Chapter 6

The Ensembl Genome Browser: Strategies for Accessing Eukaryotic Genome Data

Victoria Newman, Benjamin Moore, Helen Sparrow, and Emily Perry

Abstract

The Ensembl Genome Browser provides a wealth of freely available genomic data that can be accessed for many purposes by genetics, genomics, and molecular biology researchers. Herein we present two protocols for exploring different aspects of these data: a phenotype and its associated variants and genes, and a promoter and the epigenetic marks and protein-binding activity associated with it. These workflows illustrate a subset of the data types available through the Ensembl Browser, and can be considered a springboard for further exploration.

Key words Ensembl, Eukaryotic genomes, Phenotypes, Variants, Epigenetic mark

1 Introduction

Genome browsers are resources that integrate data at the genomic level, thereby allowing visualization of related genomic information in one space. These data can include genes, noncoding elements that regulate gene expression, genetic variation and the results of comparative genomics analyses, among other forms of annotation (Fig. 1) [1–4]. Commonly used genome browsers include Ensembl, the UCSC Genome Browser [5] and IGV [6].

The Ensembl project was initially launched in 1999 with the aim of developing methodologies for automatic annotation of (human) genomic sequence with genes and their constituent transcripts [7]. Since that time, the project has broadened substantially in scope; the Ensembl Genome Browser [8], which came online in 2000, now includes reference genomic sequence and annotation for nearly 100 chordate organisms. Ensembl is rapidly incorporating new data, including whole clades of new species' genomes and reference sequence for multiple strains of existing species, such as mouse. In addition, existing annotation is regularly augmented by the inclusion of new data sets. Ensembl's sister site, Ensembl Genomes,

Martin Kollmar (ed.), *Eukaryotic Genomic Databases: Methods and Protocols*, Methods in Molecular Biology, vol. 1757, https://doi.org/10.1007/978-1-4939-7737-6_6, © The Author(s) 2018

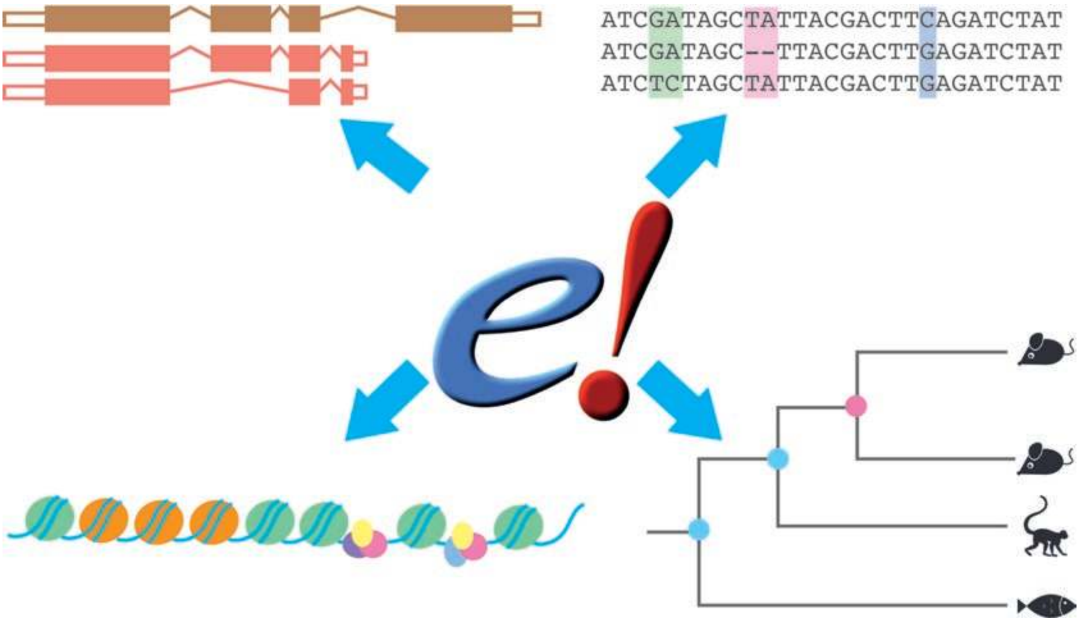
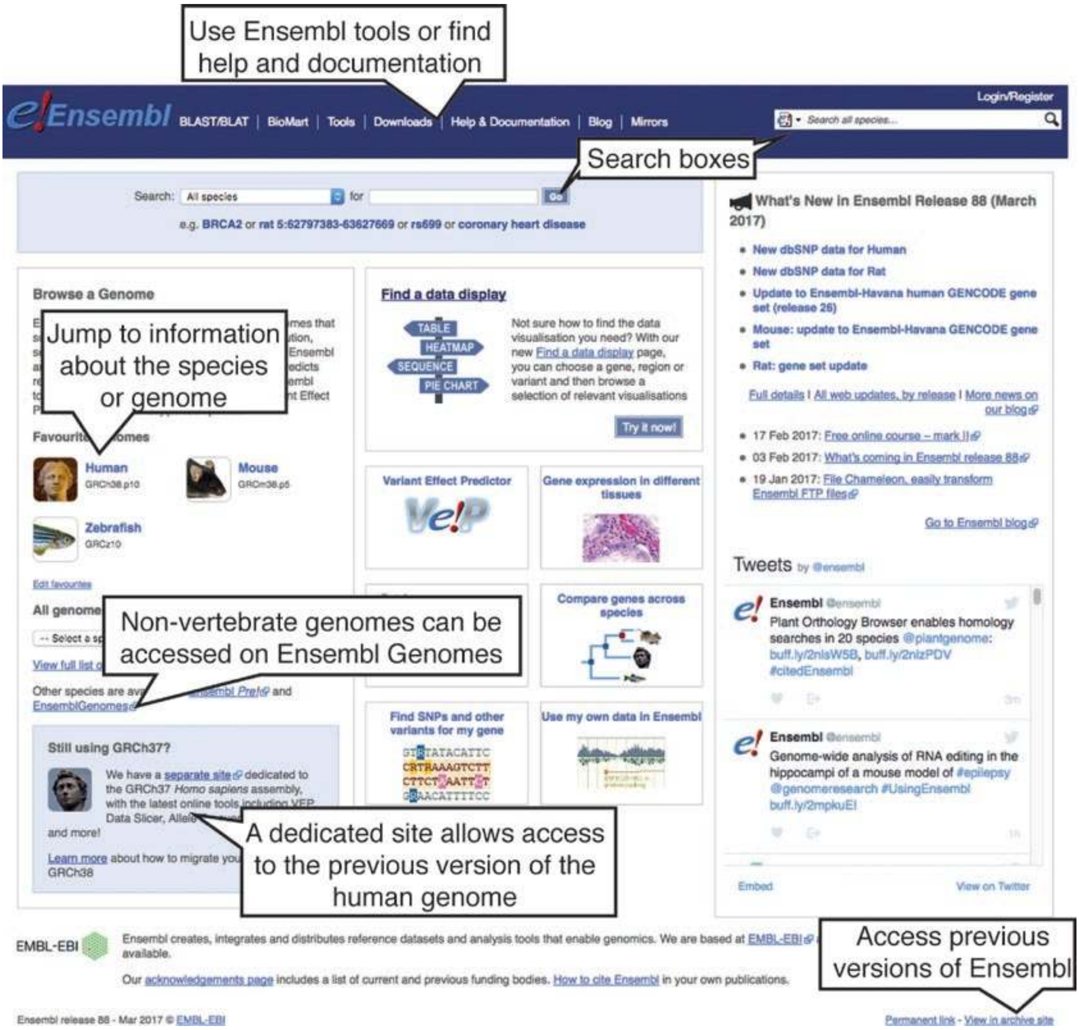

