Ensembl Browser Workshop

18th-20th May 2021

Link to Training material

helpdesk@ensembl.org

Welcome to the ‘Living Document’ for this Ensembl Browser Workshop. Everyone who is registered for the course has access to edit this document. Please feel free to use this document to ask questions to the Ensembl team throughout the workshop. If you wish to ask questions privately, please do not hesitate to contact the Ensembl Helpdesk.

The ‘Living Document’ is a great way of capturing the knowledge exchanged during the course and saving it for future use by yourselves and those who can’t attend this course. Remember - you don’t have to contribute, but any additions will be welcomed!

By the end of the course you will be able to:

- Navigate Ensembl using the web browser platform
- Access main data types in Ensembl: genomes, genes, genetic variation, and comparative genomics
- Annotate your own variant data with the Variant Effect Predictor (VEP)
- Mine Ensembl data using BioMart
- Visualise your own data in the Ensembl genome browser

Course Overview

Tuesday 18th May 2021
Introduction to Ensembl.
Ensembl Genes and Transcripts

Wednesday 19th May 2021
Ensembl Variation and the Variant Effect Predictor (VEP)
Comparative genomics

Thursday 20th May 2021
Features that regulate gene expression
Custom export of Ensembl data with BioMart
Visualising your own data in Ensembl

Trainers:

Michal Szpak, Aleena Mushtaq

Slides and Coursebook (demos and exercises) available to download:
https://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May

Feedback survey:
https://forms.gle/1g9ZLXc81DKKECBP7

We would really appreciate it if you could share your thoughts with us regarding these sessions. We are interested in your opinions, how you feel the experience has benefited you and how it could be improved. If you could find a few minutes to complete a short survey at the end of the last session it would really help us in improving the training we can deliver.

The golden rules for efficient online training
We suggest everyone follows these few simple rules for the course to run as smoothly as possible:

- Mute all your microphones - only unmute when tutor asks you to do that
- If you have questions during the course:
    - If you would like to ask a question to all the class: type your question in this shared reference document. Avoid typing in the Zoom chat as it is difficult to keep track there
- When you are ready to move on from a practical/exercises please click the yes button 

so that the tutor knows you are ready and can proceed with the course. Please remember there are more exercises than we have time for. You are not expected to complete all of them, but rather pick and choose the ones most relevant to your work and you are welcome to finish them in your own time.

Resources:
www.ensembl.org - Ensembl genome browser (chordates)
grch37.ensembl.org - Ensembl archive for browsing data associated with the human GRCh37 genome assembly
www.ensemblgenomes.org - Ensembl Genomes genome browser (non-chordates)
www.ebi.ac.uk - EMBL-EBI website
www.ebi.ac.uk/services - EMBL-EBI databases and tools
Train Online - EMBL-EBI e-learning for our databases and tools

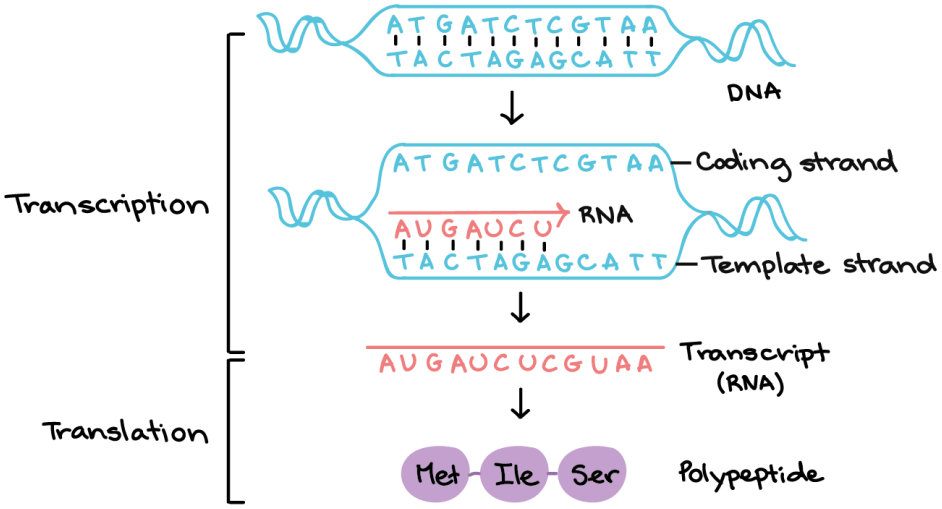
Questions

If you have any questions/problems that you would like to share and are applicable to the whole class please write them below. A tutor will answer your question.

