Using Ensembl tools for browsing ENCODE data

Aims

- Learn how to search and navigate the Ensembl website with a focus on exploring ENCODE/GENCODE data and data generated by the Ensembl Regulatory Build
- Learn how to add custom tracks in Ensembl
- Learn how to retrieve data from Ensembl using the BioMart data retrieval tool

Introduction

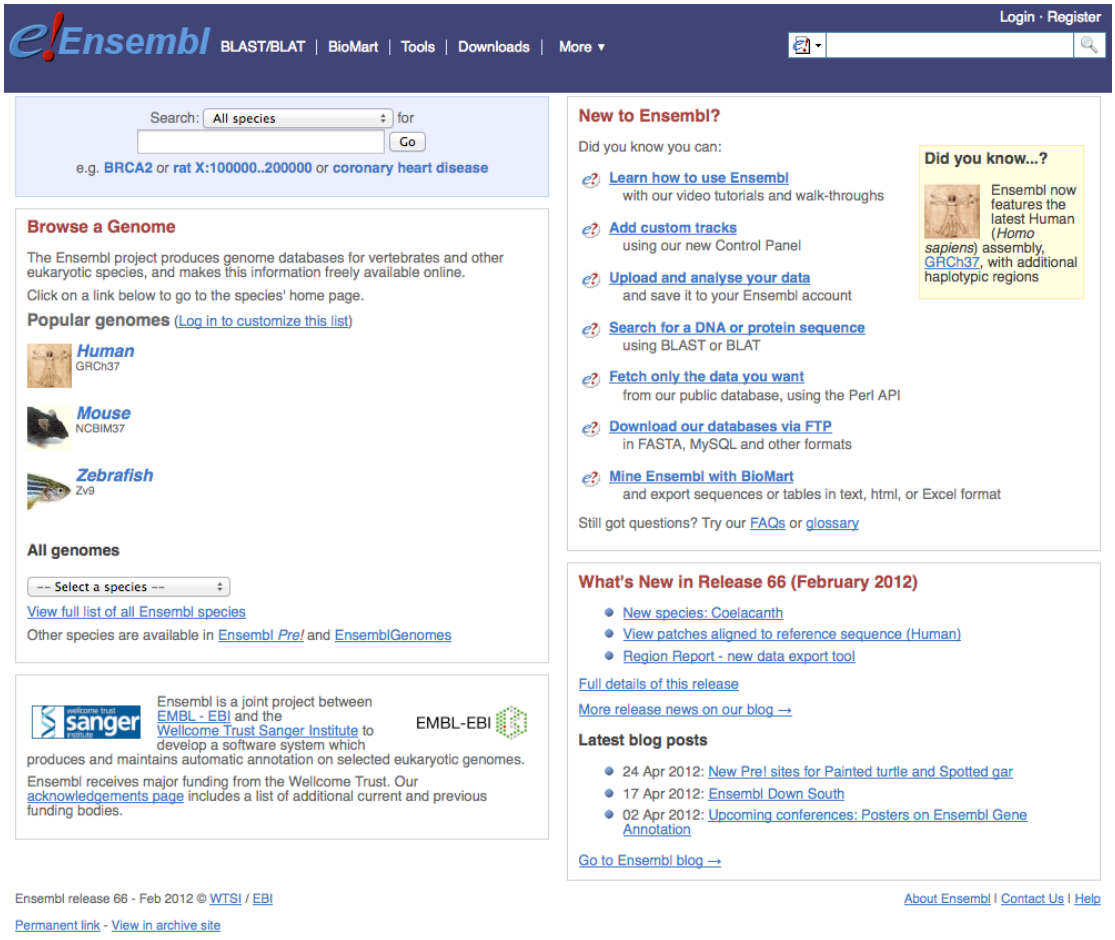
The Ensembl project (http://www.ensembl.org) provides genome resources for chordate genomes with a particular focus on human genome data as well as data for key model organisms such as mouse, rat and zebrafish. The total number of supported species is 68 as of Ensembl release 66 (February 2012). Of these, 57 species appear on the main Ensembl website and eleven species are provided on the Ensembl preview site (Pre! Ensembl; http://pre.ensembl.org) with preliminary support. For all species on the main site, we provide comprehensive, evidence-based gene annotations and comparative resources including alignments and homology, orthology and paralogy relationships based on Ensembl GeneTrees. We integrate these annotations with a large number of external data sources including InterPro, UniProt and Pfam. Eighteen of our most popular species also include dedicated variation resources derived from dbSNP, DGVa and other sources. The Ensembl Regulatory Build provides regulatory annotation on the human and mouse genomes and incorporates data from the ENCODE and Roadmap Epigenomics Program.

In addition to the data available through the Ensembl website, we provide open access to the Ensembl API and all supporting Ensembl databases to enable flexible, programmatic interaction with our data for use in genomic analysis. Data can also be accessed through Ensembl BioMart. We support those who use multiple web-based genome bioinformatics sites by providing links to the UCSC Genome Browser and NCBI's MapViewer on all our Location pages. We also support user data upload and visualization using BAM, BigWig, VCF and other common data formats.

Worked example 1 – Browser

In this worked example we will explore the human *BRCA2* (breast cancer 2, early onset) gene, with an emphasis on the Ensembl Regulatory Build and regulatory segmentation tracks.

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



First of all, we have to search for the human *BRCA2* gene:

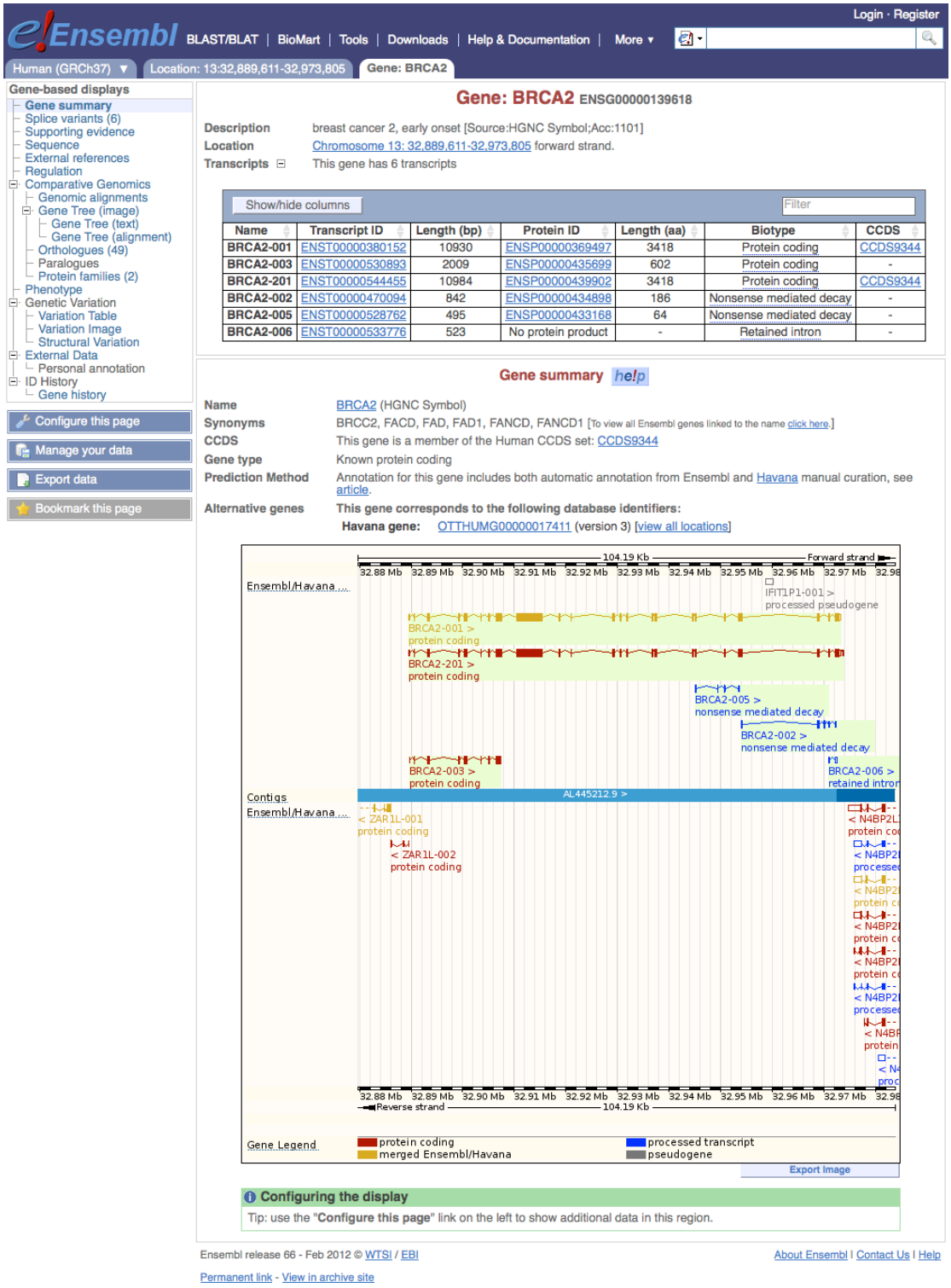
- Select 'Search: Human' and type 'brca2' in the 'for' text box.
- Click [Go].
- Click on 'Gene' on the page with search results.
- Click on 'Human'.