ATCGATAGCTATTACGACTTCAGATCTAT
ATCGATAGC--TTACGACTTGAGATCTAT
ATCTCTAGCTATTACGACTTGAGATCTAT

Fig. 1 Ensembl features. Ensembl integrates together gene annotation, genetic variation, gene regulation data, and comparative genomics onto a single genomic platform. Gene annotation is carried out in house, annotating the full intron–exon structure of coding and noncoding transcripts. Short variants, such as SNPs and indels, are pulled into Ensembl from external databases, alongside structural variants and copy-number variants. ChIP-seq and DNase-seq data is used for in-house prediction of regions of open chromatin and regulatory elements such as promoters, enhancers and CTCF binding sites on the genome and their activity in different cell types. Whole genome alignments and gene tree analysis is carried out in house to compare species in Ensembl. These data are presented alongside each other on the genome in the Ensembl browser, and can also be accessed for bulk export through BioMart, programmatically through APIs and as flat-files on the Ensembl FTP site

provides access to nonvertebrate genomes through dedicated portals for Bacteria, Fungi, Plants, Metazoa, and Protists [9, 10].

Ensembl data, annotations, and analyses are updated every 2–3 months, alongside software updates to both the public-facing website and the underlying databases. Prior releases are frozen as archive sites, and from Dec 2013 (Ensembl version 74) will remain accessible via our web interface for at least 5 years following their initial release. A dedicated site is also maintained for the GRCh37 reference human genome assembly, which is annotated with new data on a limited basis (Fig. 2) [11]; partial data from ongoing genome annotation can be accessed via the preview Pre! site.

Data from Ensembl can be accessed at multiple scales. In this chapter, we describe data access through the browser web pages and via BioMart [12], a web-based tool that allows customized retrieval of data from the Ensembl databases. However, data can also be accessed programmatically via our Perl and REST APIs [13, 14] Files containing genome-wide data are available for all species represented in Ensembl via an FTP site [15]; data from all releases



Use Ensembl tools or find help and documentation

Search boxes

Jump to information about the species or genome

Non-vertebrate genomes can be accessed on Ensembl Genomes

A dedicated site allows access to the previous version of the human genome

Access previous versions of Ensembl

Login/Register
e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search all species...
Search: All species for Go
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease
Browse a Genome
Favourite genomes
Human GRCh38.p10
Mouse GRCm38.p5
Zebrafish GRCz10
Edit favourites
All genomes
-- Select a sp
View full list o
Other species are av ... Ensembl Pre! and EnsemblGenomes
Find a data display
TABLE
HEATMAP
SEQUENCE
PIE CHART
Try it now!
Variant Effect Predictor
VEP
Gene expression in different tissues
Compare genes across species
Find SNPs and other variants for my gene
Use my own data in Ensembl
Tweets by @ensembl
Embed
View on Twitter
EMBL-EBI
Ensembl release 88 - Mar 2017 © EMBL-EBI
Permanent link - View in archive site

Still using GRCh37?

