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 (<u>http://www.ensembl.org</u>) 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 (<u>http://www.ensembl.org</u>).

	Login · Register
CENSEMBI BLAST/BLAT BioMart Tools Downloads	More 🔻 🙆 🔍
Search: All species for	New to Ensembl?
Go	Did you know you can: Did you know?
e.g. BRCA2 or rat X:100000200000 or coronary heart disease	e? Learn how to use Ensembl
	with our video tutorials and walk-throughs
Browse a Genome	Add custom tracks Iatest Human (Homo
The Ensembl project produces genome databases for vertebrates and other	using our new Control Panel sapiens) assembly, GBCb37 with additional
eukaryotic species, and makes this information freely available online.	Upload and analyse your data haplotypic regions
Click on a link below to go to the species' home page.	and save it to your Ensembl account
Popular genomes (Log in to customize this list)	Search for a DNA or protein sequence
Human GPChaz	USING BLAST OF BLAT
Ghuisr	Fetch only the data you want
Mouse	from our public database, using the Perl API
NCBIM37	Download our databases via FTP
	In FASTA, MySQL and other formats
Zebrafish Zv9	Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format
Procession of the second se	
	Still got questions? Try our <u>FAQS</u> or <u>glossary</u>
All genomes	
Select a species \$	What's New in Release 66 (February 2012)
View full list of all Ensembl species	<u>New species: Coelacanth</u>
Other species are available in Ensembl Pre! and EnsemblGenomes	View patches aligned to reference sequence (Human)
	Region Report - new data export tool
Economia e inicitat escient between	Full details of this release
Ensemblies a joint project between	More release news on our blog \rightarrow
Wellcome Trust Sanger Institute to develop a software system which	Latest blog posts
produces and maintains automatic annotation on selected eukaryotic genomes.	• 24 Apr 2012: New Pre! sites for Painted turtle and Spotted par
Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous	17 Apr 2012: Ensembl Down South
funding bodies.	 02 Apr 2012: <u>Upcoming conferences: Posters on Ensembl Gene</u> <u>Annotation</u>
	Go to Ensembl blog \rightarrow
Ensembl release 66 - Feb 2012 © WTSI / EBI	About Ensembl I Contact Us I Help
Permanent link - View in archive site	

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

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

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.