Note that, apart from the *BRCA2* gene, the search also returns genes that have the text 'BRCA2' as part of their description.

- Click on 'BRCA2 [Ensembl/Havana merge: ENSG00000139618].

This leads us to the 'Gene summary' page under the 'Gene' tab. This page shows general information about the *BRCA2* gene and all transcripts that

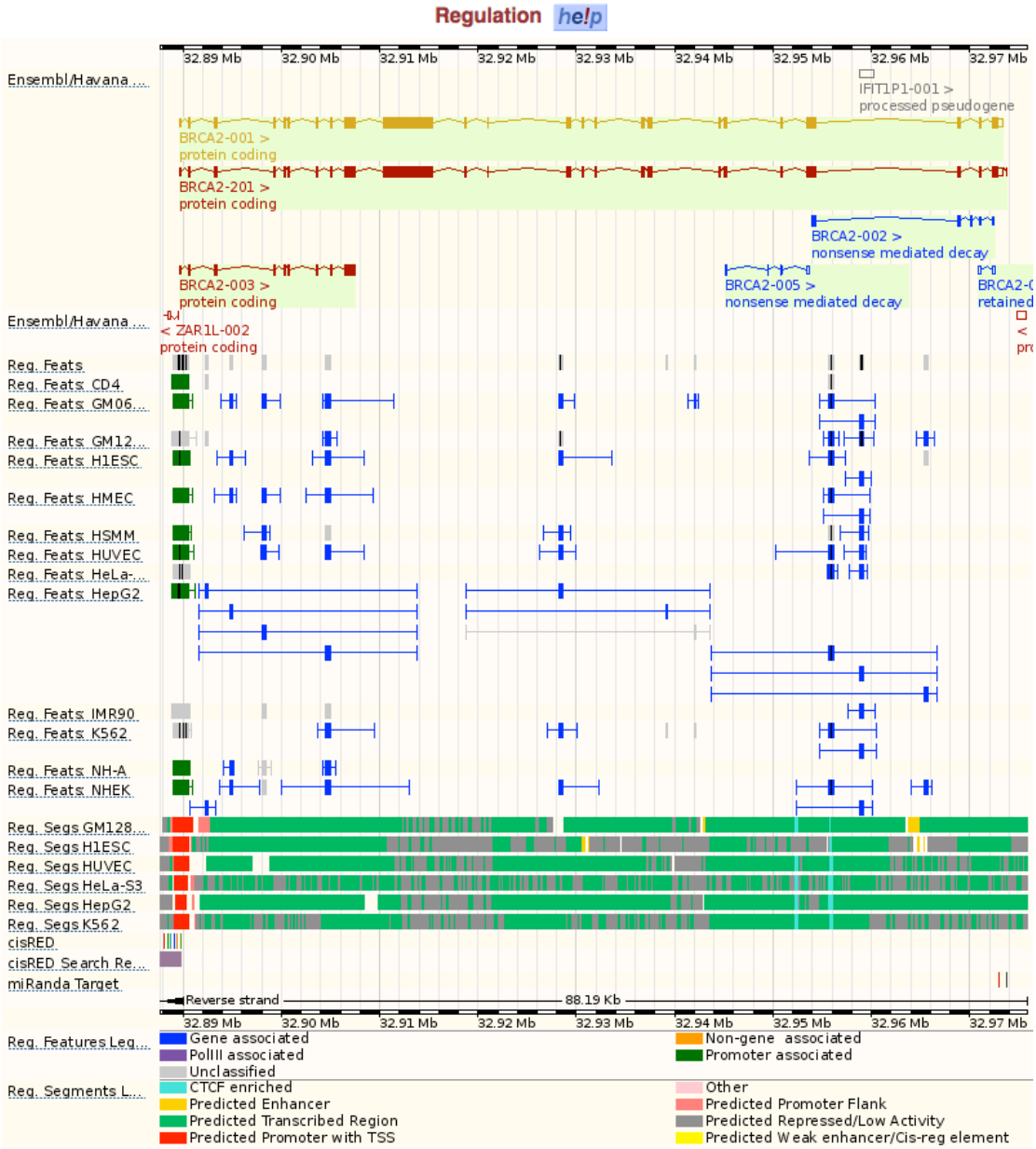
have been annotated for it as part of the GENCODE gene set. Ensembl/Havana merge transcripts are shown in golden color. Note the [*he!p*] button that opens up a help page as well as the legend at the bottom of the graphical display.



Pages (also called ‘views’) in Ensembl are organized under a number of tabs, i.e. ‘Species’, ‘Location’, ‘Gene’, ‘Transcript’, ‘Variation’ and ‘Regulation’. The various available pages under each tab are listed in the left-hand side menu.

Click on 'Regulation' in the side menu.

This leads us to the 'Regulation' page. This page shows all regulatory features for the *BRCA2* gene as predicted by the Ensembl Regulatory Build as well as the regulatory segmentation tracks.



