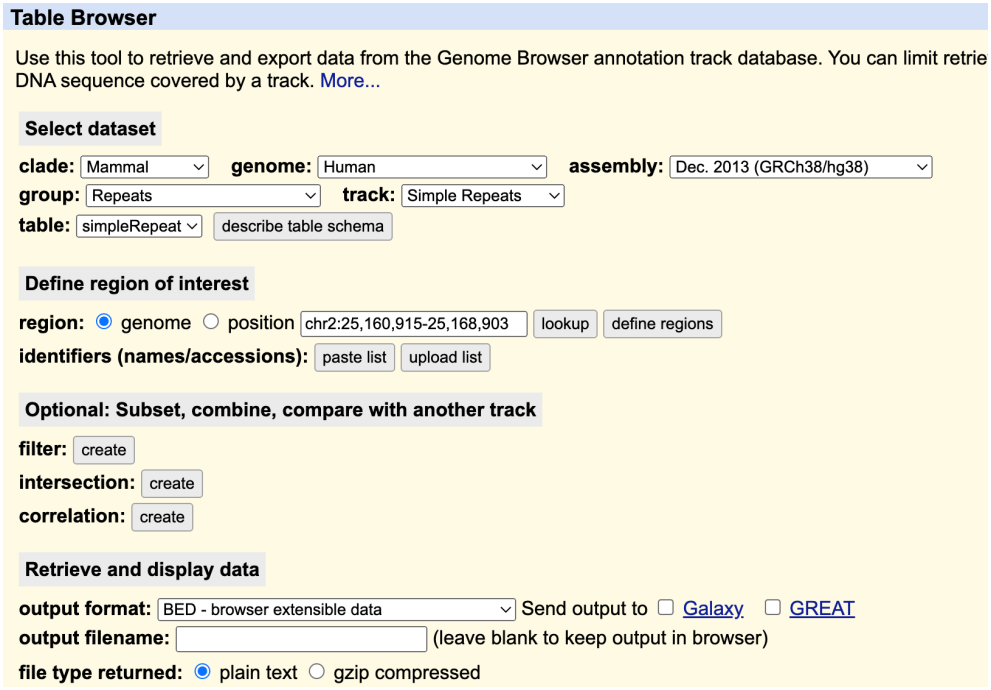
e.

There are three UniProt/SwissProt substituions in the selected transcript which are E112Q, K131R, P161Q, K575E.

Q5

a.

There are 1,039,174 simple repeats in the human genome (hg38). To get this information I picked the hg38 assembly and selected the group as “Repeats”, and track “Simple Repeats”.



Then I retrieved the results by clicking “summary/statics”.

Results:

Simple Repeats (simpleRepeat) Summary Statistics

item count	1,039,174
item bases	150,457,058 (4.84%)
item total	343,446,344 (11.04%)
smallest item	25
average item	330
biggest item	500,000
smallest score	50
average score	299
biggest score	803,864