We have a separate site dedicated to the GRCh37 *Homo sapiens* assembly, with the latest online tools including VEP, Data Slicer, Allele ... and more!

Learn more about how to migrate you ... GRCh38

Not sure how to find the data visualisation you need? With our new Find a data display page, you can choose a gene, region or variant and then browse a selection of relevant visualisations

What's New in Ensembl Release 88 (March 2017)

- New dbSNP data for Human
- New dbSNP data for Rat
- Update to Ensembl-Havana human GENCODE gene set (release 26)
- Mouse: update to Ensembl-Havana GENCODE gene set
- Rat: gene set update

Full details | All web updates, by release | More news on our blog

- 17 Feb 2017: Free online course – mark II
- 03 Feb 2017: What's coming in Ensembl release 88
- 19 Jan 2017: File Chameleon, easily transform Ensembl FTP files

Go to Ensembl blog

Ensembl @ensembl
Plant Orthology Browser enables homology searches in 20 species @plantgenome: buff.ly/2nlsW5B, buff.ly/2nlzPDV #citedEnsembl

Ensembl @ensembl
Genome-wide analysis of RNA editing in the hippocampi of a mouse model of #epilepsy @genomeresearch #UsingEnsembl buff.ly/2mpkuEI

Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI ... available.

Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Fig. 2 The Ensembl homepage. The Ensembl homepage provides access to a search function, which can retrieve information associated with, for example, genes, transcripts, proteins, variants, phenotypes, and ontology terms. In addition, links are available to Ensembl's sister site, Ensembl Genomes, as well as to the most-searched genomes and a complete list of annotated genomes. Fully annotated genomes are available on the main Ensembl site, while genomes whose annotation is in process can be browsed on the Ensembl Pre! site. Ensembl maintains web interfaces of archived versions for 5 years. These can be accessed from a link in the lower right-hand corner. Documentation and help pages can be accessed from the homepage, as well as in-house and external tools integrated into the Ensembl web interface. A dedicated page describing data-download strategies is also available and presents links to the point-and-click tool BioMart, which permits bulk download of Ensembl datasets with no requirement for programming expertise, as well as APIs and FTP site

of Ensembl can be retrieved from the FTP site, or from our databases via the Perl APIs, in perpetuity.

Beyond providing access to data related to publicly available genome annotation, Ensembl integrates a number of tools designed to process or analyze your own data. The ID History Converter converts Ensembl IDs from a previous release into their current equivalents, while the Assembly Converter maps genomic coordinates from one version of a genome assembly to another. The Variant Effect Predictor predicts the functional consequences of a set of known and/or novel variants [16]. Sequence alignment using BLAST and BLAT against Ensembl genes, genomes and proteins is also available [17, 18], along with a suite of tools developed as part of the 1000 Genomes Project [19] that can be accessed on the dedicated GRCh37 browser site [11].

In this chapter we describe two workflows showcasing a subset of the data available in the Ensembl browser and indicating possible routes to access them. First, we demonstrate a phenotype-centric search highlighting variation data associated with genes and transcripts. Secondly, we present a gene-centric search illustrating gene and transcript models, and the exploration of regulatory features in the region of a gene. In each case we also indicate strategies for data export via BioMart. Those interested in our annotation methods, in programmatic access to Ensembl data, or in exploring other forms of data and annotation are encouraged to refer to our publications [20].

2 Materials

Computer, Internet connection.

An Internet browser: recent versions of Firefox, Chrome, Safari, and Internet Explorer are supported.

3 Methods

These workflows were written using Ensembl release 88 (March 2017). There may be updates to the data or interfaces if you are using a more recent release.

3.1 WF1: Phenotype-Based Searches and Identification of Associated Genetic Variation

The Ensembl browser can be searched using a variety of terms, including genomic regions, genes, variants, or phenotypes; the following workflow describes a phenotype-based search that highlights data and annotations collated in the Phenotype, Variant, Gene, and Transcript tabs.

Non-melanoma skin cancer—principally basal cell and squamous cell carcinomas—is a relatively common pathology associated with variants in several genes [21].

1. *Getting started*: To explore the phenotype in more detail, type "non-melanoma skin cancer" into the search box on the Ensembl home page, www.ensembl.org, and click the "Go" button. The search autocomplete may retrieve direct links to

suggested results; this will allow you to proceed immediately to **step 2**.

A list of search results will be generated, with "Non-melanoma skin cancer (Human Phenotype)" appearing first. Options on the left-hand side of the page permit restriction by species and/or other categories: click on the different filters individually to apply them to the search results.

2. *Studying loci associated with a phenotype*: Click the "Non-melanoma skin cancer (Human Phenotype)" link to open the Phenotype tab.