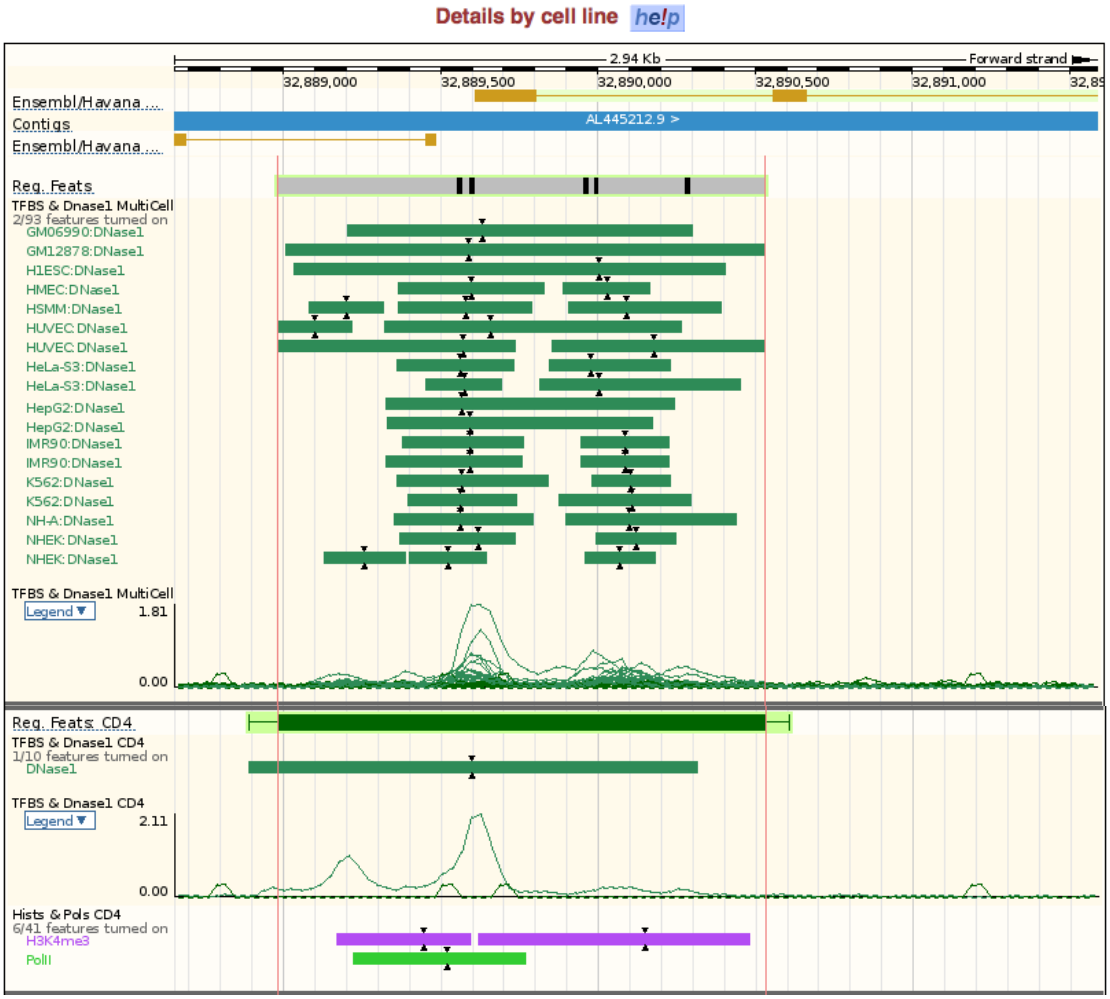
In general, clicking on any feature that is shown in an Ensembl graphical display should result in a pop-up menu with some basic information about our feature of interest and often also one or more hyperlinks to pages where more detailed information can be found.

Click in the 'Reg.Feats' track on the left most regulatory feature, that overlaps the 5' end of the *BRCA2* gene.

The resulting pop-up shows the core attributes underlying this MultiCell regulatory feature (DNase1 and transcription factors) as well as a list of the transcription factor binding site motives found in this regulatory feature with links to the JASPAR database (http://jaspar.cgb.ki.se).

Click on 'ENSR00000054736' in the pop-up menu.

This leads us to the 'Details by cell line' page under the 'Regulation' tab. This page shows the regulatory features plus some of the underlying attributes per cell line as well as the regulatory segmentation tracks.

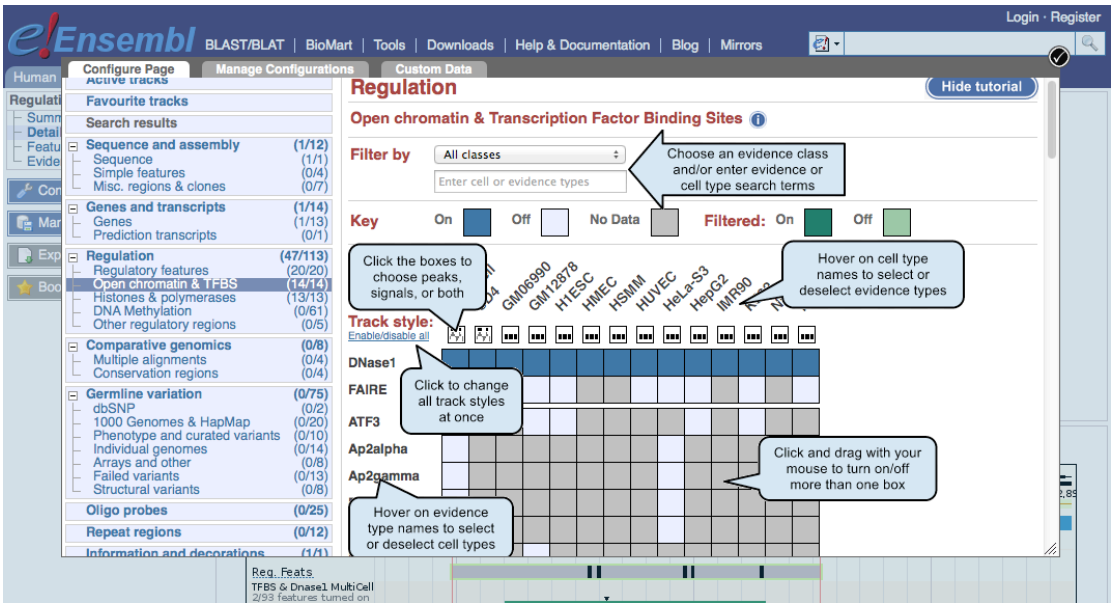


Only a sub set of the underlying attributes are shown by default. Additional attributes can be selected from the Regulation configuration matrices on the configuration page.

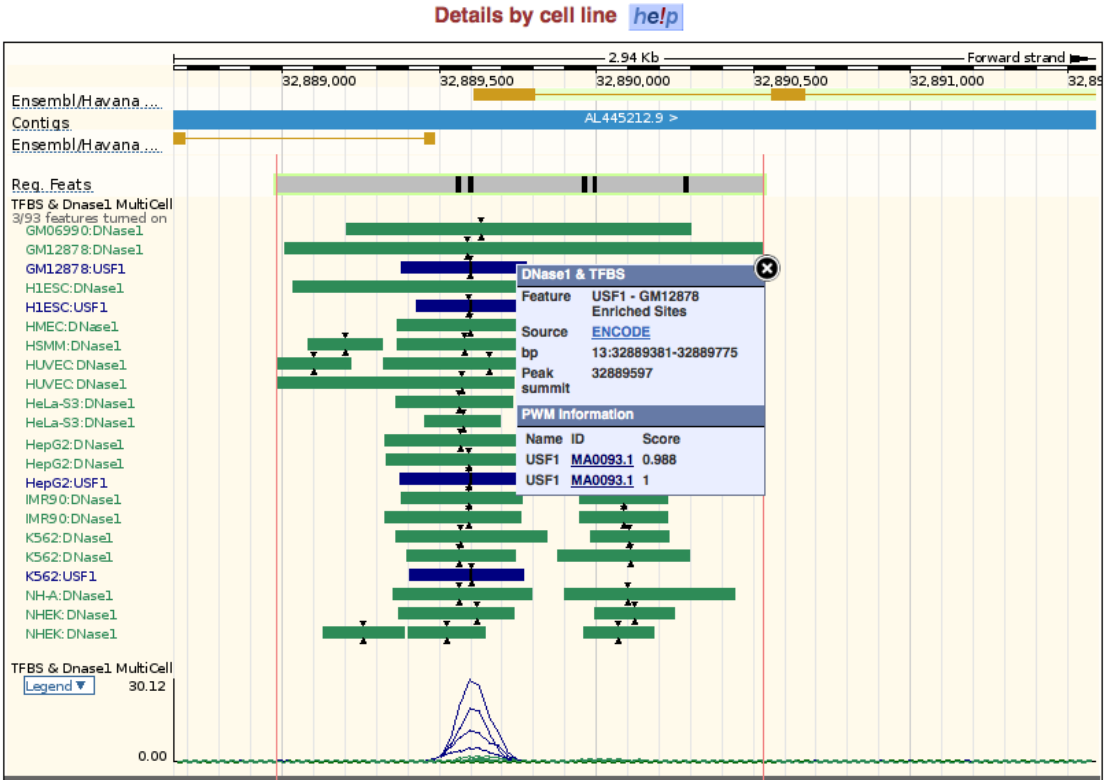
To add for example the information for the USF1 (Upstream stimulatory factor 1) transcription factor:

Click [Configure this page] in the side menu.
Click on 'Regulation – Open chromatin & TFBS'.
Click on [Hide tutorial].

