UCSC Genome browser exercise – MBV-INFX410

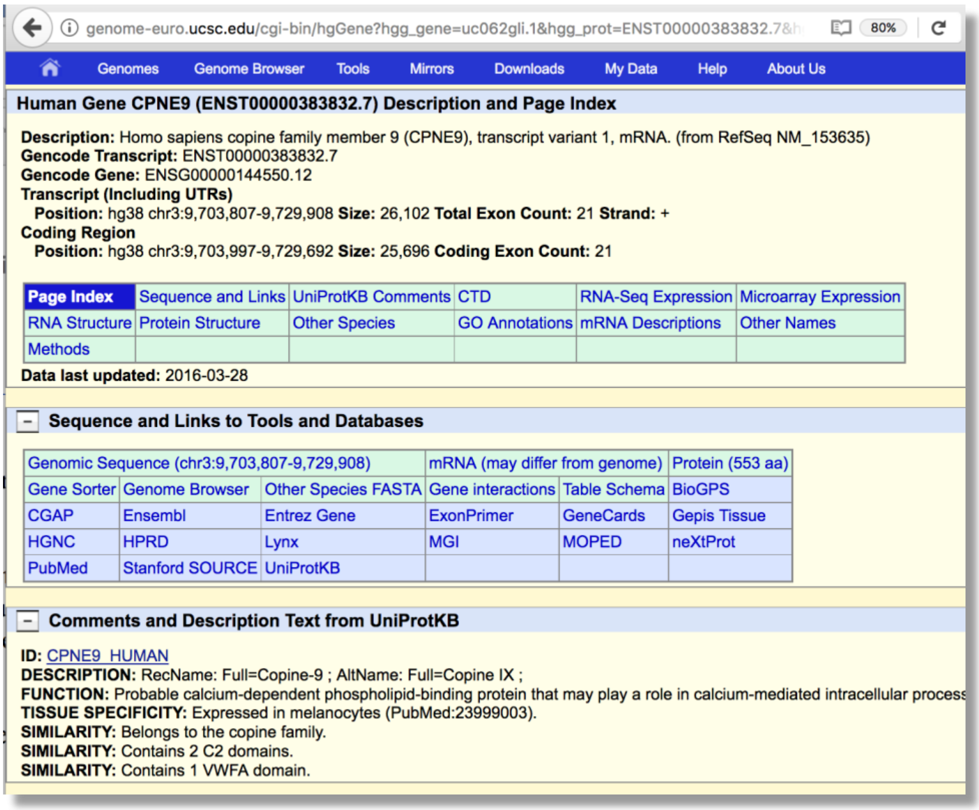
In this exercise you will very briefly look at some features of the UCSC Genome Browser (http://genome.ucsc.edu). Feel free to experiment and explore on your own!

1. Go to the UCSC Genome Browser (GB) website. Both along the top and at the bottom there are links to many tools and resources. In the top menu check out "Cite Us" under "About us" on how to cite the GB and the different tools. In the bottom menu under "Training" you find a lot of different tutorials.

2. At the front page, click on "Genomes" in the upper, left corner (you might be asked to select a different mirror site. Either is ok). You are now at the Genome Browser Gateway. Use the GRCh38 assembly of the human genome (this is the most recent release) and in the "Position/Search term" box type in "chr3:9,700,000-9,900,000". This means that you want to view chromosome 3, from base pair (bp) 9,700,000 to 9,900,000. Click the "Go" button.