Write your question after the last one you can see in this document and write your name.

1. **<Name><Why are genes strand specific?>**
   <Michal><Genes are transcribed from a specific DNA strand, so called template strand. The gene 'sense strand' is complementary to the template strand and matches the sequence of transcribed mRNA. Please note that translation of complementary strands would produce totally different proteins.>
   <<<<<I'm sorry if this is an obvious question but why when it is the complementary strand?>

   <I hope this explains it a bit more:



   You will get a different protein product depending on which strand is transcribed and translated.>

<It does, thank you very much :-) >

<You're most welcome :D >

2. **<Name><How can you analyse the effects of intronic variants when they don't encode proteins?>**
   <Michal><CADD (Combined Annotation Dependent Depletion) combines multiple annotations to predict the effect of nucleotide substitutions (including non-coding variants) incorporating properties like conservation across taxa, allele frequency and others. You can find more here: https://cadd.gs.washington.edu/>

3. **<Name><What exactly is Glorbal minor allele frequency? Why is it useful how does it very from Major?>**
   <Michal><global minor allele frequency is a frequency of the minor allele. It tells you how rare it is across the globe. It also implies the major allele frequency at the same time, as minor allele frequency + major allele frequency = 1. Allele frequency is often reported using the minor allele, as the community is often more interested in the rare alleles, which tend to be more deleterious than the common alleles.>

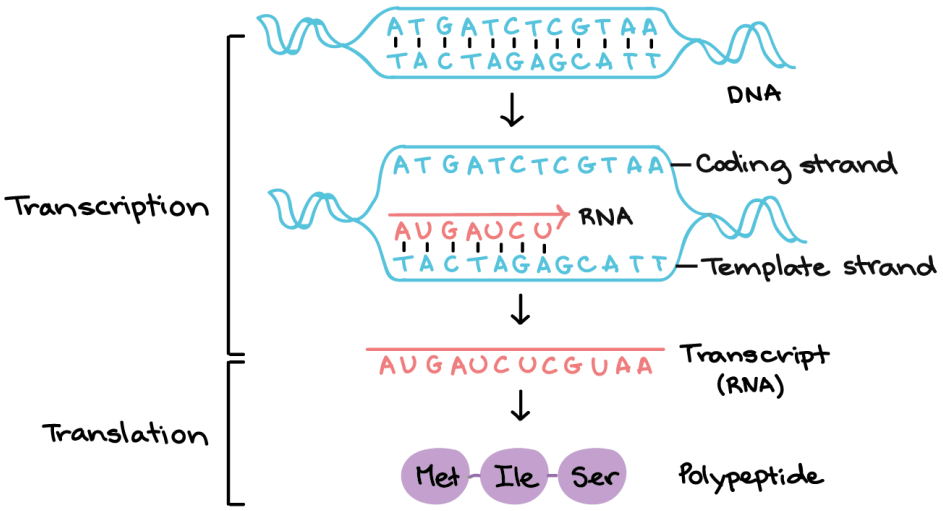
4. **<Name><G/A/C|Ancestral: G|MAF: 0.16 (A)|Highest population MAF: 0.49 - Why three alleles? And why are they not showing up int he population genetics part?>**
   <Michal><There are many multialleleic variants across the genome with more than two alleles. Often the third allele is very rare. If it's not reported in a given population genetic study, it means it haven't been detected and it's frequency was 0.>

   Much appreciated for all the answers :-)

   No worries at all, that's what we're for :)

5. **<Name><from yesterdays session, one exercise asked to check for the number of proteins in the 3 D structure.. Can you point how to get that?>**
   <Michal><I'm not sure which exercise you're referring to, but if it's GT2, then you first have to find the relevant gene, choose the MANE Select transcript from the Transcript Table. You can then find the 3D protein view in the menu on your left: http://www.ensembl.org/Homo_sapiens/Transcript/PDB?db=core;g=ENSG00000171759;r=12:102836889-102958410;t=ENST00000553106>