- Type 'usf1' in the 'Enter cell or evidence types' text box in the 'Filter by' section.
- Hover over 'USF1' in the configuration matrix and check 'Select all USF1'.
- Click (✓) to close the configuration page.



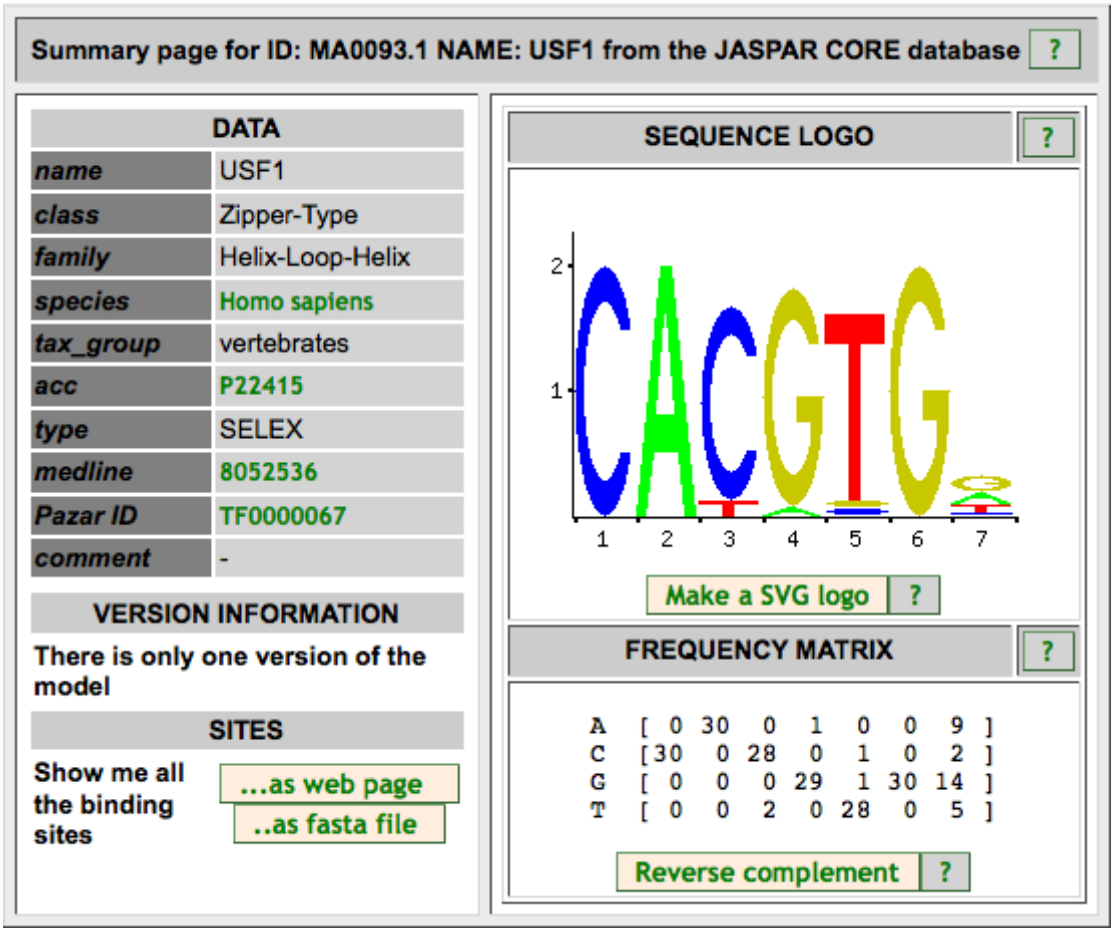
For several cell lines a block representing a region that binds the USF1 transcription factor should have been added to the display now.



Black triangles indicate the peak summit from the ChIP-Seq data. Vertical black lines indicate the position of the actual binding motif.

- Click on one of the USF1 blocks.
- Click on 'MA0093.1' in the pop-up menu.

This leads us to a page on the website of the JASPAR database that shows detailed information about the USF1 binding motif.



- Go back to the 'Details by cell line' page.

Apart from the 'Details by cell line' page there are three more pages under the 'Regulation' tab. The 'Summary' page shows our regulatory feature of interest plus all the underlying core evidence as well as the cell-specific regulatory features (without underlying evidence) and regulatory segmentation tracks. The 'Feature context' page shows our regulatory feature of interest along with neighboring regulatory features. The 'Evidence' page shows all information underlying our regulatory feature of interest in a tabular format.

- Click on 'Evidence' in the side menu.

Table columns can be hidden using the [Show/hide] columns button. Data can be ordered using the triangles next to the column header and filtered using the 'Filter' text box.

For example, to only show data with regard to USF1 binding for the MultiCell regulatory feature:

- Type 'multicell usf1' in the 'Filter' text box.

This should result in a table that only shows those rows that contain the terms 'MultiCell' and 'USF1'.

Evidence *help*

Show All entries | Show/hide columns | multicell usf1

Cell type	Evidence type	Feature name	Location
MultiCell	DNase1 & TFBS	USF1	13:32889373-32889806
MultiCell	DNase1 & TFBS	USF1	13:32889381-32889775
MultiCell	DNase1 & TFBS	USF1	13:32889406-32889769
MultiCell	DNase1 & TFBS	USF1	13:32889428-32889767
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889596-32889602
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889596-32889602
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889596-32889602
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889596-32889602
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889597-32889603
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889597-32889603
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889597-32889603
MultiCell	DNase1 & TFBS	USF1 (MA0093.1)	13:32889597-32889603

The table can be downloaded in comma-separated values (csv) format using the 'CSV' icon:

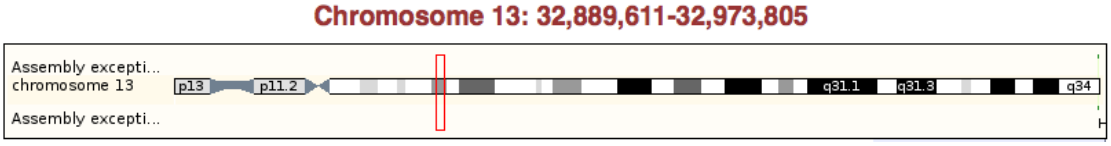
- Click on the 'CSV' icon.
- Click on 'Download what you see'.
- Open or save the csv file.

To view all the annotated genomic features (not only regulatory features) in and around the *BRCA2* gene, we have to go to the 'Location' tab.

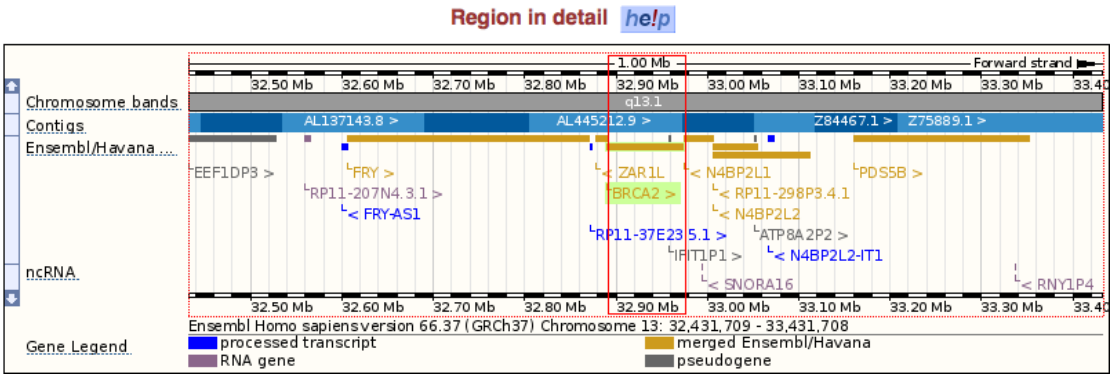
- Click on the 'Location' tab.

This leads us to the 'Region in details' page under the 'Location' tab. This page shows the genomic region of the *BRCA2* gene. It consists of three parts.