   At which band on chr 3 is this region? (25.3) There are several genes in this region. List some of them. (MTMR14, CPNE9, BRPF1, OGG1, CAMK1, etc.) Zoom in on the CPNE9 gene (e.g. by selecting a portion at the top of the track, or by dragging and centring the gene and zooming in). How many exons are there in the CPNE9 gene? (20, or a few more)

3. Click on this RefSeq gene name or somewhere on the gene track (one of the blue top tracks. This will take you to the UCSC RefSeq Gene page for CPEN9.


genome-euro.ucsc.edu/cgi-bin/hgGene?hgg_gene=uc062gli.1&hgg_prot=ENST00000383832.7&h
80%
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us

Human Gene CPNE9 (ENST00000383832.7) Description and Page Index

Description: Homo sapiens copine family member 9 (CPNE9), transcript variant 1, mRNA. (from RefSeq NM_153635)
Gencode Transcript: ENST00000383832.7
Gencode Gene: ENSG00000144550.12
Transcript (Including UTRs)
Position: hg38 chr3:9,703,807-9,729,908 **Size:** 26,102 **Total Exon Count:** 21 **Strand:** +
Coding Region
Position: hg38 chr3:9,703,997-9,729,692 **Size:** 25,696 **Coding Exon Count:** 21

Page Index	Sequence and Links	UniProtKB Comments	CTD	RNA-Seq Expression	Microarray Expression
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Other Names
Methods					

Data last updated: 2016-03-28

− **Sequence and Links to Tools and Databases**

Genomic Sequence (chr3:9,703,807-9,729,908)			mRNA (may differ from genome)		Protein (553 aa)
Gene Sorter	Genome Browser	Other Species FASTA	Gene interactions	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	Gepis Tissue
HGNC	HPRD	Lynx	MGI	MOPED	neXtProt
PubMed	Stanford SOURCE	UniProtKB			

− **Comments and Description Text from UniProtKB**

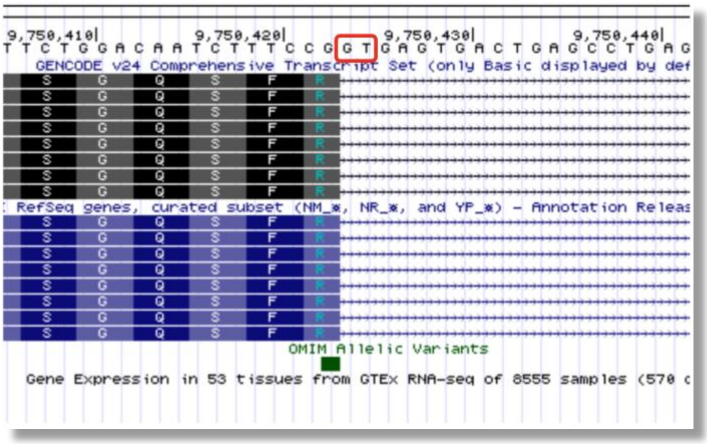
ID: CPNE9_HUMAN
DESCRIPTION: RecName: Full=Copine-9 ; AltName: Full=Copine IX ;
FUNCTION: Probable calcium-dependent phospholipid-binding protein that may play a role in calcium-mediated intracellular process
TISSUE SPECIFICITY: Expressed in melanocytes (PubMed:23999003).
SIMILARITY: Belongs to the copine family.
SIMILARITY: Contains 2 C2 domains.
SIMILARITY: Contains 1 VWFA domain.

Follow the Entrez Gene link in the second table to go to the Gene entry in the NCBI database. What is the “Official full name” of this human gene? (copine family member 9) Close this NCBI Gene window and again focus on the UCSC GB.

4. Go back to the UCSC GB Gateway by clicking on “Genomes” at the top of the webpage. This time you will search for a gene with an official gene identifier. Type “OGG1” in the “search term” box and press “Go” (Still using GRCh38 assembly of human genome).

   You get a lot of entries that match this search term. There are, for example, 8 different RefSeq curated genes. There are two main isoforms of human *OGG1*. The splice variant *α-OGG1* (that is isoform 1a, transcript NM_002542) encodes a nuclear protein with 345 amino acid residues, while the variant *β-OGG1* (that is isoform 2a) encodes a mitochondrial protein with 424 residues. Most likely the other 6 variants are “junk”, not doing anything particularly meaningful in human cells. You could have found this out by reading the literature on OGG1, but it is not usually obvious from the various sequence databases. This is an example of “noisy” or “wrong” data cluttering the databases and making it more difficult to find the useful information.

