Showing your Data on Ensembl – Exercises

by Bert Overduin

These exercises are designed to teach you how to upload or attach your own data files to the website (http://www.ensembl.org). A range of upload options are explored, including GFF, BigWig, BAM and BED file formats. Sample upload files are provided.

Note: the answers to these exercises were composed in release version 70: http://e70.ensembl.org/index.html.

Please report any discrepancies with more current versions by contacting helpdesk@ensembl.org

Exercise 1 – Attaching a GFF file (human)

Have a look at the following file:

http://www.ebi.ac.uk/~bert/n-scan_genes.gff

It contains annotations for three transcripts of the human *HFE* gene (ENSG00000010704) generated by the N-SCAN gene structure prediction software, as shown on the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr6:26087509-26095469&knownGene=pack&nscanGene=pack).

The file is in GFF (General Feature Format) format:

http://www.ensembl.org/info/website/upload/gff.html

Attach the file to Ensembl and have a look at the result.

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org/).
🖱 Go to any human-specific gene or location page.
🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.
🖱 Type 'N-SCAN genes' in the 'Name' text box.
🖱 Select 'Data format: GFF'.
🖱 Click 'Attach via URL'
🖱 Enter the URL of the file in the 'File URL' text box.
🖱 Click [Attach]

🖱 Click on 'Go to first region with data: 6:26037670-26137670'.

A new track named 'N-SCAN genes' has now been added to the 'Region in detail' page.

You may want to turn off all tracks that you added to the display in the previous exercises.

🖱 Click [Configure this page] in the side menu.
🖱 Click [Reset configuration].

To display the names of the N-SCAN genes.

🖱 Hover over the 'N-SCAN genes' track name.
🖱 Hover over the 'Change track style' icon (the cogwheel).
🖱 Select 'Labels'.

Note that, at the moment, the CDS information in the GFF file is not taken into account in Ensembl and thus no distinction between the UTRs and CDS of the transcripts can be seen.

Exercise 2 – Attaching a BigWig file (human)

The *BCL11A* (B-cell CLL/lymphoma 11A (zinc finger protein)) gene functions as a myeloid and B-cell proto-oncogene.

The files

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/wgEncodeCaltechRnaSeqGm12878R2x75Th1014Il200SigRep1V4.bigWig

and

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/wgEncodeCaltechRnaSeqK562R2x75Th1014Il200SigRep1V4.bigWig

contain RNA-Seq data for the GM12878 and K562 cell lines, respectively.

The files are in BigWig format:

https://cgwb.nci.nih.gov/goldenPath/help/bigWig.html

Attach both files to Ensembl and have a look at the result. Is the *BCL11A* gene expressed in both cell lines?

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org).
🖱 Select 'Search: Human' and type 'bcl11a' in the 'for' text box.
🖱 Click [Go].
🖱 Click on 'Gene' on the page with search results.
🖱 Click on 'Human'.
🖱 Click on '2:60678302-60780702:-1'.

You may want to turn off all tracks that you added to the display in the previous exercises.

🖱 Click [Configure this page] in the side menu.
🖱 Click [Reset configuration].
🖱 Click (✓).

🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.
🖱 Type 'GM12878_RNAseq' in the 'Name' text box.
🖱 Select 'Data format: BigWig'.
🖱 Enter the URL of the first file in the 'Provide file URL' text box.
🖱 Click [Attach].
🖱 Repeat for the second file.
🖱 Click (✓).

The *BCL11A* gene is expressed in the GM12878 cell line, while there is virtually no expression in the K562 cell line. Note that the vertical scale differs between the two attached RNA-Seq tracks.

Exercise 3 – Attaching a BAM file (human)

The following file contains alignments to the GRCh37 assembly of low coverage Illumina sequencing reads of chromosome 20 of individual HG00096 from the 'British from England and Scotland, UK' cohort (http://ccr.coriell.org/Sections/Search/Sample_Detail.aspx?Ref=HG00096&PgId=166):

http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20100901.bam

The file is in BAM format. BAM is the compressed binary version of the SAM (Sequence Alignment/Map) format, a compact and indexable representation of nucleotide sequence alignments:

http://samtools.sourceforge.net/SAM1.pdf

To display these data in Ensembl also the .bam.bai index file is needed:

http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20100901.bam.bai

The .bam.bai file should be placed in the same directory as the .bam file.

Attach the file to Ensembl and have a look at the result. Can you find any individual reads containing a nucleotide that differs from the sequence of the reference genome? And a position where individual HG00096 differs from the reference genome or where individual HG00096 is heterozygous?