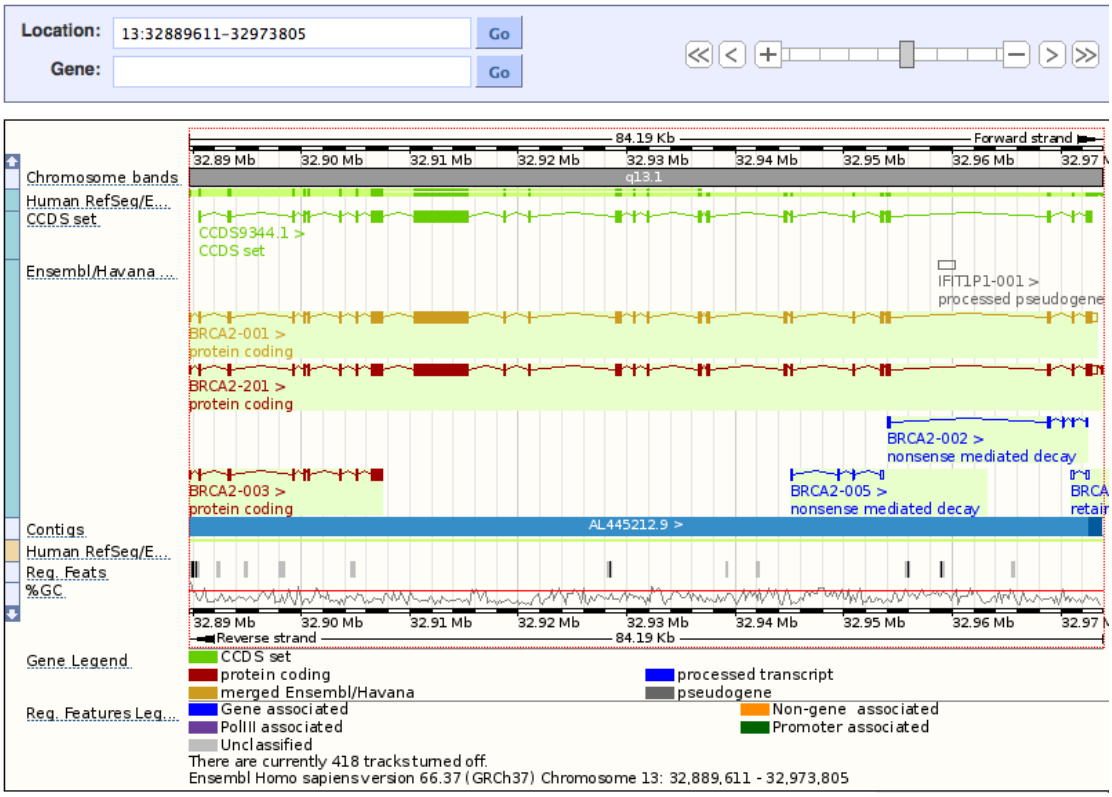
First, the complete chromosome.



Second, a 1 Mb region around our region of interest.



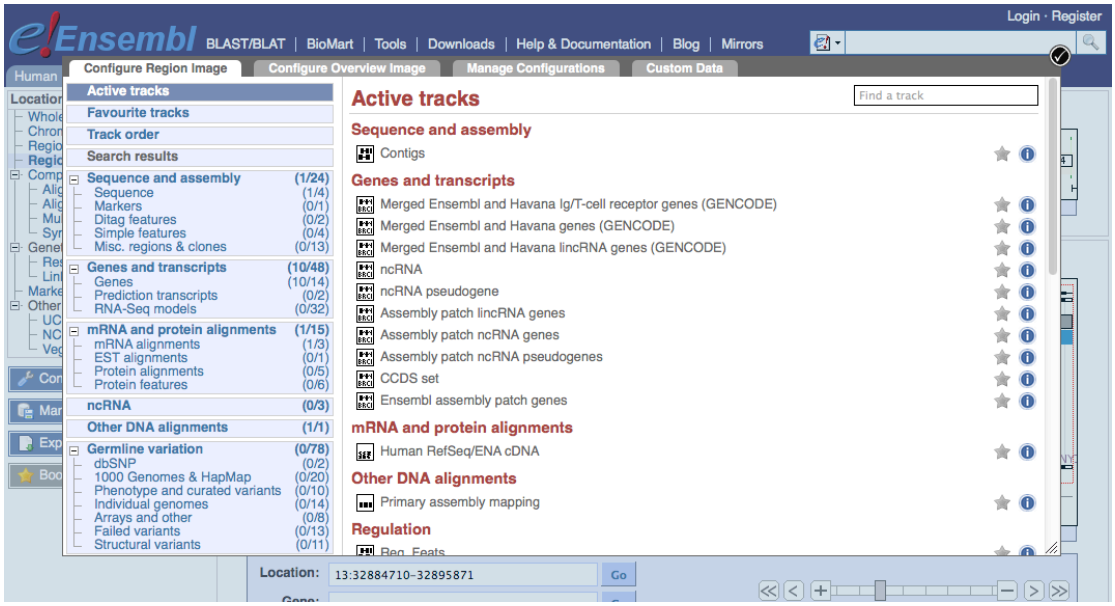
Third, our region of interest, which is in this case the *BRCA2* gene. Red boxes indicate where our region of interest is located on the 1 Mb region and where the 1 Mb region is located on the chromosome.



Zooming in and out is possible using the +/- slider at the top right of the display.

Zoom out one step (to 200000 bp) using the slider.

By default only a very limited number of tracks is shown (note that it says at the bottom the display that 'There are currently 418 tracks turned off'). Additional tracks can be added on the configuration page.



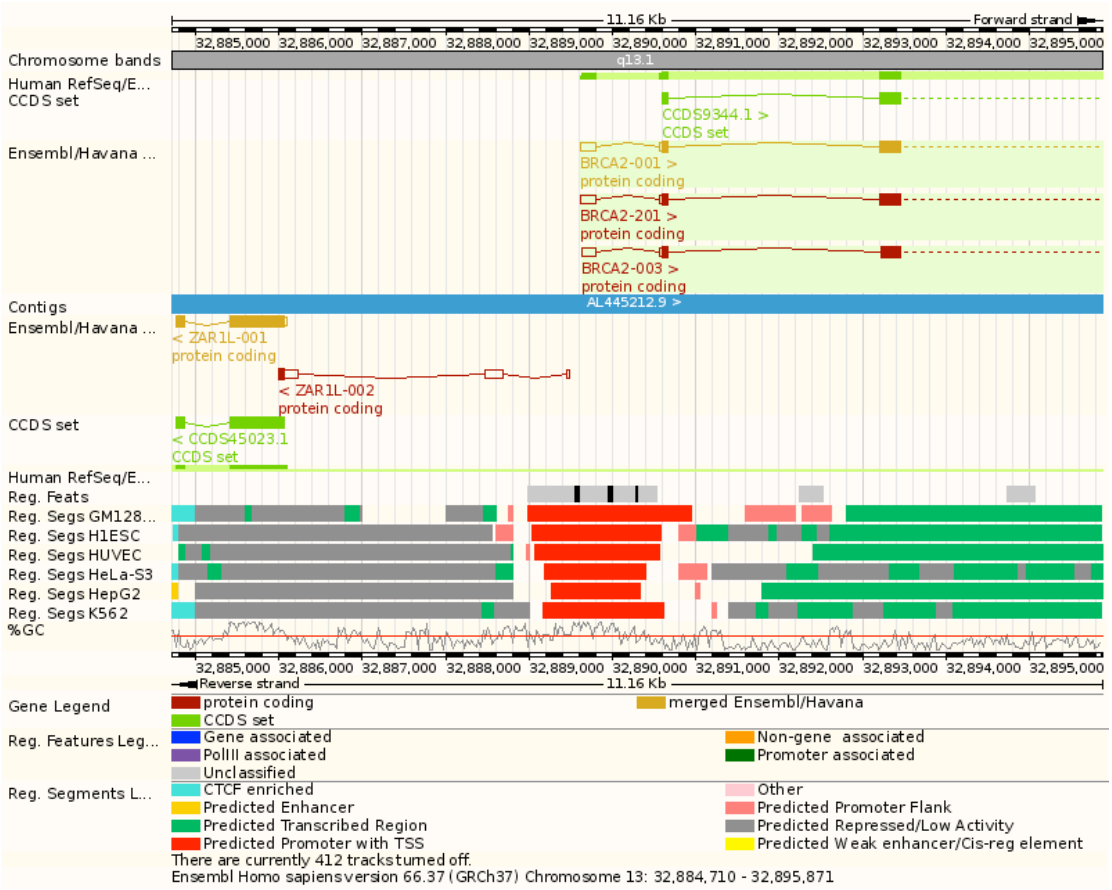
For example, to add the regulatory segmentation tracks:

- Click [Configure this page] in the side menu.
- Type 'segmentation' in the 'Find a track' text box.
- Select all six 'Reg. Segs' tracks.
- Click (✓).