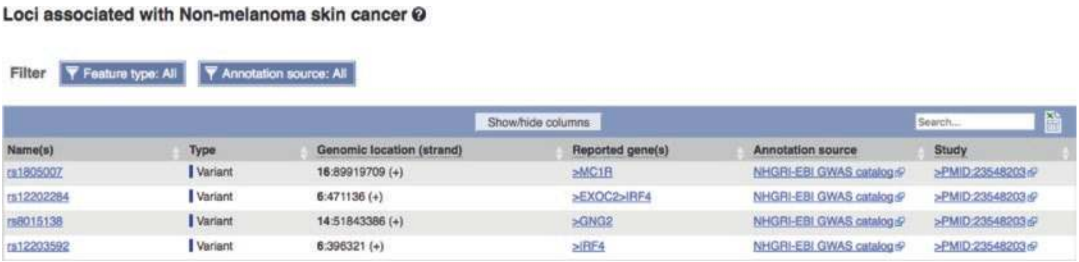
   The loci associated with non-melanoma skin cancer are presented in tabular form; their external identifiers, genomic coordinates and associated genes, and the publications in which they were initially described are all listed. Links are provided to further information about the annotation source and relevant publications (in this case, the GWAS catalog [22] *and PubMed* [23]*;* Fig. 3*).*

3. *Studying a variant*: One of the variants associated with non-melanoma skin cancer, rs1805007, falls within the *MC1R* gene. Click the "rs1805007" link to load the Variant tab.

   The Variant tab collates data relating specifically to the variant of interest (A full list of the databases from which Ensembl imports variation data can be found in the documentation [24].*).*

   An overview of the data is found at the top of the Variant tab (Fig. 4A)*, while a table indicating the phenotypes associated with the variant can be found lower down the page.*

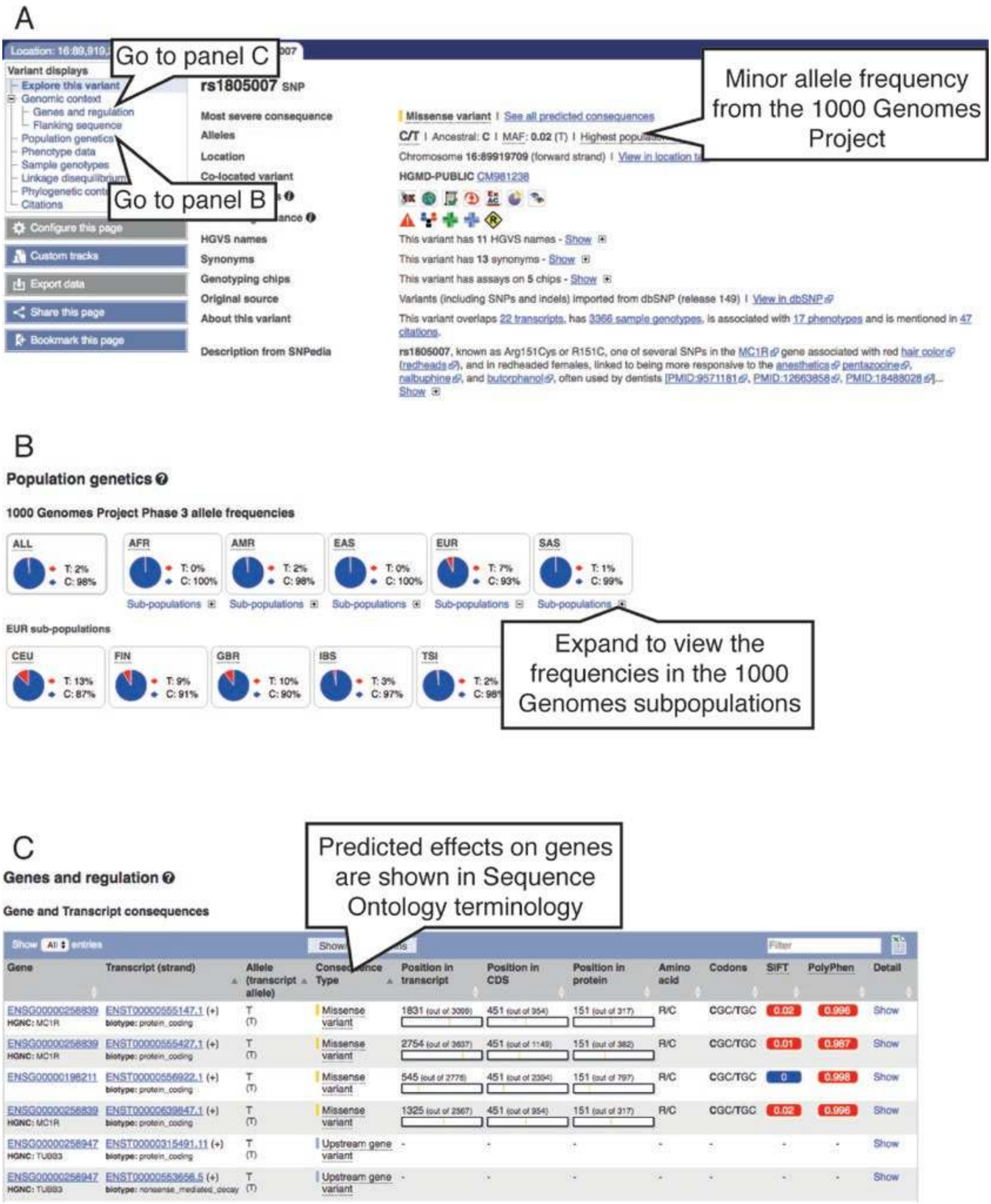
   The most severe consequence linked to rs1805007 is "missense_variant", indicating that the alternative allele at this locus lead to an amino acid substitution. All consequences of the rs1805007 variant can be explored by clicking on the "See all predicted consequences" link. Ensembl uses Sequence Ontology terms to describe variant consequences [25].