							Login · Register			
CEnsembl	LAST/BLAT Bid	Mart Tools Dow	vnloads Help &	Documentation	More 🔻 🛃		Q			
Human (GBCh37)	12.22 880 611-2	2 973 805 Gene: I	BRCA2							
Gene-based displays	1. 13.32,869,011-3	2,575,805								
- Gene summary			Gene	: BRCA2 ENSG	00000139618					
 Splice variants (6) Supporting evidence 	Description	breast cancer 2, e	arly onset [Source	HGNC Symbol:Acc:	11011					
- Sequence	Location	Chromosome 13:	32,889,611-32,97	3,805 forward strand.						
 External references Begulation 	Transcripts 🖃	This gene has 6 tr	anscripts							
Comparative Genomics										
Genomic alignments Gene Tree (image)	Show/hid	de columns				Filter				
Gene Tree (text)	Name	Transcript ID	Length (bp) 💧	Protein ID	Length (aa) 💧	Biotype	CCDS ≜			
 Gene Tree (alignment) Orthologues (49) 	BRCA2-001	ENST0000380152	10930	ENSP00000369497	3418	Protein coding	CCDS9344			
- Paralogues	BRCA2-003	ENST0000530893	2009	ENSP00000435699	602	Protein coding	-			
 Protein families (2) Phenotype 	BRCA2-201	ENST0000544455	10984	ENSP00000439902	3418	Protein coding	CCDS9344			
Genetic Variation	BRCA2-002	ENST00000470094	842	ENSP00000434898	186	Nonsense mediated decay	-			
- Variation Image	BRCA2-005	ENST00000533776	523	No protein product	-	Retained intron				
Structural Variation	2.110712 000		020	no proton product						
Personal annotation				Gono summary	holp					
□ ID History				dene summary	nep					
	Name	BRCA2 (HGN	IC Symbol)							
Configure this page	Synonyms	BRCC2, FAC	D, FAD, FAD1, FA	NCD, FANCD1 [To vie	w all Ensembl genes	s linked to the name <u>click here</u> .]				
🕞 Manage your data	CCDS	This gene is a	a member of the F	luman CCDS set: <u>CC</u>	<u>DS9344</u>					
Expect data	Gene type Prediction Meth	Appotation fo	n coaing r this appe include	es both automatic ann	otation from Eng	sombl and Havana manual cu	uration see			
	Frediction Meth	article.	r this gene include	s both automatic ann		sembi anu <u>mavana</u> manuai cu	Jiau01, 566			
🚖 Bookmark this page	Alternative gen	es This gene co	rresponds to the	e following database	identifiers:					
		Havana gen	e: OTTHUMG	00000017411 (version	3) [view all loca	itions]				
104,19 Kb Farward stranc 32,88 Mb 32,89 Mb 32,90 Mb 32,91 Mb 32,92 Mb 32,93 Mb 32,94 Mb 32,95 Mb 32,96 Mb 32,97 Mb										
	Ensembl	/Havana				IFIT1P1-001 >				
			n		······	processed piseudo	gene			
			BRCA2-001 >							
					***	~+/ # #/10				
			protein coding							
					H B					
					n	onsense mediated decay				
						BRCA2-002 >				
			r ra			nonsense mediated de	cay			
			BRCA2-003 > protein coding			BRCA	<2-006 > ned intror			
	Contigs			AL445212.9) >					
	Ensembl	/Havana < ZAR1L-0	рі			<	N4BP2L			
		protein co	ding			P	rotein col			
		< 2	AR1L-002				< N4BP2I			
		pro	terricounig							
							< N4BP21 protein co			
							protein co			
							< N4BP2I			
							protein co			
							< N4BP21			
							W-1			
							< N4BF protein			
							proc			
		32.88 Mb ————————————————————————————————————	32.89 Mb 32.90 Ml e strand	32.91 Mb 32.92 Mb 10	32.93 Mb 32.94 N 4.19 Kb	4b 32.95 Mb 32.96 Mb 32.97 I	Mb 32.98			
	Gene Le	gend protei	in coding		processed	transcript				
		merg	eu chsembi/Havai	10	pseudoge	Export Image				
	0.0	iourino the disular								
	Cont	iguring the display	enell link on the 1	A to obout - dalla	data in this as it.	-				
	Tip: use	the "Configure this p	age" link on the li	en to snow additional	uata in this regio	n.				
	Ensembl release 66	6 - Feb 2012 © WTSI / F	BI			About Ensembl I	Contact Us I Help			
	Bormonont link Vi	aw in archive site				stood chadhor	<u></u> i <u></u>			

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.

A 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 (<u>http://jaspar.cgb.ki.se</u>).

Click on 'ENSR0000054736' 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.



Details by cell line help

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:

- The side menu.
- A Click on 'Regulation Open chromatin & TFBS'.
- Click on [Hide tutorial].

 $^{\circ}$ Type 'usf1' in the 'Enter cell or evidence types' text box in the 'Filter by' section.

 ${}^{\prime \ominus}$ Hover over 'USF1' in the configuration matrix and check 'Select all USF1'.

 \mathcal{A} Click (\checkmark) to close the configuration page.