The six regulatory segmentation tracks should now have been added to the display, as well as a color legend.

Zooming in on a particular region is possible by drawing a box around the desired region using your mouse or trackpad.

- Draw a box of about 10 kb around the region at the 5' end of the *BRCA2* gene that, according to the segmentation tracks, is a 'Predicted Promoter with TSS' (shown in red).
- Click on 'Jump to region' in the pop-up menu.



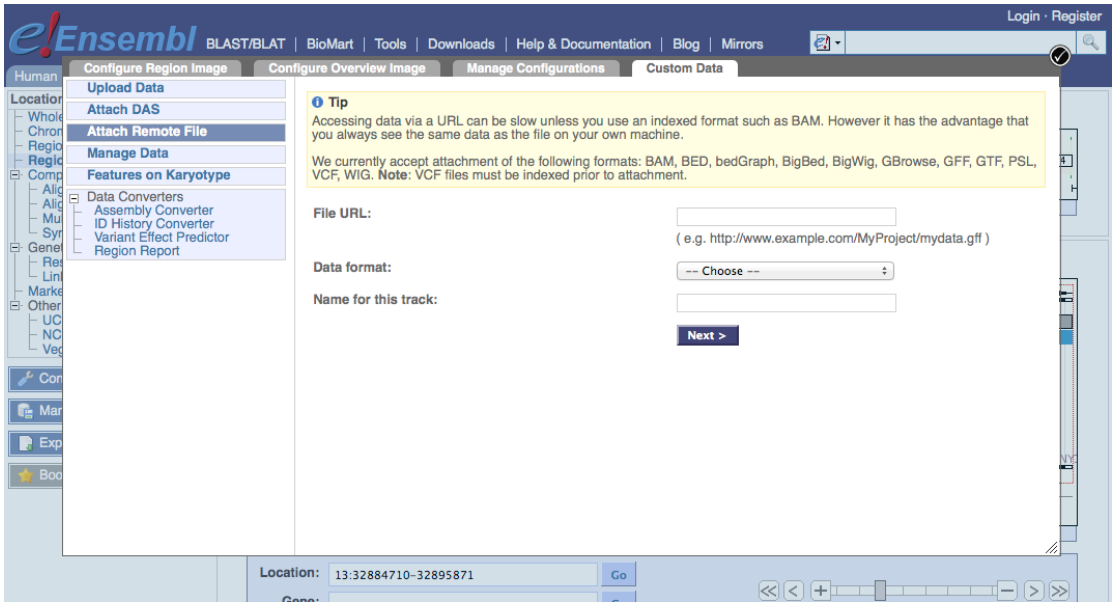
Tracks can be ordered by clicking on the bar in front of the track title and dragging the track to the desired location.

Individual tracks can be removed by hovering over the track title and clicking on the 'Turn track off' icon (i.e. the red circle with the white cross) in the pop-up menu.

Returning to the default settings is possible by clicking [Reset configuration] on the configuration page.

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

Finally, tracks with custom data in many different formats (e.g. BAM, BED, BigBed, BigWig, GFF, VCF) can be added to the display using the [Manage your data] button in the side menu.


Login · Register
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Configure Region Image
Configure Overview Image
Manage Configurations
Custom Data
Upload Data
Attach DAS
Attach Remote File
Manage Data
Features on Karyotype
Data Converters
Assembly Converter
ID History Converter
Variant Effect Predictor
Region Report
Tip
Accessing data via a URL can be slow unless you use an indexed format such as BAM. However it has the advantage that you always see the same data as the file on your own machine.
We currently accept attachment of the following formats: BAM, BED, bedGraph, BigBed, BigWig, GBrowse, GFF, GTF, PSL, VCF, WIG. Note: VCF files must be indexed prior to attachment.
File URL:
(e.g. http://www.example.com/MyProject/mydata.gff)
Data format:
-- Choose --
Name for this track:
Next >
Location:
13:32884710-32895871
Go

Worked example 2 – Biomart

In this worked example we will retrieve a list of all human genes in the GENCODE geneset that are located on the Y chromosome and that are protein-coding. Of these genes we will export the Ensembl Gene ID, Ensembl Transcript ID, gene biotype, transcript biotype, name and description and CCDS ID (http:// http://www.ncbi.nlm.nih.gov/CCDS/).

Step 1 – Dataset:

- Go to the Ensembl homepage (http://www.ensembl.org).
- Click on the 'BioMart' link on the toolbar.

Start with all human Ensembl genes.

- Choose the 'Ensembl Genes 66' database.
- Choose the 'Homo sapiens genes (GRCh37.p6)' dataset.

Step 2 – Filters:

Now filter for the genes on the Y chromosome.

- Click on 'Filters' in the left panel.
- Expand the 'REGION' section by clicking on the + box.
- Select 'Chromosome – Y'. Make sure the check box in front of the filter is ticked, otherwise the filter won't work.

Note that what you filtered for is confirmed in the side menu.

- Click the [Count] button on the toolbar.

This should give you 513 / 56478 Genes.

Now filter further for genes that are protein-coding.

- Expand the 'GENE' section by clicking on the + box.
- Select 'Gene type – protein_coding'.
- Click the [Count] button on the toolbar.

This should give you 53 / 56478 Genes.

Step 3 – Attributes:

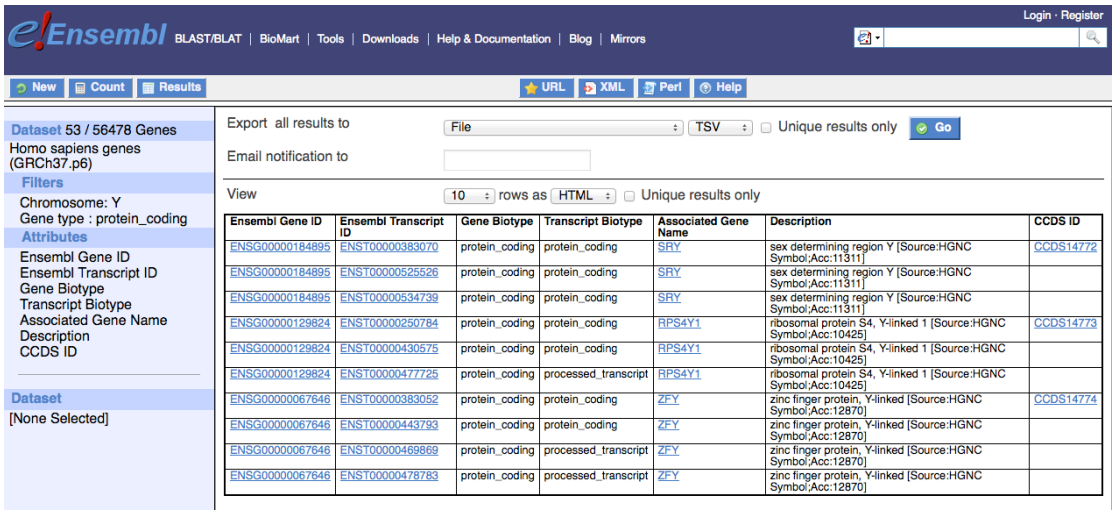
Specify the attributes to be included in the output (note that a number of attributes will already be selected by default).