Region and Timing Statistics

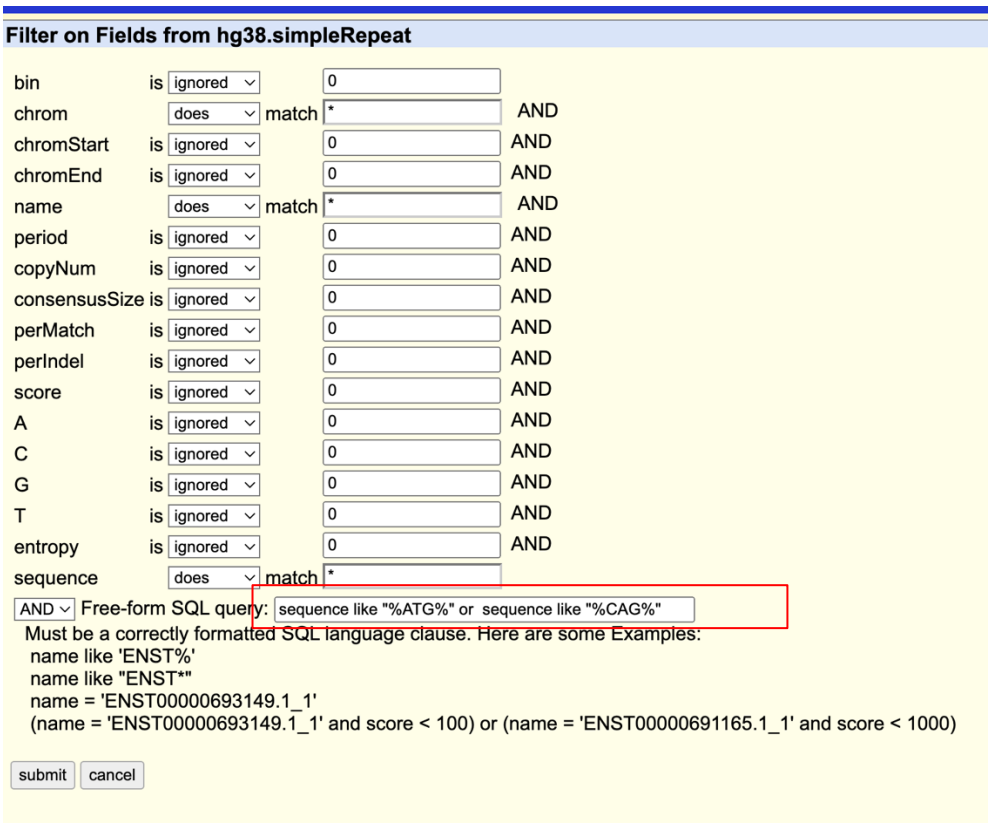
region	genome
bases in region	3,272,116,950
bases in gaps	161,348,343
load time	4.75
calculation time	0.52
free memory time	0.00
filter	off
intersection	off

---

b.

I used the same dataset selection as the previous question.

Then for filtering the sequences containing “ATG” or “CAG”, I wrote an SQL query as follows:



The results showed **309,841** such sequences:

Simple Repeats (simpleRepeat) Summary Statistics

item count	309,841
item bases	120,076,290 (3.86%)
item total	287,765,317 (9.25%)
smallest item	25
average item	929
biggest item	500,000
smallest score	50
average score	798
biggest score	803,864

Region and Timing Statistics

region	genome
bases in region	3,272,116,950
bases in gaps	161,348,343
load time	4.84
calculation time	0.50
free memory time	0.00
filter	on
intersection	off

---

c.

Again, I used the same dataset selection, and updated the filter as follows to retrieve the CG repeats on chromosome 1 and AG repeats on chromosome 2:

Filter on Fields from hg38.simpleRepeat

bin	is ignored		0	
chrom	does	match	*	AND
chromStart	is ignored		0	AND
chromEnd	is ignored		0	AND
name	does	match	*	AND
period	is ignored		0	AND
copyNum	is ignored		0	AND
consensusSize	is ignored		0	AND
perMatch	is ignored		0	AND
perIndel	is ignored		0	AND
score	is ignored		0	AND
A	is ignored		0	AND
C	is ignored		0	AND
G	is ignored		0	AND
T	is ignored		0	AND
entropy	is ignored		0	AND
sequence	does	match	*	

AND Free-form SQL query: (chrom = "chr1" and sequence = "CG") or (chrom = "chr2"

Must be a correctly formatted SQL language clause. Here are some Examples:
name like 'ENST%'
name like "ENST*"
name = 'ENST00000693149.1_1'
(name = 'ENST00000693149.1_1' and score < 100) or (name = 'ENST00000691165.1_1' and score < 1000)

submit cancel

My SQL query as the field doesn't fully show:

(chrom = "chr1" and sequence = "CG") or (chrom = "chr2" and sequence = "AG")

Then again, I visualized the results using the “summary/statistics” button, which showed **281** such repeats:

Simple Repeats (simpleRepeat) Summary Statistics

item count	281
item bases	13,375 (0.00%)
item total	13,375 (0.00%)
smallest item	25
average item	48
biggest item	566
smallest score	50
average score	73
biggest score	228

Region and Timing Statistics

region	genome
bases in region	3,272,116,950
bases in gaps	161,348,343
load time	1.07
calculation time	0.07
free memory time	0.00
filter	on
intersection	off

---

d.

a- SELECT COUNT(*)
FROM hg38.simpleRepeat

b- SELECT COUNT(*)
FROM hg38.simpleRepeat
WHERE sequence like “%ATG%” or sequence like “%CAG%”

c- SELECT COUNT(*)
FROM hg38.simpleRepeat
WHERE (chrom = "chr1" and sequence = "CG") or (chrom = "chr2" and sequence = "AG")