   Follow the RefSeq Gene link marked “OGG1 at chr3:9749944-9757407 – (NM_002542)”. In the Genome Viewer, zoom in on exon 1. Use both the “zoom in” buttons and the “drag select” to zoom option (zoom in until you see the nucleotides at the top). What is the sequence of the start codon? (ATG, as always, almost...) What is the position in the chromosome of the first protein coding nucleotide? (9,750,287) What are the last 9 nucleotides of the 5’ UTR? (GCTGTGGAA) What are the two first and last nucleotides of intron 1? (GT and AG)

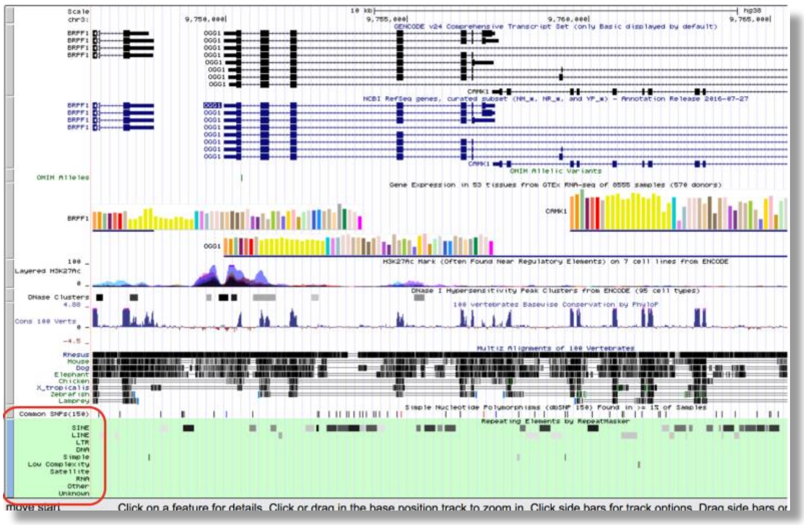

9,750,410| 9,750,420| 9,750,430| 9,750,440|
T T C T G G A C A A T C T T T C C G G T G A G T G A C T G A G C C T G A G
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by def
S G Q S F R
RefSeq genes, curated subset (NM_*, NR_*, and YP_*) - Annotation Releas
S G Q S F R
OMIM Allelic Variants
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 c

   Is it surprising to find GT and AG at the start and end of the intron? (No, most introns are GT-AG introns)

5. One codon is split between exons 1 and 2. What is this codon and which amino acid does it code for? (CGG = Arg) Codon table: http://en.wikipedia.org/wiki/DNA_codon_table

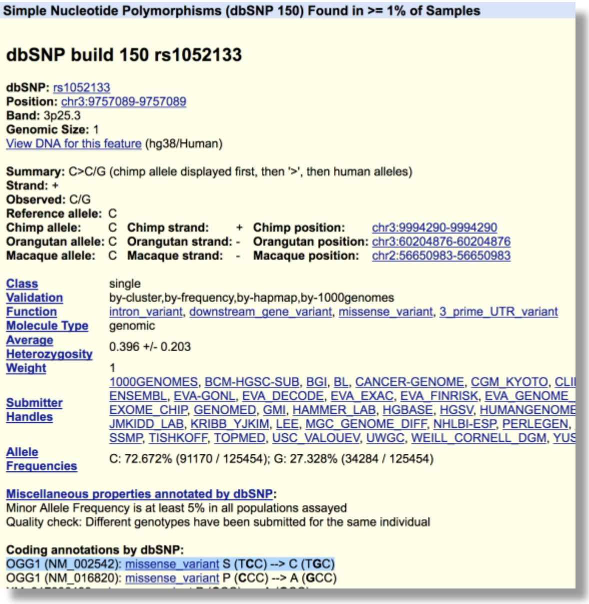
6. Zoom out again to see the full *OGG1* gene. Scroll down to the “Repeats” category and make sure that “RepeatMasker” is set to “full” view. Press “refresh”. Are there any