- Click on 'Attributes' in the left panel.
- Expand the 'GENE' section by clicking on the + box.
- Select, in addition to the attributes 'Ensembl Gene ID' and 'Ensembl Transcript ID' that are already selected by default, 'Transcript Biotype', 'Gene Biotype', 'Associated Gene Name' and 'Description'.
- Expand the 'EXTERNAL' section by clicking on the + box.
- Select 'CCDS ID'.

Step 4 – Results:

Have a look at a preview of the results (only 10 rows of the results will be shown!).

- Click the [Results] button on the toolbar.



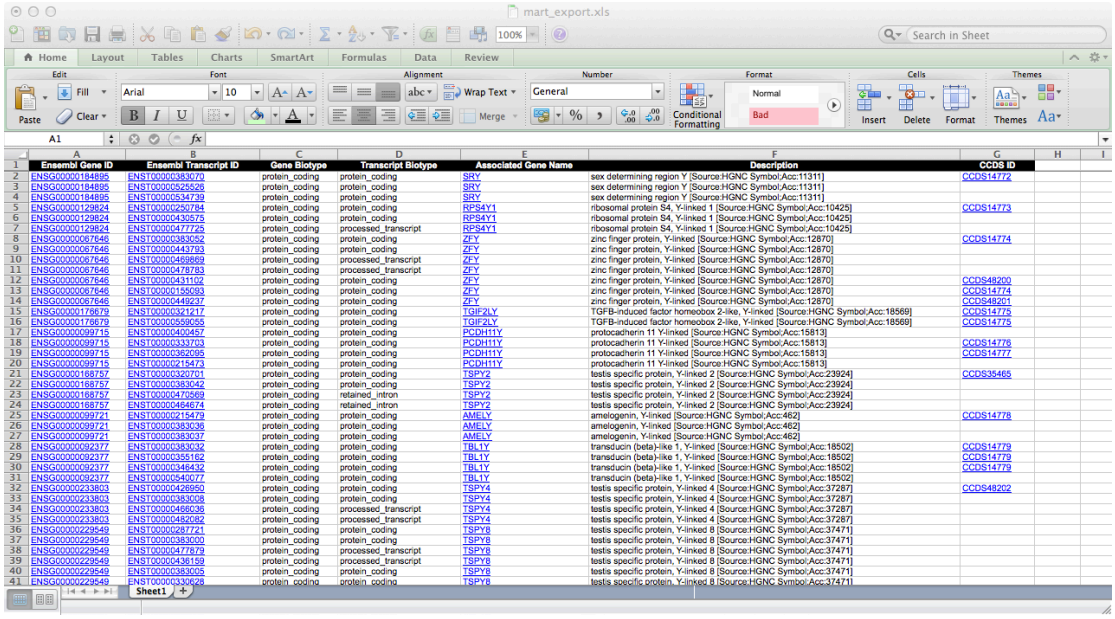
Ensembl Gene ID	Ensembl Transcript ID	Gene Biotype	Transcript Biotype	Associated Gene Name	Description	CCDS ID
ENSG00000184895	ENST00000383070	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	CCDS14772
ENSG00000184895	ENST00000525526	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	
ENSG00000184895	ENST00000534739	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	
ENSG00000129824	ENST00000250784	protein_coding	protein_coding	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	CCDS14773
ENSG00000129824	ENST00000430575	protein_coding	protein_coding	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	
ENSG00000129824	ENST00000477725	protein_coding	processed_transcript	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	
ENSG00000067646	ENST00000383052	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS14774
ENSG00000067646	ENST00000443793	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	
ENSG00000067646	ENST00000469869	protein_coding	processed_transcript	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	
ENSG00000067646	ENST00000478783	protein_coding	processed_transcript	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	

If you are happy with how the results look in the preview, output all the results.

- Select 'View All rows as HTML' or export all results to a file. To export the result to an Excel spreadsheet, select the 'XLS' format.

Note that when you select 'View All rows as HTML' your results will be shown under a new tab or in a new window, depending on your internet browser (and its settings).

Although you have filtered for only 53 genes, your results will contain more than 53 rows. This is because several of the genes have more than one transcript. Consequently the results contain a separate row for each of these transcripts. Also note that not all transcripts of a gene with biotype protein-coding necessarily have the biotype protein-coding.



Ensembl Gene ID	Ensembl Transcript ID	Gene Biotype	Transcript Biotype	Associated Gene Name	Description	CCDS ID
ENSG00000184895	ENST00000383070	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	CCDS14772
ENSG00000184895	ENST00000525526	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	
ENSG00000184895	ENST00000534739	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	
ENSG00000129824	ENST00000250784	protein_coding	protein_coding	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	CCDS14773
ENSG00000129824	ENST00000430575	protein_coding	protein_coding	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	
ENSG00000129824	ENST00000477725	protein_coding	processed_transcript	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	
ENSG00000067646	ENST00000383052	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS14774
ENSG00000067646	ENST00000443793	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	
ENSG00000067646	ENST00000469869	protein_coding	processed_transcript	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	
ENSG00000067646	ENST00000478783	protein_coding	processed_transcript	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	
ENSG00000067646	ENST00000431102	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS48200
ENSG00000067646	ENST00000155093	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS14774
ENSG00000067646	ENST00000449237	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS48201
ENSG00000178679	ENST00000321217	protein_coding	protein_coding	TGIF2LY	TGFB-induced factor homeobox 2-like, Y-linked [Source:HGNC Symbol;Acc:18569]	CCDS14775
ENSG00000178679	ENST00000559055	protein_coding	protein_coding	TGIF2LY	TGFB-induced factor homeobox 2-like, Y-linked [Source:HGNC Symbol;Acc:18569]	CCDS14775
ENSG00000099715	ENST00000400457	protein_coding	protein_coding	PCDH11Y	protocadherin 11 Y-linked [Source:HGNC Symbol;Acc:15813]	
ENSG00000099715	ENST00000333703	protein_coding	protein_coding	PCDH11Y	protocadherin 11 Y-linked [Source:HGNC Symbol;Acc:15813]	CCDS14776
ENSG00000099715	ENST00000362095	protein_coding	protein_coding	PCDH11Y	protocadherin 11 Y-linked [Source:HGNC Symbol;Acc:15813]	CCDS14777
ENSG00000099715	ENST00000215473	protein_coding	protein_coding	PCDH11Y	protocadherin 11 Y-linked [Source:HGNC Symbol;Acc:15813]	
ENSG00000168757	ENST00000320701	protein_coding	protein_coding	TSPY2	testis specific protein, Y-linked 2 [Source:HGNC Symbol;Acc:23924]	CCDS35465
ENSG00000168757	ENST00000383042	protein_coding	protein_coding	TSPY2	testis specific protein, Y-linked 2 [Source:HGNC Symbol;Acc:23924]	
ENSG00000168757	ENST00000470569	protein_coding	retained_intron	TSPY2	testis specific protein, Y-linked 2 [Source:HGNC Symbol;Acc:23924]	
ENSG00000168757	ENST00000464674	protein_coding	retained_intron	TSPY2	testis specific protein, Y-linked 2 [Source:HGNC Symbol;Acc:23924]	
ENSG00000099721	ENST00000215479	protein_coding	protein_coding	AMELY	amelogenin, Y-linked [Source:HGNC Symbol;Acc:462]	CCDS14778
ENSG00000099721	ENST00000383036	protein_coding	protein_coding	AMELY	amelogenin, Y-linked [Source:HGNC Symbol;Acc:462]	
ENSG00000099721	ENST00000383037	protein_coding	protein_coding	AMELY	amelogenin, Y-linked [Source:HGNC Symbol;Acc:462]	
ENSG00000092377	ENST00000383032	protein_coding	protein_coding	TBL1Y	transducin (beta)-like 1, Y-linked [Source:HGNC Symbol;Acc:18502]	CCDS14779
ENSG00000092377	ENST00000355162	protein_coding	protein_coding	TBL1Y	transducin (beta)-like 1, Y-linked [Source:HGNC Symbol;Acc:18502]	CCDS14779
ENSG00000092377	ENST00000346432	protein_coding	protein_coding	TBL1Y	transducin (beta)-like 1, Y-linked [Source:HGNC Symbol;Acc:18502]	CCDS14779
ENSG00000092377	ENST00000540077	protein_coding	protein_coding	TBL1Y	transducin (beta)-like 1, Y-linked [Source:HGNC Symbol;Acc:18502]	
ENSG00000233803	ENST00000426950	protein_coding	protein_coding	TSPY4	testis specific protein, Y-linked 4 [Source:HGNC Symbol;Acc:37287]	CCDS48202
ENSG00000233803	ENST00000383008	protein_coding	protein_coding	TSPY4	testis specific protein, Y-linked 4 [Source:HGNC Symbol;Acc:37287]	
ENSG00000233803	ENST00000466036	protein_coding	processed_transcript	TSPY4	testis specific protein, Y-linked 4 [Source:HGNC Symbol;Acc:37287]	
ENSG00000233803	ENST00000482082	protein_coding	processed_transcript	TSPY4	testis specific protein, Y-linked 4 [Source:HGNC Symbol;Acc:37287]	
ENSG00000229549	ENST00000287721	protein_coding	protein_coding	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	
ENSG00000229549	ENST00000383000	protein_coding	protein_coding	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	
ENSG00000229549	ENST00000477879	protein_coding	processed_transcript	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	
ENSG00000229549	ENST00000436159	protein_coding	processed_transcript	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	
ENSG00000229549	ENST00000383005	protein_coding	protein_coding	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	
ENSG00000229549	ENST00000330628	protein_coding	protein_coding	TSPY8	testis specific protein, Y-linked 8 [Source:HGNC Symbol;Acc:37471]	