Loci associated with Non-melanoma skin cancer

Name(s)	Type	Genomic location (strand)	Reported gene(s)	Annotation source	Study
rs1805007	Variant	16:89919709 (+)	>MC1R	NHGRI-EBI GWAS catalog	>PMID:23548203
rs12202284	Variant	6:471136 (+)	>EXOC2>IRF4	NHGRI-EBI GWAS catalog	>PMID:23548203
rs8015138	Variant	14:51843386 (+)	>GNG2	NHGRI-EBI GWAS catalog	>PMID:23548203
rs12203592	Variant	6:396321 (+)	>IRF4	NHGRI-EBI GWAS catalog	>PMID:23548203

Fig. 3 The Ensembl phenotype tab. The Ensembl phenotype tab allows you to explore the phenotype ontology associated with a phenotype and any loci (variants, QTLs, or genes) linked to the phenotype. Loci associated with the phenotype shown in a table on the Associated loci page. The buttons above the table allow filtering. Links take you to the database and/or paper where the link between locus and phenotype was made



rs1805007 SNP

Most severe consequence: Missense variant | See all predicted consequences

Alleles: C/T | Ancestral: C | MAF: 0.02 (T) | Highest popula

Location: Chromosome 16:89919709 (forward strand) | View in location t

Co-located variant: HGMD-PUBLIC CM981238

HGVS names: This variant has 11 HGVS names - Show

Synonyms: This variant has 13 synonyms - Show

Genotyping chips: This variant has assays on 5 chips - Show

Original source: Variants (including SNPs and indels) imported from dbSNP (release 149) | View in dbSNP

About this variant: This variant overlaps 22 transcripts, has 3366 sample genotypes, is associated with 17 phenotypes and is mentioned in 47 citations.

Description from SNPedia: rs1805007, known as Arg151Cys or R151C, one of several SNPs in the MC1R gene associated with red hair color (redheads), and in redheaded females, linked to being more responsive to the anesthetics pentazocine, nalbuphine, and butorphanol, often used by dentists [PMID:9571181, PMID:12663858, PMID:18488028]... Show

Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen	Detail
ENSG00000258839 HGNC: MC1R	ENST00000555147.1 (+) biotype: protein_coding	T (T)	Missense variant	1831 (out of 3009)	451 (out of 954)	151 (out of 317)	R/C	CGC/TGC	0.02	0.996	Show
ENSG00000258839 HGNC: MC1R	ENST00000555427.1 (+) biotype: protein_coding	T (T)	Missense variant	2754 (out of 3607)	451 (out of 1149)	151 (out of 382)	R/C	CGC/TGC	0.01	0.987	Show
ENSG00000198211	ENST00000556922.1 (+) biotype: protein_coding	T (T)	Missense variant	545 (out of 2776)	451 (out of 2394)	151 (out of 797)	R/C	CGC/TGC	0	0.998	Show
ENSG00000258839 HGNC: MC1R	ENST00000639847.1 (+) biotype: protein_coding	T (T)	Missense variant	1325 (out of 2567)	451 (out of 954)	151 (out of 317)	R/C	CGC/TGC	0.02	0.996	Show
ENSG00000258947 HGNC: TUBB3	ENST00000315491.11 (+) biotype: protein_coding	T (T)	Upstream gene variant	-	-	-	-	-	-	-	Show
ENSG00000258947 HGNC: TUBB3	ENST00000553656.5 (+) biotype: nonsense_mediated_decay	T (T)	Upstream gene variant	-	-	-	-	-	-	-	Show

Fig. 4 The Ensembl variation tab. The Ensembl variation tab provides a wealth of information about a particular variant, such as a SNP or indel. (A) A variant summary shown on all pages in the variant tab, including alleles, MAF, and evidence status. The menu at the left-hand side provides links to all the pages providing information on the variant. (B) Pie charts from the Population Genetics page, showing the allele frequencies for the variant in the 1000 Genomes populations. (C) The Genes and Regulation table, listing all genes affected by the variant with details of sequence ontology consequences, position in the gene and protein, and SIFT and Polyphen scores for amino acid changes (where relevant)

Below the consequence, you can see that the reference allele of rs1805007 at the genomic position 16:89919709 is C, and one alternative allele, T, has been observed. Minor allele frequency (MAF) has been calculated for the alternative allele, which was observed in 1000 Genomes Project participants: it was identified in 2% of participants in that study [2, 26].

Navigating to the Variant tab from the Phenotype tab automatically loads a table containing the phenotype data relating to this variant, as mentioned above. Tanning ability, sensitivity to sun, and fair hair and skin color have all been associated with the variant, as has basal cell carcinoma, a form of non-melanoma skin cancer. Collectively, these phenotypes are consistent with the observed linkage between fair complexions and sensitivity to sun exposure.

4. The menu on the left presents additional options. Click "Population genetics" to view allele frequencies in global populations.

