jonkl@medisin.uio.no

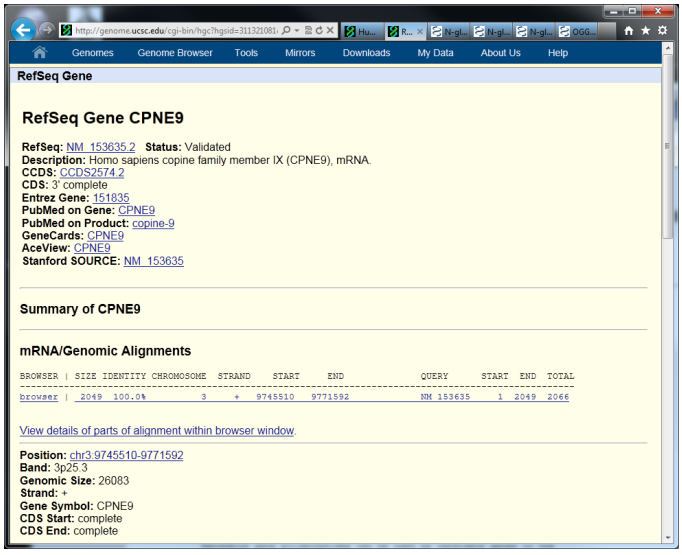
UCSC Genome Browser exercise – MBV-INFX410

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 left hand side there are links to many tools and resources. In the left hand menu check out "Cite Us" and "Training". What do you find following these links? (How to cite the UCSC GB in publications and tutorials & user guides)
2. At the front page, click on "Genomes" in the upper, left corner. You are now at the *Genome Browser Gateway*. Click on "Click here to reset the browser....". If this is the first time you are using the UCSC GB, nothing much will happen, but this is a way to reset everything and turn off strange or wrong settings if you make some in the future.

   Use ***the GRCh37 assembly*** of the human genome and in the "search term" box type in "chr3:9,700,000+200,000". This means that you want to view chromosome 3, base pair (bp) 9,700,000 and then 200,000 bp after that. Click the "submit" button.

   At which band on chr 3 is this region? (3p25.3) There are several genes in this region. List some of them. (MTMR14, CPNE9, BRPF1, OGG1, CAMK1, etc.) Zoom in on the CPNE9 gene. How many exons are there in the CPNE9 gene? (20, or a few more)
3. Click on this RefSeq gene, somewhere on an exon or intron. This will take you to the UCSC RefSeq Gene page for CPEN9.

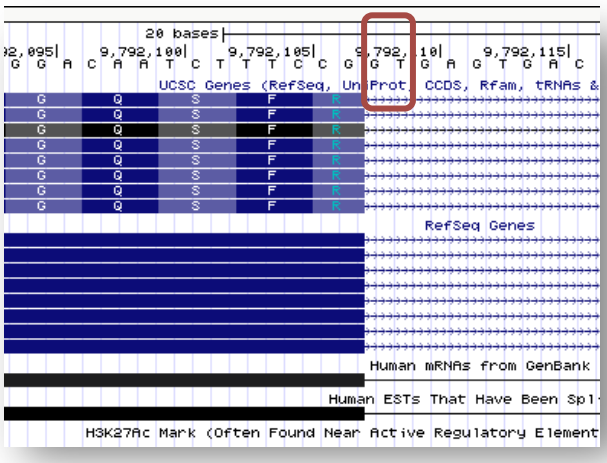


Follow the **Entrez Gene** link to go to the Gene entry in the NCBI database. What is the "Official full name" of this human gene? (copine family member IX) Close this NCBI Gene window and again focus on the UCSC GB.

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 "submit" (Still using GRCh37 assembly of human genome).

   You get a lot of entries that match this search term. There are, for example, 8 different RefSeq 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 chr3:9791628-9799089 - (NM_002542) N-glycosylase/DNA lyase isoform 1a". In the Genome Viewer, zoom in on exon 1. Use both the "zoom in" buttons and the "drag select" to zoom option. 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? (9,791,971) 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)

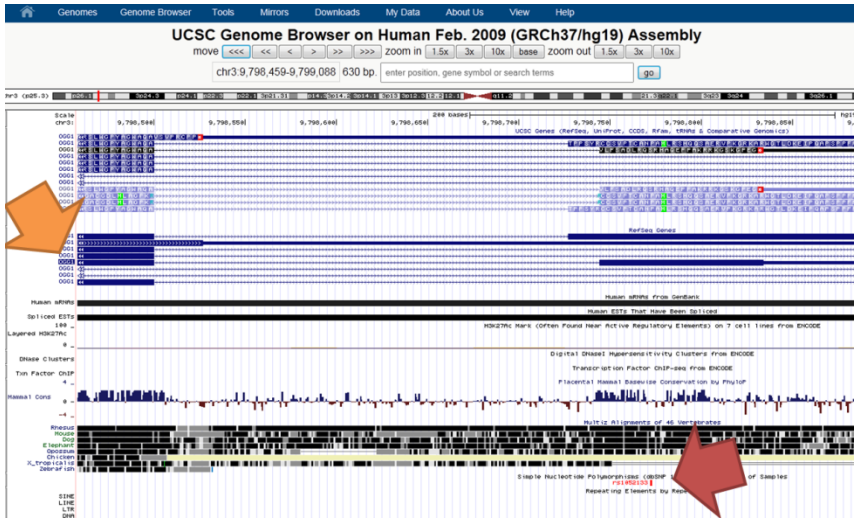


   Is it surprising to find GT and AG at the start and end of the intron? (No, most introns are GT-AG introns)