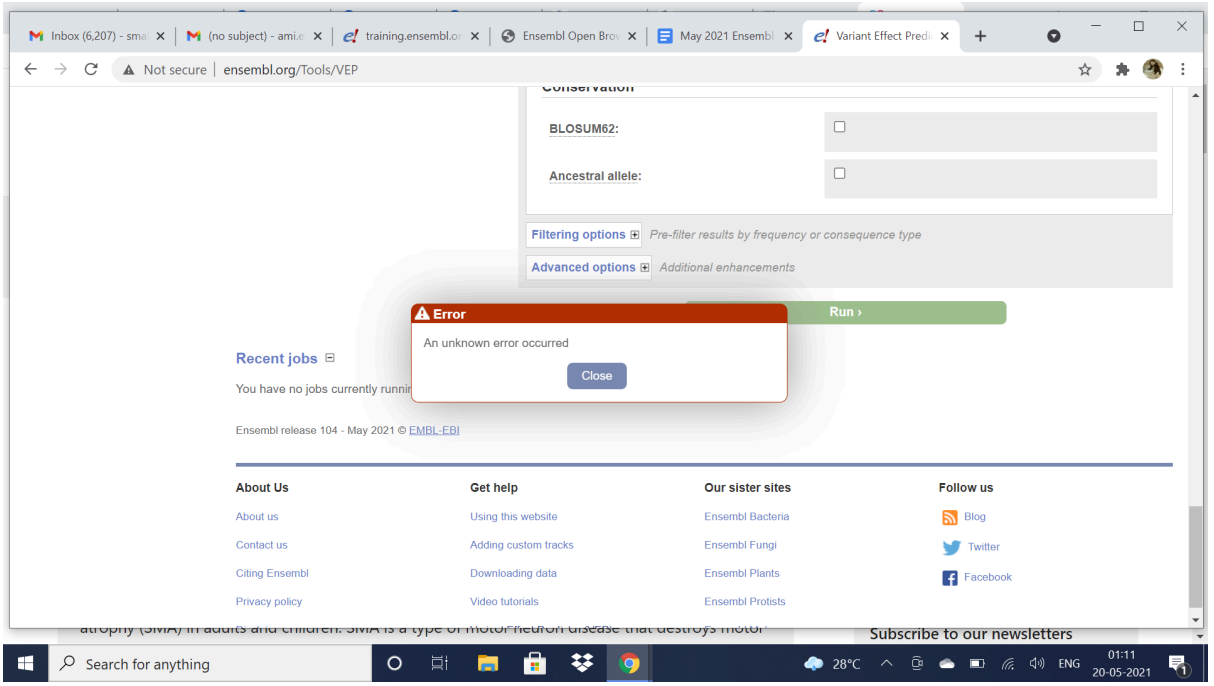
6. **<Name><MTHFR gene is located on the reverse strand - why is it only on the reverse?>**
   <Michal><It's coding sequence (sense) is on the reverse strand - that's how it evolved. I hope this diagram will help you understand it:

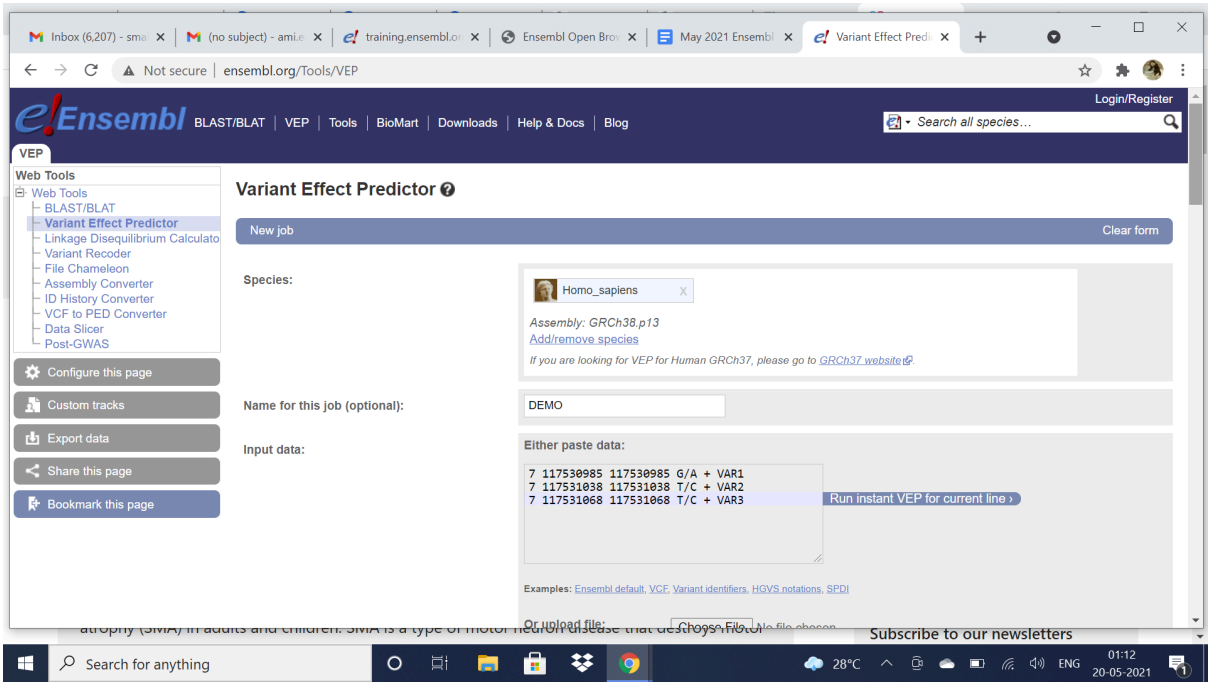


Also, please have a look at the first question above.>

7. **<Name><How is VEP uploaded as a track?>**
   <Michal><Shown on the screen - simply click on the genomic coordinates in the VEP result table to go to the location tab with your variants plotted along the genome.>

8. **<Name><**





>

<Michal><Interesting… looks fine to me. Have you tried refreshing the page? I just run your data on my machine and it runs fine.>
<YUP .. worked after refreshing ..thank you>

<Glad it's working now. Seems some technical glitch.>

9. **<Name> BLASTz/LASTz alignments I can't find this**
   <Michal><Please navigate to the Location Tab, then go to Configure this page -> Comparative genomics section and BLASTz/LASTz alignments

   Got it, thanks. I went to the side part and clicked configure this page - totally off

   @Michal and @Aleena Thank you :-)

10. **<Name><can you help me with where to find " constrained element"?>**
    <Michal><Please navigate to the Location Tab, then go to Configure this page -> Comparative genomics section and Conservation regions>

    Got it .. thank you !

11. <Name><

Gene & Transcript	Variant	Location	Zygosity	Disorder (OMIM)	Inheritance	Classification
MRPL44 NM_022915.5	c.748G>C (p.Val250Leu)	Exon 3	Homozygous	?Combined oxidative phosphorylation deficiency 16	Autosomal Recessive	Uncertain Significance
ABCC6 NM_001171.6	c.1762A>G (p.Ile588Val)	Exon 13	Homozygous	Pseudoxanthoma elasticum	Autosomal Recessive	Uncertain Significance

VARIANT INTERPRETATIONS

MRPL44 chr2:224828572G>C - Uncertain Significance.
A homozygous missense variant (c.748G>C) in exon 3 of the *MRPL44* gene that results in the amino acid substitution from Valine to Leucine at codon 250 (p.Val250Leu) was identified. The observed variant is present in both the 1000 Genomes and gnomAD databases with the minor allele frequency of 0.1022% and 0.0098%. The reference base is not conserved across the species and *in-silico* predictions by Polyphen is tolerated and SIFT is damaging. The Missense Badness Score and MPC value for this variant are 0.34 and 0.14, respectively. Missense Badness Score is the normalized fold difference of missense substitutions between observed and expected variants from ExAC dataset. This score is then combined with orthogonal deleteriousness metrics into one score called MPC (for Missense badness, PolyPhen2, and Constraint) designed to classify whether a missense variants is deleterious. Variants with MPC ≥ 2 have a rate nearly 6 times higher in cases than in controls. While those with intermediate MPC values (1 ≤ MPC < 2) have a more modest excess in cases. **Based on the above evidence this variant has been classified as Variant of uncertain significance according to the ACMG guidelines.**

Combined oxidative phosphorylation deficiency-16 (COXPD16) is caused by homozygous mutation in the *MRPL44* gene on chromosome 2. This disorder ischaracterized by pale hypertrophic heart with diffuse, mild microvesicular steatosis in cardiomyocytes, fatty degeneration in liver, mild microvesicular fatty degeneration, and elevated liver enzymes and mild granular pigmentation in the retina.