		Login · Register
$e_{\mathbf{I}}$		art Tools Downloads Help & Documentation Blog Mirrors 🛛 🔯 - 💦 🔍
Human	Configure Page Manage Configuratio	ns Custom Data
Regulati	Favourite tracks	Regulation
- Sumn	Search results	Open chromatin & Transcription Factor Binding Sites 1
Featu Evide	Sequence and assembly (1/12) Sequence (1/1) Simple features (0/4) Misc. regions & clones (0/7)	Filter by All classes Enter cell or evidence types Choose an evidence or cell type search terms
🕞 Mar	□ Genes and transcripts (1/14) □ Genes (1/13) □ Prediction transcripts (0/1)	Key On Off No Data Filtered: On Off
📑 Exp	 Regulation (47/113) Regulatory features (20/20) Open chromatin & TFBS (14/14) Histones & polymerases (13/13) DNA Methylation (0/61) Other regulatory regions (0/5) 	Click the boxes to choose peaks, signals, or both Track style:
	Comparative genomics (0/8) Multiple alignments (0/4) Conservation regions (0/4)	
	Bermline variation (0/75) dbSNP (0/2) 1000 Genomes & HapMap (0/2) Phenotype and curated variants (0/10) Individual genomes (0/14) A rrays and other (0/8) Failed variants (0/13)	ATF3 all track styles at once Ap2alpha Click and drag with your mouse to turn on horr mouse to turn on horr
	Structural variants (0/8) Oligo probes (0/25)	Hover on evidence
	Repeat regions (0/12)	or deselect cell types
l	Information and decorations (1/1) Reg. Feats TFBS & Dnasel M 2/33 features tur	ubicell

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



Details by cell line help

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

- \mathcal{A} 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.

A 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 rentries multicell usf1 Show/hide columns CSV Cell type Evidence type Feature name Location MultiCell DNase1 & TEBS USF1 13:32889373-32889806 MultiCell **DNase1 & TFBS** USF1 13:32889381-32889775 MultiCell DNase1 & TFBS USF1 13:32889406-32889769 MultiCell USF1 13:32889428-32889767 **DNase1 & TFBS** MultiCell **DNase1 & TFBS** USF1 (MA0093.1) 13:32889596-32889602 13:32889596-32889602 DNase1 & TEBS USF1 (MA0093.1) MultiCell 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 DNase1 & TFBS 13:32889597-32889603 MultiCell USF1 (MA0093.1) 13:32889597-32889603 MultiCell **DNase1 & TFBS** USF1 (MA0093.1)

The table can be downloaded in comma-separated values (csv) format using 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.

Chromosome 13: 32,889,611-32,973,805

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

					Regio	n in detai	he!p							
							- 1.00 Mb -				– Forward stra	nd 💻		
1	Characteristic hands		32.50 Mb	32.60 Mb	32.70 Mb	32.80 Mb	32.90 Mb	33.00 Mb	33.10 Mb	33.20 Mb	33.30 Mb	33,40		
	Chromosome bands		A1	1071400	-	AL 4452	q13.1		794467	1 > 775000	1 >			
	Contigs			13/143.0 >		ALTH32			204407	275665.	1 -			
	Ensembl/Havana			•					_					
		EEF1DP3	3 >	^L FRY >		۴.	ZAR1L	< N4BP2L1	- PC	DS5B >				
			^L RI	P11-207N4.3	1>		BRCA2 >	<pre>< RP11-2</pre>	298P3.4.1					
				< FRY-AS	L			^L < N4BP2	L2					
			R#11-37E23 5.1 > LATP8A2P2 >											
							LI F	TIP1 >	: N4BP2L2-IT	1				
	ncRNA.							- SNORAL	.6		L< RN	IY1P4		
•			32.50 Mb	32.60 Mb	32.70 Mb	32.80 Mb	32.90 Mb	33.00 Mb	33.10 Mb	33.20 Mb	33.30 Mb	33.40		
		Ensembl	Homo sap	oiensversion	66.37 (GRCh	137) Chromos	ome 13: 32	2,431,709 - 33	431,708					
	Gene Legend	proc	essed tra	anscript			r	merged Enser	mbl/Havana					
		RNA	gene				p	pseudogene						

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.

∠
 ^A 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:

- A Click [Configure this page] in the side menu.
- Type 'segmentation' in the 'Find a track' text box.

^[⊕] 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.

 $^{\circ}$ 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).

The 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 popup 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].

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.