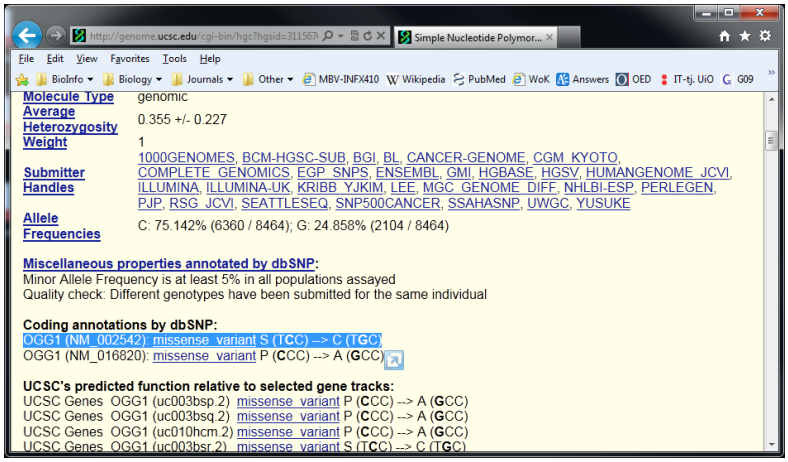
5. One codon is split between exons 1 and 2. What is this codon and which amino acid does it code for? (CGG = Arg)

Hint: Check here, 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 change "RepeatMasker" to "full" view. Press "refresh" to get this modification. Are there any predicted repeating elements in *OGG1*? (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)
7. Zoom in on the 3' exon of *α-OGG1* (that is isoform 1a, the last exon). 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, like this: OGG1 (orange arrow below). How many exons are there for this isoform? (7) Are there any common SNPs in this exon? (Yes, look at the "Common SNPs" track. There is a red bar at position 9,798,773. See red arrow below)

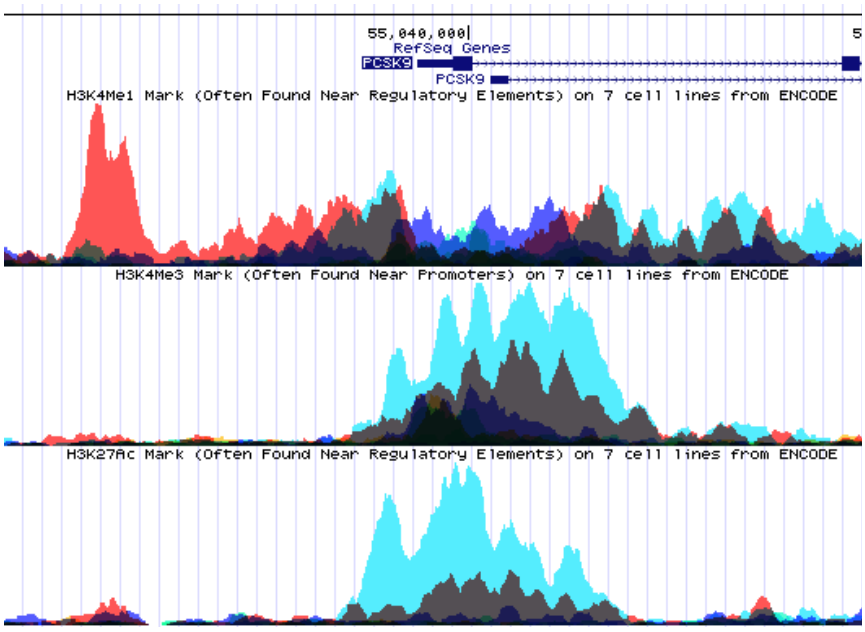


8. Click on the little red bar in the Common SNPs track to alter the display. What is the identifier for this SNP? (rs1052133) Click again on the SNP to go to the UCSC SNP page. NM_002542 is the *α-OGG1* transcript. Is this a silent variant? (No, it is a missense variant leading to a Ser (TCC) to Cys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



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” and direction with “reverse”.
10. Search for the human gene PCSK9 in GRCh38. Choose to display the NM_174936 RefSeq transcript (mapped to the genome). 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 “Layered H3K27Ac”. 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”. 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.



    H3K27Ac, H3K4Me1, and H3K4Me3 are all marks of “active chromatin” and found near promoters and other regulatory elements. Does it seem that the 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

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 from ENCODE (RNA-Seq, Caltech RNA-seq, apparently only available for the GRCh37 genome). Do the RNA-Seq results correlate with the levels of the active histone marks found above?

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?

11. From the MBV-INF4410/9410 exam in 2012 (note that "most recent" below means "most recent in 2012"!): The zebra finch ortholog of human *OGG1* is found on chromosome 12 and the Ensembl identifier for this gene is ENSTGUG00000008997. Use both the UCSC and Ensembl genome browsers to find the answers to the following questions: What is the most recent genome assembly available for the zebra finch in these genome browsers? How many Ensembl transcripts are there for this zebra finch gene and what are their identifiers (Transcript IDs)? Are there any gaps in the most recent zebra finch genome assembly within 1000 base pairs of ENSTGUG00000008997? How many, and what are their locations/positions? If you take a screenshot of a genome browser with the relevant information, this might make your explanations better and easier to understand.
12. An excellent, free online tutorial can be found here:

    http://www.openhelix.com/ucsc