ABCC6 chr16:16282705T>C - Uncertain Significance.
A homozygous missense variant (c.1762A>G) in exon 13 of the *ABCC6* gene that results in the amino acid substitution from Isoleucine to Valine at codon 558 (p.Ile588Val) was identified. The observed variant is not present in both the 1000 Genomes and gnomAD databases. The reference base is conserved across the primates and *in-silico* predictions by Polyphen and SIFT are tolerated. The Missense Badness Score and MPC value for this variant are 0.19 and 0.06, respectively. Missense Badness Score is the normalized fold difference of missense substitutions between observed and expected variants from ExAC dataset. This score is then combined with orthogonal deleteriousness metrics into one score called MPC (for Missense badness, PolyPhen-2, and Constraint) designed to classify whether a missense variants is deleterious. Variants with MPC ≥ 2 have a rate nearly 6 times higher in cases than in controls. While those with intermediate MPC values (1 ≤ MPC< 2) have a more modest excess in cases. **Based on the above evidence this variant has been classified as Variant of uncertain significance according to the ACMG guidelines.**

Can you help me how i can check these variants on vep? I tried but got a completely different gene on using the coordinates (2 224828572 G/C). Since the transcript is not from Ensembl i think so it would be different?>
<Michal><I think your coordinates are for GRCh37, rather than GRCh38. To run VEP on GRCh37, please visit:
http://grch37.ensembl.org/info/docs/tools/vep/index.html
Always make sure which assembly your coordinates refer to, otherwise you might get into troubles.>

Its actually GRCh38.. I did check that ;)

<So I just double checked, as your variants were reported on RefSeq transcripts (NM_022915.5). This transcript coordinates on GRCh38 are:

RefSeq transcript NM_022915.5
Location Chromosome 2: 223,957,463-223,967,714

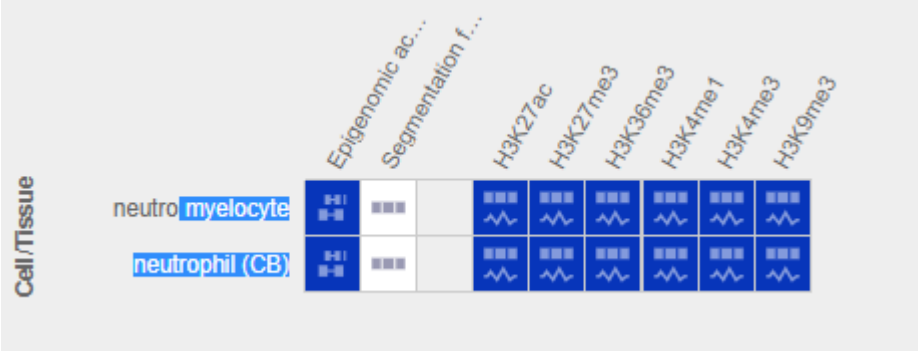
Your position '2:224828572' is far far away. It cannot be GRCh38 - maybe someone made a mistake in the report you're using?
Also, when using VEP, you can specify if you want your variants to be reported on Ensembl (GENCODE) transcripts, or RefSeq. I'd suggest using RefSeq in that case for consistency.