Working with ENCODE data

		Login · F	Register
e_{1}	ENSEMBI BLAST/BL/	BioMart Tools Downloads Help & Documentation Blog Mirrors 🛛 😰 🗸	
Human	Configure Region Image	nfigure Overview Image Manage Configurations Custom Data	
Leastier	Upload Data		
	Attach DAS	ОТр	
- Chron	Attach Remote File	Accessing data via a UHL 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.	
- Regio	Manage Data	We automatic execute the sharest of the following formatic PAM, DED, he (Crant, BisBed, BisWig, OBseuro, OEE, OTE, D2)	_
E Comp	Features on Karyotype	VCF, WIG. Note: VCF files must be indexed prior to attachment.	-
	Data Converters		н
– Mu	 Assembly Converter ID History Converter 	File URL:	
⊢ Syr ⊟ Genel	 Variant Effect Predictor Region Report 	(e.g. http://www.example.com/MyProject/mydata.gff)	
- Res	- Region Report	Data format:	
- Marke			
⊡ Other		Name for this track:	
		Next >	
- Veg			
Je Con			
📑 Mar			
Exp			
			NY
1 Boo			
			\square
	Lo	ation: 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://www.ncbi.nlm.nih.gov/CCDS/).

Step 1 – Dataset:

Go to the Ensembl homepage (<u>http://www.ensembl.org</u>).

A Click on the 'BioMart' link on the toolbar.

Start with all human Ensembl genes.

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

Step 2 – Filters:

Now filter for the genes on the Y chromosome.

A Click on 'Filters' in the left panel.

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.

- The 'GENE' section by clicking on the + box.
- → Click the [Count] button on the toolbar.

This should give you 53 / 56478 Genes.

<u>Step 3 – Attributes:</u>

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

Attributes' in the left panel.

The 'GENE' section by clicking on the + box.

A 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'.

TEXPANDE 'EXTERNAL' section by clicking on the + box.

<u>Step 4 – Results:</u>

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.

CENSEMDI BLAST	BLAT BioMart Too	ls Downloads He	elp & Documentati	ion Blog Mirrors		🖓 •	Q,					
New Count Results			*	URL 👂 XML 🔮	Perl 💿 Help							
Dataset 53 / 56478 Genes	Export all results	to	File		\$ TSV \$	Unique results only of Go						
Homo sapiens genes (GRCh37.p6)	Email notification	to										
Filters	10.000											
Chromosome: Y	View		10 ÷ rows a	as HTML ÷ 0	nique results only							
Gene type : protein_coding	Ensembl Gene ID	Ensembl Transcript	Gene Biotype	Transcript Biotype	Associated Gene	Description	CCDS ID					
Attributes Ensembl Gene ID	ENSG00000184895	ENST00000383070	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]	CCDS14772					
Ensembl Transcript ID Gene Biotype	ENSG00000184895	ENST00000525526	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]						
Transcript Biotype	ENSG00000184895	ENST00000534739	protein_coding	protein_coding	SRY	sex determining region Y [Source:HGNC Symbol;Acc:11311]						
Associated Gene Name Description	ENSG00000129824	ENST00000250784	protein_coding	protein_coding	RPS4Y1	ribosomal protein S4, Y-linked 1 [Source:HGNC Symbol;Acc:10425]	CCDS14773					
CCDS ID	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]						
Dataset	ENSG0000067646	ENST00000383052	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]	CCDS14774					
[None Selected]	ENSG0000067646	ENST00000443793	protein_coding	protein_coding	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]						
	ENSG0000067646	ENST00000469869	protein_coding	processed_transcript	ZFY	zinc finger protein, Y-linked [Source:HGNC Symbol;Acc:12870]						
	ENSG0000067646	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 proteincoding necessarily have the biotype protein-coding.