Note: These exercises are based on Ensembl version 66 (February 2012). After in future a new version has gone live, version 66 will still be available for at least three years at http://e66.ensembl.org. If your answer doesn't correspond with the given answer, please consult the instructor.

Exercise 1 – Browser / Regulatory Build & segmentation

The *HLA-DRB1* and *HLA-DQA1* genes are part of the human major histocompatibility complex class II (MHC-II) region and are located about 44 kb from each other on chromosome 6. In the paper 'The human major histocompatibility complex class II HLA-DRB1 and HLA-DQA1 genes are separated by a CTCF-binding enhancer-blocking element' (Majumder *et al.* J Biol Chem. 2006 Jul 7;281(27):18435-43) a region of high acetylation located in the intergenic sequences between *HLA-DRB1* and *HLA-DQA1* is described. This region, termed XL9, coincided with sequences that bound the insulator protein CCCTC-binding factor (CTCF). Majumder *et al.* hypothesize that the XL9 region may have evolved to separate the transcriptional units of the *HLA-DR* and *HLA-DQ* genes.

(a) Go to the region from bp 32,540,000 to 32,620,000 on human chromosome 6

(b) Is there a regulatory feature annotated in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes that has CTCF binding data as (part of) its core evidence?

(c) Has CTCF binding been detected at this position in all cell types analyzed?

(d) Is the region that shows CTCF binding also a region of high acetylation, as found by Majumder *et al.*?

(e) Is the CTCF binding region reflected in the regulatory segmentation tracks?

Answer

(a)

- Go to the Ensembl homepage (http://www.ensembl.org).
- Select 'Search: Human' and type '6:32540000-32620000' in the 'for' text box.
- Click [Go].

If you didn't yet turn turn off all tracks that you added to the display in the worked example:

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

Click (✓).

(b)

Click on the regulatory features shown in the 'Reg. Feats' track that are located in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes. The resulting pop-ups show, amongst others, the core attributes underlying the regulatory features.

… or …

Click [Configure this page] in the side menu.
Click on 'Regulation – Open chromatin & TFBS'.
Click [Hide tutorial].
Click on the 'Track style' box in the 'MultiCell' column and select 'Both'.
Click (✓).

Optional: if you want to remove the DNase1 data to get a "cleaner" display:

Click [Configure this page] in the side menu.
Click on 'Regulation – Open chromatin & TFBS'.
Hover over 'DNase1' in the configuration matrix and uncheck 'Select all DNase1'.
Click on 'Track style: Enable/disable all' and select 'Off'.
Click on the 'Track style' box in the 'MultiCell' column and select 'Both'.
Click (✓).

Yes, there is one regulatory feature, i.e. ENSR00000488025, that has CTCF binding data as part of its core evidence.

(c)

If you haven't done this already in part (b):

Click [Configure this page] in the side menu.
Click on 'Regulation – Open chromatin & TFBS'.
Click on the 'Track style' box in the 'MultiCell' column and select 'Both'.
Click (✓).

CTCF binding has been detected at this position in all the cell types analyzed, with the exception of IMR90 and K562.

(d)

Click [Configure this page] in the side menu.

- Click on 'Regulation – Histones & polymerases'.
- Filter for all acetylation tracks by typing 'ac' in the 'Enter cell or evidence types' text box in the 'Filter by' section.
- Click and drag with your mouse to turn on all acetylation boxes in the configuration matrix.
- Click (✓).

Yes, the region that shows CTCF binding is also a region of high acetylation of histone 2, 3 and 4, at least in CD4 cells.

(e)

- Click [Configure this page] in the side menu.
- Click on 'Regulation – Regulatory features'.
- Click on 'Enable/disable all Segmentation features' and select 'On'.
- Click (✓).

Yes, the CTCF binding region is reflected in the segmentation tracks for five of the cell types studied, as shown by the light blue coloring, which indicates a 'CTCF enriched' segment.

Exercise 2 – Browser / Adding custom tracks

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 exercise as follows:

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

- Click [Manage your data] in the side menu.
- Click on 'Attach Remote File'.
- Enter the URL of the first file in the 'File URL' text box.
- Select 'Data format: BigWig'.
- Enter 'GM12878_RNAseq' in the 'Name for this track' text box.
- Click [Next>].
- Click [Save].
- Click (✓).
- Repeat for the second file.

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.

To remove the attached data sets:

- Click [Manage your data] in the side menu.
- Click for each data set on the trash can icon.
- Click (✓).

Exercise 3 – BioMart

A gene desert located on chromosome 8q24 is associated with multiple cancer types. One of the closest genes is the *MYC* proto-oncogene. Several studies suggest that the 8q24 region harbors regulatory elements that regulate the expression of *MYC* (Chromosome 8q24-Associated Cancers and MYC. Grisanzio C, Freedman ML. Genes Cancer. 2010 Jun;1(6):555-9.).

Generate for the above region (8:128573000-128745000) a list of all regulatory features predicted by the Ensembl Regulatory Build for the GM12878 cell line. Include the feature type, genomic coordinates and Regulatory Stable ID.

Answer

- Go to the Ensembl homepage (http://www.ensembl.org).
- Click on the 'BioMart' link on the toolbar.

- Choose the 'Ensembl Regulation 66' database.
- Choose the 'Homo sapiens genes (GRCh37.p6)' dataset.

- Click on 'Filters' in the left panel.
- Expand the 'REGULATORY FEATURES' section by clicking on the + box.
- Select 'Chromosome: 8'.
- Enter 'Base pair Start (bp): 128573000' and 'End (bp): 128745000'.
- Select 'Cell Type: GM12878 '.

- Click on 'Attributes' in the left panel.
- Deselect 'Feature Set' and 'Feature Type Description'.
- Select 'Regulatory Stable ID'.

- Click the [Results] button on the toolbar.
- Select 'View All rows as HTML' or export all results to a file.

There are 91 predicted regulatory features in the 8q24 gene desert, 89 of which are of the feature type 'unclassified' and one of the type 'gene associated' and 'promoter associated' each.