Here are the VEP results using GRCh37 and RefSeq transcripts:
http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/Results?tl=kwzMDgnyHeVZUI2t-7327078

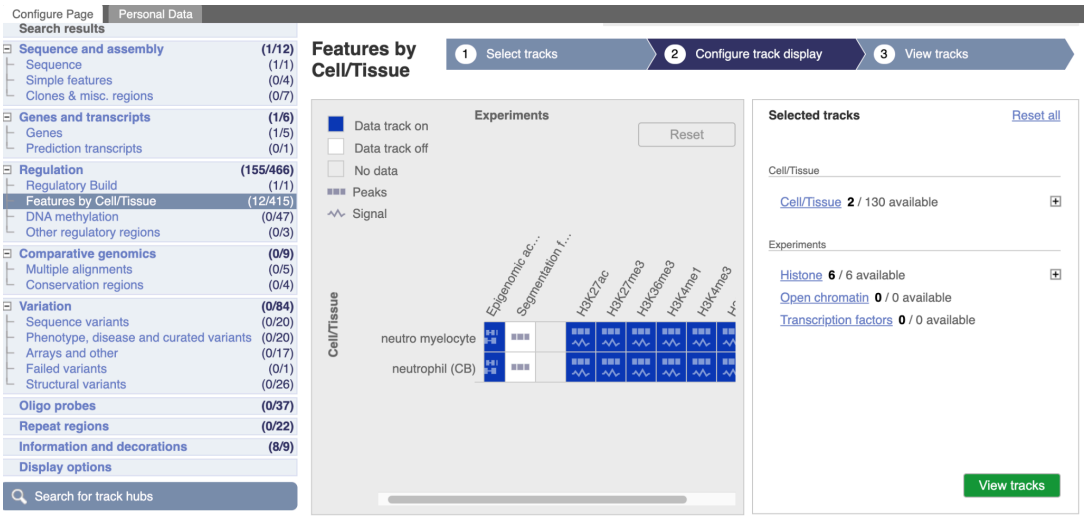
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
2_224828572_G/C	2:224828572-224828572	C	missense_variant	MODERATE	MRPL44	65080	Transcript	NM_022915.5	protein_coding	3/4	758	748	250	V/L	GTT/CTT	rs569652787	1

Which matches your variant: V/L, c.748, genomic 2:224828572, same transcript
>Thanks a ton.. Will recheck with the lab!

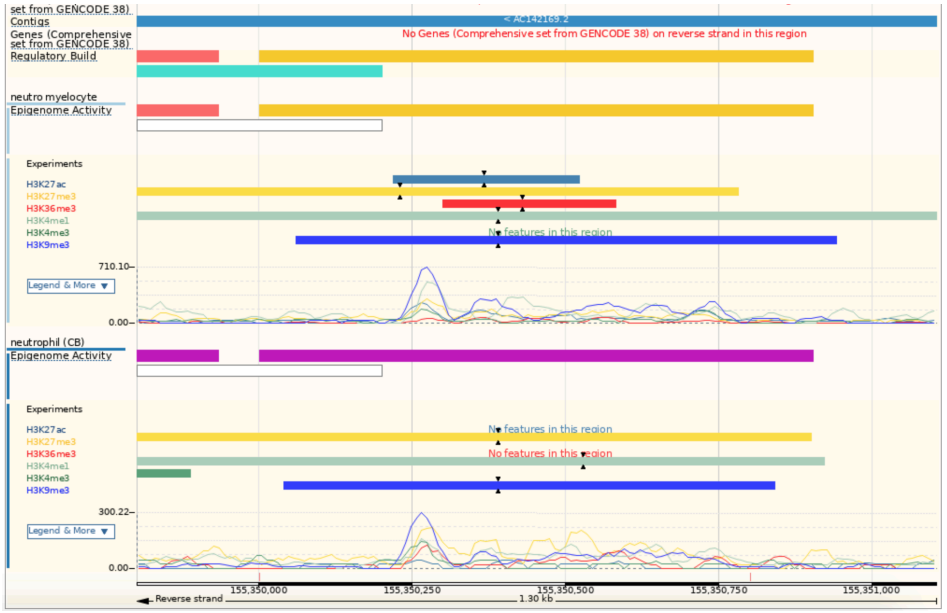
12. **<Name><In relation to exercise 1 does this look right. This looks like I have modifications in all>**



<Michal>This is just data availability view. Please click on 'View tracks' to see how the histone modifications look like in these two cell lines.