0) mart_export	xls									
2	111		2 🖌 🖻 🗳	🔮 🕼 • 🕅 • 🔰	• 🌆 • 🍸	• 🕼 🛅 📑 100	* •					C	λ ≖ (Sear	ch in She	et .		
1	Home	Layou	it Tables	Charts SmartArt	Formulas	Data Review										1	∧ ÷-
	Edit			Font		Alignment	N	umber		Format			Cells		Then	nes	
r 🗎		Fill 🔻	Arial	• 10 • A• A•		abc 🔻 🚍 Wrap Text 🔻	General	*		Normal			-		And		
] • •								11 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			i i i	THE T		80000 Y		
Pa	ste 🥥	Clear *	BIU	• 🌭 • <u>A</u> •		🗧 🎾 📄 Merge 🕚	- 100	> \$\$.0 \$.0	Conditional	Bad	Ŭ	Insert	Delete	Format	Themes	Aa-	
	41		0.0.6.6.						Formatting								
	AI	•	5 (- Jx			1	-			-							
1	Engom	A bl Cone ID	Encombl Trans	crint ID Gene Biotype	Transcript	Biotune Associate	L Gone Name			Percription					C ID	н	
2	ENSGOOD	00184895	ENST00000383070	protein coding	protein coding	SRY	d Gene Rune	sex determining re	aion Y (Source:H	GNC Symbol:Acc:11	3111			CCD	\$14772		<u> </u>
3	ENSG000	00184895	ENST00000525526	protein coding	protein coding	SRY		sex determining re	gion Y [Source:H	GNC Symbol:Acc:11	3111						
4	ENSG000	00184895	ENST00000534739	protein coding	protein coding	SRY		sex determining re	aion Y (Source:H	GNC Symbol:Acc:11	3111						
5	ENSG000	00129824	ENST00000250784	protein_coding	protein_coding	RPS4Y1		ribosomal protein \$	54, Y-linked 1 [So	urce:HGNC Symbol;	Acc:1042	5]		CCD	514773		
6	ENSG000	00129824	ENST00000430575	protein_coding	protein_coding	RPS4Y1		ribosomal protein \$	34, Y-linked 1 [So	urce:HGNC Symbol;	Acc:1042	5]					
7	ENSG000	00129824	ENST00000477725	protein_coding	processed_trans	cript RPS4Y1		ribosomal protein \$	34, Y-linked 1 [So	urce:HGNC Symbol;	Acc:1042	5]					
8	ENSG000	00067646	ENST00000383052	protein_coding	protein_coding	ZFY		zinc finger protein,	Y-linked [Source:	HGNC Symbol;Acc:	12870]			CCD	514774		
9	ENSG000	00067646	ENST00000443793	protein_coding	protein_coding	ZFY		zinc finger protein,	Y-linked [Source:	HGNC Symbol;Acc:	12870]						
10	ENSG000	00067646	ENST00000469869	protein_coding	processed_trans	cript ZFY		zinc finger protein,	Y-linked [Source:	HGNC Symbol;Acc:	12870]						
11	ENSGOOD	00067646	ENST00000478783	protein_coding	processed_trans	cript ZFY		zinc finger protein,	Y-linked [Source:	HGNC Symbol;Acc:	12870]			0000040000			
12	ENSCOOL	00087646	ENS10000431102	protein_coding	protein_cooing	ZET		zinc ringer protein,	T-linked [Source:	HGNC Symbol;Acc:	12870]			CCDS48200			
14	ENRCOOL	00067646	ENET00000440227	protein_coding	protein_cooling	2F1 7EV		zinc ringer protein,	Y linked (Source:	HONC Symbol Acc.	128701			000040004			
15	ENSCOOL	00176670	ENST00000321217	protein_coding	protein_coding	TOIE2LY		TGER-induced feel	In homeohox 2.1	ke Vlieked (Source)	HONC S	mhol-Acc-1	85601	CCD848201			
16	ENSGOOD	00176679	ENST00000559055	protein_coding	protein_coding	TGIE2LY		TGEB-induced fact	or homeohox 2-li	ke, Y-linked (Source)	HGNC S	mbol:Acc:1	85601	CCD	\$14775		
17	ENSGOOD	00099715	ENST00000400457	protein_coding	protein_coding	PCDH11Y		protocadherin 11 Y	linked [Source:H	GNC Symbol: Acc:15	8131	111000,0 100.1	00001	000			
18	ENSG000	00099715	ENST00000333703	protein coding	protein coding	PCDH11Y		protocadherin 11 Y	linked (Source:H	GNC Symbol:Acc:1	813]			CCD	514776		