   On this page, data from the 1000 Genomes [26], *HapMap* [27], *and NHLBI Exome Sequencing* [28] *Projects and the Exome Aggregation Consortium (ExAC)* [29] *are displayed. The data from the 1000 Genomes Project are shown at the top* (Fig. 4B)*; the pie-charts represent allele frequencies for different superpopulations. Allele frequencies for subpopulations within each superpopulation can be viewed by clicking the "Subpopulations" link beneath the corresponding superpopulation. Allele and genotype frequencies among 1000 Genomes Project participants can also be found in tabular form immediately below the graphical views.*

   The frequency of the T variant allele in 1000 Genomes Project participants is highest among European subgroups, and individuals homozygous for the variant also occur only in these subgroups. This is expected given the phenotypes associated with the variant (Fig. 4B).

5. To explore genes and transcripts with which the variant is associated, click "Genes and regulation" in the menu on the left.

   As we saw previously, the variant lies within the MC1R *gene; the summary table here indicates that it overlaps two independent transcripts of this gene as a missense variant and is a downstream gene variant of a third transcript. Other genes and transcripts affected by the variant, as well as the associated consequences, are also shown* (Fig. 4C).

   In a second table, called "Gene expression correlations," you can find a list of genes whose expression has been found by the GTEx Project to be affected by the variant of interest [30].

   Finally, any regulatory features or motifs in which the variant falls will be listed in two separate tables at the bottom of the page. There are no regulatory features or motifs that overlap the variant rs1805007.

6. *Studying a gene and its transcripts*: Click "ENSG00000258839" in the Genes and regulation table to go directly to the Gene tab, which collates gene-related information, for *MC1R*.

   Navigating to the Gene tab from the Variant tab loads the Variant table, which lists all variants in the Ensembl database that fall within the gene itself or in the region 5 kb upstream or downstream of the gene. The top of the page presents a short overview of MC1R*, including a description of the gene, its genomic location and synonyms, and an option to show a table of all its transcripts. This information can also be found at the top of all subsequent views within the Gene tab. As in other tabs in the Ensembl browser, the menu to the left of the Gene tab presents links to a variety of additional data and annotations* (Fig. 5A).

7. Click "Summary" in the left-hand menu.

   General information about the gene, including a description, synonyms and the genomic location, can be found in this view. A graphical model of the gene's transcripts is shown at the bottom (Fig. 5A).

8. For the complete set of phenotypes associated with *MC1R*, click the "Phenotypes" link in the left-hand menu.

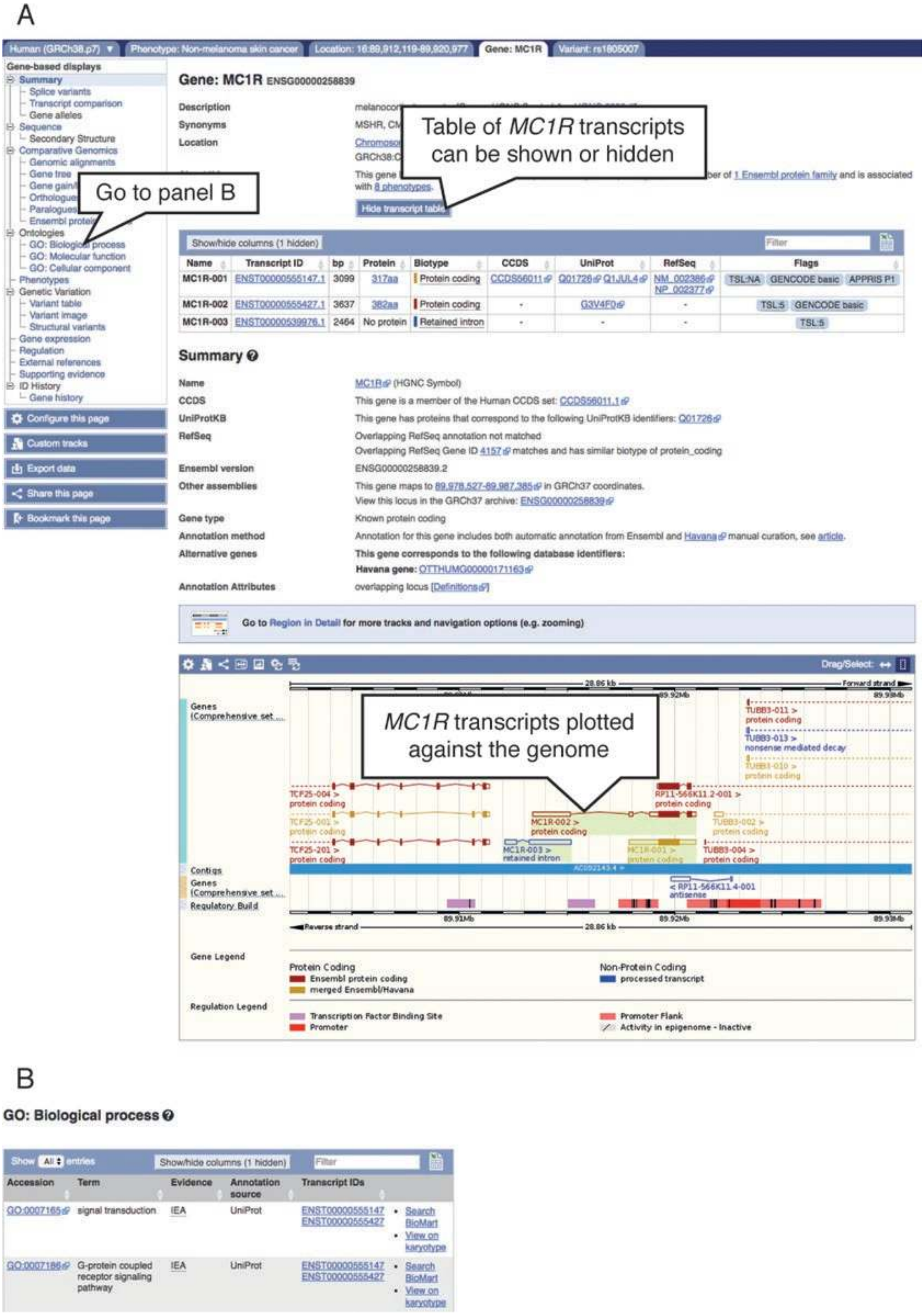
   The three tables list phenotypes associated with the gene, with variants in the gene, and with other species' orthologues of the gene, as predicted by the Ensembl comparative genomics pipeline [3]. *Several phenotypes have been linked to rs1805007, and the* MC1R *gene also plays a role in coat and skin pigmentation in other organisms, suggesting a conserved function.*