predicted repeating elements in *OGG1*? (The title of each track is located in the center of the screen. Look for "Repeating Elements by RepeatMasker. You can also drag and change the order of tracks) (SINEs in introns 3 and 4, possibly in introns 1 and 2. And in the last intron of the long variants between 9,800,000 and 9,807,000)


Scale
chr3:
10 kb
hg38
9,750,000
9,755,000
9,760,000
9,765,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
BRPF1
OGG1
CAMK1
NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*) - Annotation Release 2016-07-27
OMIM Alleles
OMIM Allelic Variants
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
4.88
Cons 100 Verts
100 vertebrates Basewise Conservation by PhyloP
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(150)
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
move start
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

7. Look at *α-OGG1* (this is the splice variant you clicked on in the search page to get to the Genome Viewer. It is highlighted in the RefSeq gene list with a solid background on the gene name (see previous image)). How many exons are there for this isoform? (7) Zoom in on the 3'-exon. Are there any common SNPs in this exon? (Yes, look at the "Simple Nucleotide Polymorphisms (dbSNP 150)" track. There is a red bar at position 9,757,089. If you can't see this track make sure it is activated in the menus below)

8. Click on the little red bar in the Common SNPs track to alter the display. What is the identifier for this SNP? (rs1052133) Click again on the SNP to go to the UCSC SNP page. NM_002542 is the *α-OGG1* transcript. Is this a silent variant? (No, it is a missense variant leading to a Ser (TCC) to Cys (TGC) mutation)



Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples

dbSNP build 150 rs1052133

dbSNP: rs1052133
Position: chr3:9757089-9757089
Band: 3p25.3
Genomic Size: 1
View DNA for this feature (hg38/Human)

Summary: C>C/G (chimp allele displayed first, then '>', then human alleles)
Strand: +
Observed: C/G
Reference allele: C
Chimp allele: C **Chimp strand:** + **Chimp position:** chr3:9994290-9994290
Orangutan allele: C **Orangutan strand:** - **Orangutan position:** chr3:60204876-60204876
Macaque allele: C **Macaque strand:** - **Macaque position:** chr2:56650983-56650983

Class	single
Validation	by-cluster,by-frequency,by-hapmap,by-1000genomes
Function	intron_variant, downstream_gene_variant, missense_variant, 3_prime_UTR_variant
Molecule Type	genomic
Average Heterozygosity	0.396 +/- 0.203
Weight	1
Submitter Handles	1000GENOMES, BCM-HGSC-SUB, BGI, BL, CANCER-GENOME, CGM_KYOTO, CLI ENSEMBL, EVA-GONL, EVA_DECODE, EVA_EXAC, EVA_FINRISK, EVA_GENOME_ EXOME_CHIP, GENOMED, GMI, HAMMER_LAB, HGBASE, HGSV, HUMANGENOME JMKIDD_LAB, KRIBB_YJKIM, LEE, MGC_GENOME_DIFF, NHLBI-ESP, PERLEGEN, SSMP, TISHKOFF, TOPMED, USC_VALOUEV, UWGC, WEILL_CORNELL_DGM, YUS
Allele Frequencies	C: 72.672% (91170 / 125454); G: 27.328% (34284 / 125454)