19	ENSGOOD	00099715	ENST00000362095	protein coding	protein coding	PCDH11Y		protocadherin 11 Y	linked [Source:H	GNC Symbol;Acc:15	813]			CCD	\$14777		
20	ENSG000	00099715	ENST00000215473	protein_coding	protein_coding	PCDH11Y		protocadherin 11 Y	linked [Source:H	GNC Symbol;Acc:15	813]						
21	ENSG000	00168757	ENST00000320701	protein_coding	protein_coding	TSPY2		testis specific prote	in, Y-linked 2 [So	urce:HGNC Symbol	Acc:2392	4]		CCD	<u>535465</u>		
22	ENSG000	00168757	ENST00000383042	protein_coding	protein_coding	TSPY2		testis specific prote	in, Y-linked 2 [So	urce:HGNC Symbol	Acc:2392	4]					
23	ENSG000	00168757	ENST00000470569	protein_coding	retained_intron	TSPY2		testis specific prote	in, Y-linked 2 [So	urce:HGNC Symbol	Acc:2392	4]					
24	ENSGOOD	00168757	ENST00000464674	protein_coding	retained_intron	TSPY2		testis specific prote	in, Y-linked 2 [So	urce:HGNC Symbol	Acc:2392	[4]					
25	ENSGOOD	00099721	ENST00000215479	protein_coding	protein_coding	AMELY		amelogenin, Y-linki	ed [Source:HGN0	Symbol;Acc:462]				CCD	514778		
20	ENSGUUU	00099721	ENS100000383035	protein_coding	protein_coding	AMELY		amelogenin, Y-linki	ed [Source:HGNC	Symbol;Acc:462]							
20	ENECODE	000000000000000000000000000000000000000	ENS100000383037	protein_coding	protein_cooling	TRUAN		amelogenin, t-linki	to 1 V linked IPr	Symbol;Acc:462]	A	201		000	244770		
20	ENROOOD	00002377	ENGT00000363032	protein_coding	protein_cooling	TRLAV		transducin (beta) li	ke 1, 1-linked [Sc	urce.HONC Symbol	Acc: 1000	12]		CCD	214770		
30	ENSGOOD	00092377	ENST00000346432	protein_coding	protein_coding	TBL1Y		transducin (beta)-li	ke 1. Y-linked [Sc	urce:HGNC Symbol	Acc: 1850	121		CCD	\$14779		
31	ENSG000	00092377	ENST00000540077	protein coding	protein coding	TBL1Y		transducin (beta)-li	ke 1. Y-linked [Sc	urce:HGNC Symbol	Acc: 1850	121					
32	ENSGOOD	00233803	ENST00000426950	protein coding	protein coding	TSPY4		testis specific prote	in, Y-linked 4 (So	urce:HGNC Symbol	Acc:3728	7		CCD	\$48202		
33	ENSG000	00233803	ENST00000383008	protein coding	protein coding	TSPY4		testis specific prote	in, Y-linked 4 (So	urce:HGNC Symbol	Acc:3728	7					
34	ENSG000	00233803	ENST00000466036	protein coding	processed trans	cript TSPY4		testis specific prote	in, Y-linked 4 [So	urce:HGNC Symbol	Acc:3728	7					
35	ENSG000	00233803	ENST00000482082	protein_coding	processed_trans	cript TSPY4		testis specific prote	in, Y-linked 4 [So	urce:HGNC Symbol	Acc:3728	7					
36	ENSG000	00229549	ENST0000287721	protein_coding	protein_coding	TSPY8		testis specific prote	in, Y-linked 8 [So	urce:HGNC Symbol	Acc:3747	1]					
37	ENSGOOD	00229549	ENST0000383000	protein_coding	protein_coding	TSPY8		testis specific prote	in, Y-linked 8 [So	urce:HGNC Symbol	Acc:3747	1]					
38	ENSG000	00229549	ENS100000477879	protein_coding	processed_trans	cript TSPY8		testis specific prote	in, Y-linked 8 [So	urce:HGNC Symbol	Acc:3747	1				-	
39	ENSGOOD	00229549	ENS10000436159	protein_coding	processed_trans	cnpt TSPY8		testis specific prote	in, Y-linked 8 [So	urce:HGNC Symbol	ACC:3747	1					
40	ENSCOOL	00229549	ENST00000320628	protein_coding	protein_coding	TSPY8		testis specific prote	in, T-linked 8 [So	urce:HGNC Symbol	ACC:3747	1				-	
-		4 4 5 5	Sheet1 +	arotein cound	Diotoni Cound	13410		Tobes auditine prote		area.nonc aviitor	nu.3/4/						

