Data Upload Exercises

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

Note: the answers to these exercises were composed in release version 61:

http://Feb2011.archive.ensembl.org/index.html

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

CUSTOM ANNOTATION

Exercise 1 – Attaching a GFF file

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).
- Click on the Leonardo da Vinci picture or the word 'Human' next to it.
- Click [Manage your data] in the side menu.
- Click on 'Attach URL Data'.
- Enter the URL of the file in the 'File URL' text box.
- Enter 'N-SCAN genes' in the 'Name for this track' text box.
- Click [Next>].
- Click 'Go to first region with data: 6:26037670-26137670'.

A new track named ‘N-SCAN genes’ should now have been added to the ‘Region in detail’ page.

To display the names of the transcripts:

- Click [Configure this page] in the side menu.
- Select ‘N-SCAN genes – Labels’ under ‘User attached data’.
- Click (✓).

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 WIG file

Have a look at the following file:

http://www.ebi.ac.uk/~bert/medip-chip_cd4.wig

It contains methylation values in the genomic region of the human *ICAM3* gene (ENSG00000076662) in CD4 cells (Rakyan *et al.* An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res. 2008 Sep;18(9):1518-29).

The file is in WIG (Wiggle) fixedStep format:

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

Upload the data to Ensembl and have a look at the result.

Answer

- Go to the Ensembl homepage (http://www.ensembl.org).
- Click on the Leonardo da Vinci picture or the word ‘Human’ next to it.
- Click [Manage your data] in the side menu.
- Click on ‘Attach URL Data’.
- Enter the URL of the file in the ‘File URL’ text box.
- Enter ‘MeDIP-chip CD4’ in the ‘Name for this track’ text box.
- Click [Next>].
- Click ‘Go to first region with data: 6:10393775-10493775.

A new track named ‘MeDIP-chip CD4’ should now have been added to the ‘Region in detail’ page.

Exercise 3 – Attaching a BAM file

The following file contains low coverage 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 sequencing was done on an Illumina machine and the reads were aligned to the GRCh37 assembly.

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

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 Leonardo da Vinci picture or the word 'Human' next to it.
- Click [Manage your data] in the side menu.
- Click on 'Attach BAM File'.
- Enter the URL of the file in the 'BAM File URL' text box.
- Enter 'HG00096 in the 'Name for this track' text box.
- Click [Next>].
- Click (✓).
- Go to any 'Region in detail' page for chromosome 20.
- Click [Configure this page] in the side menu.
- Click on 'User attached data'.
- Select 'HG00096 – Unlimited'.
- Click (✓).

A new track named 'HG00096' should have 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.

If you want to compare the reads to the reference genome sequence:

- Click [Configure this page] in the side menu.
- Click on 'Sequence'.
- Select 'Sequence'.
- Click (✓).

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 4 – Creating an annotated karyotype

Have a look at the following file:

http://www.ebi.ac.uk/~bert/caspase_genes.txt

It contains the genomic positions of the human caspase genes.

The file is in BED format:

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

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 Leonardo da Vinci picture or the word 'Human' next to it.
- Click on 'Sample entry points – Karyotype' in the side menu.
- Click [Manage your data] in the side menu.
- Click on 'Upload Data'.
- Enter 'Caspase genes' in the 'Name for this upload (optional)' text box.
- Select 'Data format: BED'.

- Copy/paste the data from the caspase_genes.txt file in the 'Paste file' text box.
- Click [Upload].
- Click the 'Karyotype panel' tab.
- Click on 'User attached data'.
- Select 'Caspase genes – Arrow on righthand side'.
- Click (✓).

The positions of the caspase genes should now be 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 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.
- Click on the Leonardo da Vinci picture or the word 'Human' next to it.
- Click [Manage your data] in the side menu.
- Click on 'Upload Data'.
- Enter the name for your track in the 'Name for this upload (optional)' text box.
- Select 'Data format: BED'.
- Click [Browse…] behind 'Upload file:'.
- Select the text file you just created.
- Click [Upload].
- Click 'Go to first region with data:'.

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

Exercise 6 – Removing custom annotation

Remove your attached and uploaded annotations.

Answer

🖱 Click [Manage your data] in the side menu.
🖱 Click for each dataset on 'Delete'.
🖱 Click (✓).

Your annotations should be removed now.