9. Click on the "GO: Biological process" link in the left-hand menu.

   The GO, or Gene Ontology, terms related to biological processes which have been associated with the transcripts of the MC1R *gene are displayed in the table* (Fig. 5B) [31, 32]. *Each row of the table contains the GO accession number, a description of the GO term, and the evidence codes, annotation source and stable IDs of transcripts associated with that GO term. Hover over the evidence codes to see their definitions.*

   MC1R*-encoded proteins are involved in signal transduction and the melanin biosynthesis pathway, and are located in the plasma membrane, consistent with a role in pigmentation.*

Fig. 5 (continued) against the genome. The central contig indicates the genome. Positive stranded genes, such as *MC1R* are depicted above the contig. Strand is also indicated by an arrow alongside the transcript name indicating the direction of transcription, and by introns, which are shown pointing upward on positive stranded genes and downward on negative stranded genes. Some transcripts have been removed from this image for size. On all pages in the gene tab, a menu on the left-hand side lists all the pages available for looking at a gene. (B) Three pages are available for looking at the GO terms associated with a gene, conforming to the three categories of terms, Biological process, Molecular function, and Cellular component. These are listed for each gene, including which transcript they are associated with and how they were annotated



Go to panel B

Table of *MC1R* transcripts can be shown or hidden

Gene: MC1R ENSG00000258839

Description melanocor
Synonyms MSHR, CM
Location Chromoso
GRCh38:C
This gene ber of 1 Ensembl protein family and is associated with 8 phenotypes.

Hide transcript table

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
MC1R-001	ENST00000555147.1	3099	317aa	Protein coding	CCDS56011	Q01726 Q1JUL4	NM_002386 NP_002377	TSL:NA GENCODE basic APPRIS P1
MC1R-002	ENST00000555427.1	3637	382aa	Protein coding	-	G3V4F0	-	TSL:5 GENCODE basic
MC1R-003	ENST00000539976.1	2464	No protein	Retained intron	-	-	-	TSL:5

Summary

Name	MC1R (HGNC Symbol)
CCDS	This gene is a member of the Human CCDS set: CCDS56011.1
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: Q01726
RefSeq	Overlapping RefSeq annotation not matched Overlapping RefSeq Gene ID 4157 matches and has similar biotype of protein_coding
Ensembl version	ENSG00000258839.2
Other assemblies	This gene maps to 89,978,527-89,987,385 in GRCh37 coordinates. View this locus in the GRCh37 archive: ENSG00000258839
Gene type	Known protein coding
Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.
Alternative genes	**This gene corresponds to the following database identifiers:** **Havana gene:** OTTHUMG00000171163
Annotation Attributes	overlapping locus [Definitions]

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

MC1R transcripts plotted against the genome

GO: Biological process

Show All entries Show/hide columns (1 hidden) Filter

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0007165	signal transduction	IEA	UniProt	ENST00000555147 ENST00000555427	• Search BioMart • View on karyotype
GO:0007186	G-protein coupled receptor signaling pathway	IEA	UniProt	ENST00000555147 ENST00000555427	• Search BioMart • View on karyotype

Fig. 5 The Ensembl gene tab. The Ensembl gene tab provides a number of views to look at different aspects of a gene. (A) The gene summary page includes a graphical depiction of the transcripts of the gene, shown

Two further links in the menu at left provide GO term associations regarding the Molecular Function and Cellular Component corresponding to transcripts of the MC1R *gene* (Fig. 5A).

10. Click on the "External references" link in the left-hand menu.

    Links to records in external databases such as EntrezGene [33], *HGNC* [34], *and MIM Gene and MIM Morbid* [35] *can be found on this page.*

11. *Studying a transcript*: Click the "Show transcript table" button in the "Transcripts" section at the top of the page.

    A tabular view of the individual transcripts comprising the gene model can be seen (for more information on the Ensembl gene annotation strategy, see ref. 1*). This table displays information about transcript length and biotype, as well as links to the entries in the CCDS* [36], *UniProt* [37], *and RefSeq* [33] *databases that correspond to particular transcripts.*

    The level of support for a transcript prediction, and its biological relevance, can be inferred from the matching evidence records and associated flags.