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org/).
🖱 Click on the picture of human (Donatello's St. George) or the word 'Human' next to it.
🖱 Click on 'View karyotype'.
🖱 Click on chromosome 20 in the karyotype.
🖱 Click on 'Jump to location View' in the pop-up menu.

You may want to turn off all tracks that you added to the display in the previous exercises.

🖱 Click [Configure this page] in the side menu.
🖱 Click [Reset configuration].
🖱 Click (✓).

🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.
🖱 Type 'HG00096' in the 'Name' text box.
🖱 Select 'Data format: BAM'.
🖱 Enter the URL of the file in the 'Provide file URL' text box.
🖱 Click [Attach].

A new track named 'HG00096' has now been added to the 'Region in detail' page.

🖱 Zoom in to see the actual reads.

Individual reads are shown in grey, with the consensus sequence shown above the reads in colour.

Nucleotides that differ from the sequence of the reference genome are shown in red:

http://www.ensembl.org/Homo_sapiens/Location/View?db=core&r=20:44861207-44861246

An example of a position where individual HG00096 is heterozygous:

http://www.ensembl.org/Homo_sapiens/Location/View?db=core&r=20:44854706-44854746

Exercise 3 – Attaching a BAM file (plants)

The following file contains alignments to the TAIR10 assembly of RNAseq reads of a wild type *Arabidopsis thaliana* seedling (http://www.ebi.ac.uk/ena/data/view/SRR070570&display=html):

http://www.ebi.ac.uk/~bert/SRR070570.bam

The file is in BAM format. BAM is the compressed binary version of the SAM (Sequence Alignment/Map) format, a compact and indexable representation of nucleotide sequence alignments:

http://samtools.sourceforge.net/SAM1.pdf

To display these data in Ensembl also the .bam.bai index file is needed:

http://www.ebi.ac.uk/~bert/SRR070570.bam.bai

The .bam.bai file should be placed in the same directory as the .bam file.

Attach the file to Ensembl and have a look at the result. Compare the expression of a gene that is expected to be constitutively highly expressed, e.g. *RBCS1A* (ribulose bisphosphate carboxylase small chain 1A), to one that is not expected to be constitutively expressed, e.g. *PR1* (pathogenesis-related protein 1).

Answer

🖱 Go to the Ensembl Plants homepage (http://plants.ensembl.org/).
🖱 Click on the picture of *Arabidopsis thaliana* or the word 'Arabidopsis thaliana' next to it.

🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.
🖱 Type 'SRR070570' in the 'Name' text box.
🖱 Select 'Data format: BAM'.
🖱 Enter the URL of the file in the 'Provide file URL' text box.
🖱 Click [Attach].

🖱 Go any region on the 'Region in detail' page under the 'Location' tab.
🖱 Click [Configure this page] in the side menu.
🖱 Click on 'Your data'.
🖱 Select 'SRR070570 - Unlimited'.
🖱 Click (✓).

A new track named 'SRR070570' has now been added to the display. The track shows the coverage as well as the actual reads.

🖱 Zoom in to see the actual reads.

Individual reads are shown in grey, with the consensus sequence shown above the reads in colour.

Nucleotides that differ from the sequence of the reference genome are shown in red. For example:

http://plants.ensembl.org/Arabidopsis_thaliana/Location/View?db=core&r=5%3A13957459-13957508

🖱 Type 'RBCS1A' in the 'Gene:' text box.
🖱 Click [Go].

The *RBCS1A* gene is highly expressed. Note that only a small part of the data is shown.

🖱 Type 'PR1' in the 'Gene:' text box.
🖱 Click [Go].

The *PR1* gene is not expressed at all.

Exercise 4 – Attaching a VCF file (mouse) – NOT WORKING AS SNPS ARE MAPPED TO NCBIM37 ASSEMBLY

The following file contains variants from the Sanger Mouse Genomes Project (http://www.sanger.ac.uk/resources/mouse/genomes/):

ftp://ftp-mouse.sanger.ac.uk/current_snps/20111102-snps-all.vcf.gz

The file is in VCF (Variant Call Format) format. VCF is a tab delimited format for storing variant calls and individual genotypes:

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40

To display these data in Ensembl also the .vcf.gz.tbi index file is needed:

ftp://ftp-mouse.sanger.ac.uk/current_snps/20111102-snps-all.vcf.gz.tbi

The .vcf.gz.tbi file should be placed in the same directory as the .vcf.gz file.