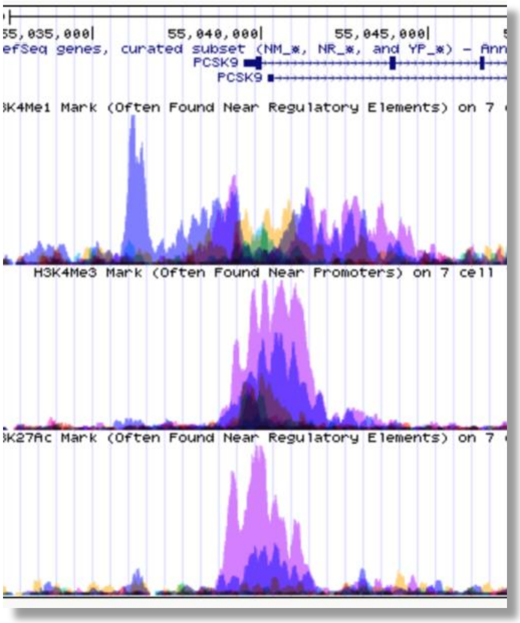
Miscellaneous properties annotated by dbSNP:
Minor Allele Frequency is at least 5% in all populations assayed
Quality check: Different genotypes have been submitted for the same individual

Coding annotations by dbSNP:
OGG1 (NM_002542): missense_variant S (T**C**C) --> C (T**G**C)
OGG1 (NM_016820): missense_variant P (**C**CC) --> A (**G**CC)

9. Experiment a bit more on your own. Move tracks up and down and add new tracks in various display formats. You can always go back to the default setup by clicking on the buttons marked "default tracks" and "default order". You can also change the look of the Genome Viewer by clicking "configure".

10. Search for the human gene PCSK9 in GRCh38 (NM_174936). At which chromosome band is this gene located (1p32.3). Center the start codon in the middle of the genome viewer window and zoom to see roughly 30 kbp around the start codon. Hide all tracks except "RefSeq Genes" and "ENCODE Regulation". Click on the "Integrated Regulation from ENCODE" track under the "Regulation" track set. Show H3K27Ac, H3K4Me1, and H3K4Me3 histone marks as "full" and click "submit". Right click on "Layered H3K4Me1" at the left and choose "Configure". Set "track height" to 100 and "Data view scaling" to "auto-scale to data view". Try out the various "Overlay methods". To view the results, click "submit". Also try to switch off some subtracks, for example, show only the embryonic stem cell line data (H1-hESC). Turn on all subtracks again and also modify the tracks for H3K27Ac and H3K4Me3 to get then same look for all three histone marks.

    Make a nice picture that shows the distribution of histone marks around the PCSK9 5' end, for example like this.


55,035,000|
55,040,000|
55,045,000|
efSeq genes, curated subset (NM_*, NR_*, and YP_*) - Ann
PCSK9
PCSK9
3K4Me1 Mark (Often Found Near Regulatory Elements) on 7
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell
3K27Ac Mark (Often Found Near Regulatory Elements) on 7

    H3K27Ac, H3K4Me1, and H3K4Me3 are all marks of "active chromatin" and found near promoters and other regulatory elements. Does it seem that th e PCSK9 gene is more "active" in any of the cell lines? (There are high levels of H3K4Me3 in NHEK and K562, but low in GM12878 and HUVEC, while the levels are intermediate in the 5 others. H3K4Me1 is high in K562, but also in NHEK and H1-hESC. It is low in GM12878. Also H3K27Ac is high in K562 and in particular in NHEK. The gene is apparently more active in NHEK cells, and slightly less in K562, but inactive in other cell lines, in particular GM12878.)

    You can also check out the expression tracks, e.g. Human mRNAs and "full", you will see mRNA reads mapped to the genome (spliced if they cross over introns).

Some more to do: For example, □

- Explore RYR2 (splicing, conservation, SNPs, and more).
- Explore LDLR or some other gene you are interested in (splicing, transcription factor binding, histone marks, CpG methylation, DNA methylation, DNaseI hypersensitivity clusters, and more). What are all these?

An excellent, free online tutorial can be found here: http://www.openhelix.com/ucsc