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 <u>http://e66.ensembl.org</u>. 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)

The Go to the Ensembl homepage (<u>http://www.ensembl.orq</u>).

"∂ Select 'Search: Human' and type '6:32540000-32620000' in the 'for' text box.

^A Click [Go].

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

A Click [Configure this page] in the side menu.

Click [Reset configuration].

∽ Click (√).

(b)

 \mathcal{A} 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.

A Click on 'Regulation - Open chromatin & TFBS'.

A Click on the 'Track style' box in the 'MultiCell' column and select 'Both'.

∽ Ĉlick (√).

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

The click [Configure this page] in the side menu.

A Click on 'Regulation - Open chromatin & TFBS'.

"Hover over 'DNase1' in the configuration matrix and uncheck 'Select all DNase1'.

A Click on 'Track style: Enable/disable all' and select 'Off'.

The 'Click on the 'Track style' box in the 'MultiCell' column and select 'Both'. The Click (\checkmark).

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):

The side menu.

A Click on 'Regulation - Open chromatin & TFBS'.

The 'Click on the 'Track style' box in the 'MultiCell' column and select 'Both'. \mathcal{C} Click (\checkmark).

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.

A 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)

A Click [Configure this page] in the side menu.

A Click on 'Regulation - Regulatory features'.

 \mathcal{A} Click on 'Enable/disable all Segmentation features' and select 'On'.

∽ Ĉlick (✓).

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/wgEncodeCalt echRnaSeq/wgEncodeCaltechRnaSeqGm12878R2x75Th1014II200SigRep1V 4.bigWig

and

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCalt echRnaSeq/wgEncodeCaltechRnaSeqK562R2x75Th1014II200SigRep1V4.bi gWig

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

- The Go to the Ensembl homepage (<u>http://www.ensembl.org</u>).
- A Select 'Search: Human' and type 'bcl11a' in the 'for' text box.
- ^A Click [Go].
- A Click on 'Gene' on the page with search results.
- ^A 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:

- A Click [Configure this page] in the side menu.
- Click [Reset configuration].
- ∽ Click (√).
- A Click [Manage your data] in the side menu.
- Attach Remote File'.
- ${}^{\prime \ominus}$ Enter the URL of the first file in the 'File URL' text box.

- ∽ Click (√).
- ${}^{\circ}$ 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:

A Click [Manage your data] in the side menu.

- ${}^{\prime \ominus}$ 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 (<u>http://www.ensembl.org</u>).
- A Click on the 'BioMart' link on the toolbar.
- A Choose the 'Ensembl Regulation 66' database.
- The Choose the 'Homo sapiens genes (GRCh37.p6)' dataset.
- The 'REGULATORY FEATURES' section by clicking on the + box.
- ~ Enter 'Base pair Start (bp): 128573000' and 'End (bp): 128745000'.
- Attributes' in the left panel.
- A Deselect 'Feature Set' and 'Feature Type Description'.
- A 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.