Attach the file to Ensembl and have a look at the result. Is the set of variants from the Sanger Mouse Genomes Project less or more comprehensive than the set of mouse variants present in Ensembl?

Answer

🖰 Go to the Ensembl homepage (http://www.ensembl.org/).
🖰 Click on the picture of the mouse or the word 'Mouse' next to it.
🖰 Click on 'View Karyotype'.
🖰 Click on any chromosome in the karyotype.
🖰 Click on 'Jump to location View' in the pop-up menu.

You may want to turn off all tracks that you added to the display in the previous exercises.

🖰 Click [Configure this page] in the side menu.
🖰 Click [Reset configuration].
🖰 Click (✓).

🖰 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.
🖰 Type 'Sanger variants" in the 'Name' text box.
🖰 Select 'Data format: VCF'.
🖰 Enter the URL of the file in the 'Provide file URL' text box.
🖰 Click [Attach].

🖰 Click [Configure this page] in the side menu.
🖰 Type 'variants' in the 'Find a track' text box.
🖰 Select 'Sequence variants (dbSNP and all other sources)'.
🖰 Click (✓).

A new track named 'Sanger variants' has now been added to the 'Region in detail' page. This track contains many more variants than the 'Sequence variants (dbSNP and all other sources)' track. Once the Sanger Mouse Genomes Project has submitted these variants to dbSNP they will subsequently be imported into Ensembl and shown in the 'Sequence variants (dbSNP and all other sources)' track.

🖰 Zoom in to see the individual variants.

Exercise 5 – Creating an annotated karyotype (human)

This is a list of all human caspase genes:

CASP1, *CASP2*, *CASP3*, *CASP4*, *CASP5*, *CASP6*, *CASP7*, *CASP8*, *CASP9*, *CASP10*, *CASP12*, *CASP14*

Create a figure of the human karyotype showing the genomic position of the caspase genes.

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org/).

🖱 Click on the human icon to go to the human homepage

🖱 Click on 'View Karyotype'.

🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.

🖱 Click on 'Features on Karyotype'.

🖱 Enter the list of caspase genes in the 'ID(s)' text box.

🖱 Click [Show features].

The positions of the caspase genes are now shown in the karyotype by red triangles. Note that some of the genes (on chromosome 2 and 11) are so close to each other that they cannot be shown by separate triangles.

Exercise 5 – Creating an annotated karyotype (mouse)

This is a list of all mouse caspase genes:

Casp1, *Casp2*, *Casp3*, *Casp4*, *Casp6, Casp7*, *Casp8*, *Casp9*, *Casp12*, *Casp14*

Create a figure of the mouse karyotype showing the genomic position of the caspase genes.

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org).

🖱 Click on the mouse icon to go to the mouse homepage

🖱 Click on 'View Karyotype'.

🖱 Click [Add your data] (or [Manage your data] if you already have custom data in Ensembl) in the side menu.

🖱 Click on 'Features on Karyotype'.

🖱 Enter the list of caspase genes in the 'ID(s)' text box.

🖱 Click [Show features].

The positions of the caspase genes are now shown in the karyotype by red triangles. Note that some of the genes (on chromosome 9) are so close to each other that they cannot be shown by separate triangles.

Exercise 6 – Creating and uploading a BED file

Create a small text file containing some annotation in BED format (http://www.ensembl.org/info/website/upload/bed.html) and upload it to Ensembl.

Note that BED offers the simplest format, with only three required fields, i.e. chromosome, start and end.

Answer

🖱 Create a text file with your annotation in for example Notepad or TextEdit and save it on your computer.
🖱 Go to the Ensembl homepage (http://www.ensembl.org/).
🖱 Go to any page for your favourite species.
🖱 Click [Manage your data] in the side menu.
🖱 Click on 'Upload Data'.
🖱 Type the name for your track in the 'Name for this upload (optional)' text box.
🖱 Select 'Data format: BED'.
🖱 Click [Choose File] behind 'Upload file:'.
🖱 Select the text file you just created.
🖱 Click [Upload].
🖱 Click 'Go to first region with data:'.

Your data are now shown as a new track on the 'Region in detail' page.

Exercise 7 – Removing custom annotation

Remove your attached and uploaded annotations.

Answer

🖱 Go to the Ensembl homepage (http://www.ensembl.org/).
🖱 Go to any page for your favourite species.
🖱 Click [Manage your data] in the side menu.
🖱 Click for each attached / uploaded data set on the trash can icon.

🖰 Click (✓).

Your annotations have now been removed.