You can see the epigenetic differences here:

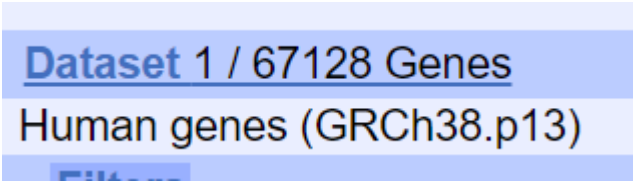


You can see that neutrophil (CB) line is missing the red (H3K36me3) and blue (H3K27ac) histone modifications

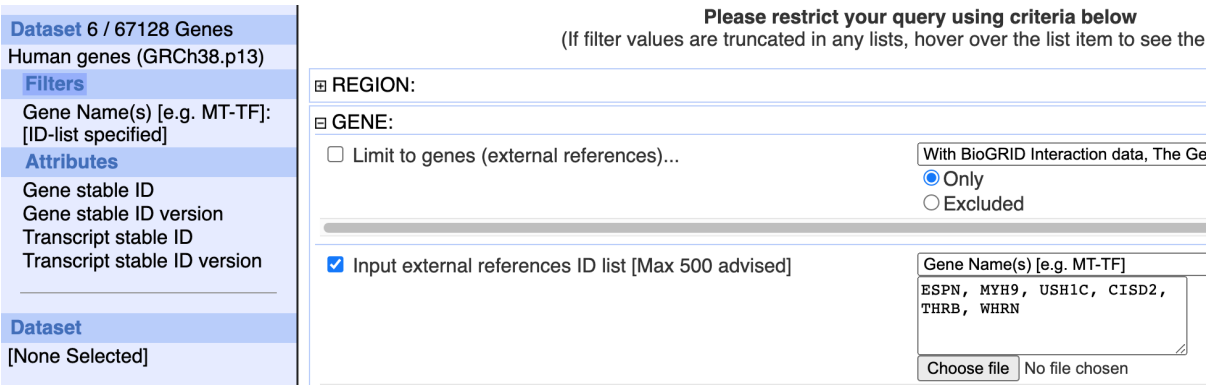
13. **<Name><Can't find the motifs track>**
    <Michal><You can see the motif tracks on the Regulation Tab Summary view, e.g.
    http://www.ensembl.org/Homo_sapiens/Regulation/Summary?db=core;fdb=funcgen;r=7:155349001-155351801;rf=ENSR00001133586
    >

14. **<Name><I'm really sorry, I put in the 6 genes, took away the space, checked it and I still only get the 1 gene.**

15. 



    <Michal><Have you updated the count, by clicking on the 'Count' button again?>



    Yes did that, refreshed page and redid it - still same.

    <Could you send me a screenshot of the whole window?>

16. **<Name><Question>**
    <Tutor><Answer>

17. **<Name><Question>**
    <Tutor><Answer>

18. <Name><Question>
    <Tutor><Answer>

19. <Name><Question>
    <Tutor><Answer>

20. <Name><Question>
    <Tutor><Answer>

21. <Name><Question>
    <Tutor><Answer>

22. <Name><Question>
    <Tutor><Answer>

23. <Name><Question>
    <Tutor><Answer>

24. <Name><Question>
    <Tutor><Answer>

25. <Name><Question>
    <Tutor><Answer>

26. <Name><Question>
    <Tutor><Answer>

27. <Name><Question>
    <Tutor><Answer>

28. <Name><Question>
    <Tutor><Answer>

29. <Name><Question>
    <Tutor><Answer>

30. <Name><Question>
    <Tutor><Answer>

31. <Name><Question>
    <Tutor><Answer>

Other resources

This section of the Living Document provides additional resources that might be useful to you in developing skills and knowledge in the course topic area.

Train online

EMBL-EBI provides an e-learning platform called Train online. Train online provides free courses on Europe's most widely used data resources, created by experts at EMBL-EBI and collaborating institutes. You do not need to have any previous experience of bioinformatics to benefit from this training. We want to help you to become a confident user of our data resources; we are not trying to train you to become a bioinformatician.

Tutorials of interest might include:

- **Ensembl: Quick Tour**
- **Ensembl: Browsing Genomes**
- **Ensembl Genomes (non-chordates): Quick Tour**
- **Ensembl REST API**

Webinar series

The EMBL-EBI training team also run regular webinar series featuring the EBI resources. See the Training pages for more information. You can also catch up on any webinars that you might have missed in Train online or on Youtube.