12. Click the "ENST00000555147.1" link in the Transcript table; ENST00000555147.1 is the Ensembl stable ID for the *MC1R-001* transcript.

    The MC1R-001 *transcript's biotype is listed as "protein-coding," and the transcript is colored golden in the graphical view. This indicates that it has been independently annotated with identical coordinates by both the Ensembl automated gene annotation and the HAVANA manual gene annotation methods* [1] (Fig. 5A).

    We are now located in the Transcript tab, which is visible in the blue navigation bar at the top of the page, next to the Gene tab. From the left-hand menu of the Transcript tab you can access complete, spliced or translated transcript sequences ("Exons," "cDNA," and "Protein," respectively), as well as graphical and tabular representations of annotated protein domains ("Protein summary" and "Domains & features," respectively). "General identifiers" provides links to related records in external repositories (Fig. 6A).

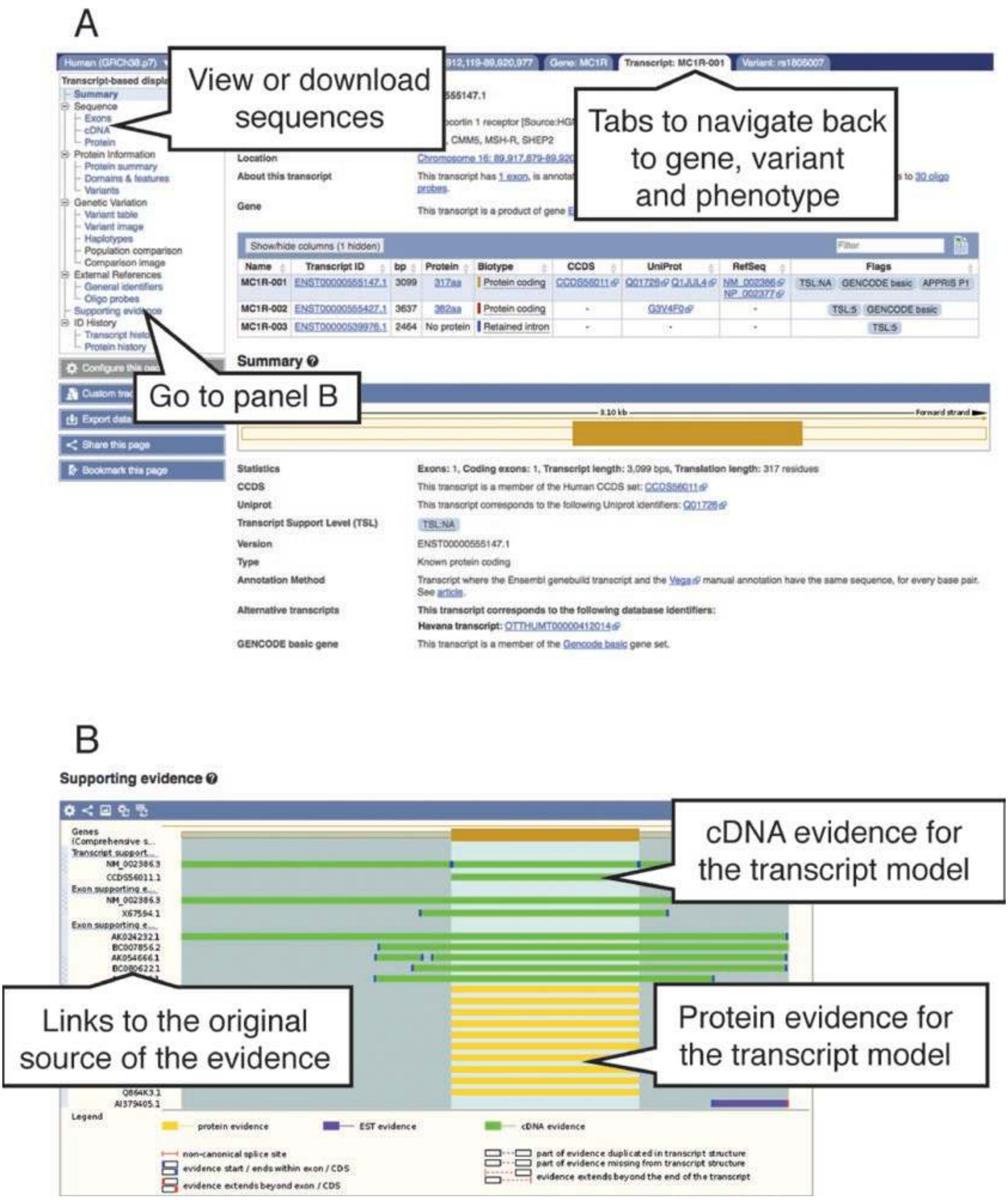
    You can now click on "Hide Transcript table" in the Gene section at the top of the page to remove the Transcript table from the page view.

13. Click on the "Supporting evidence" link in the left-hand menu.

    This page displays the records used in the annotation in graphical form; all records are hyperlinked to the original data in RefSeq, UniProt, and ENA [38] (Fig. 6B).

14. Click the "Variant table" link in the left-hand menu.

    This table displays the set of variants associated with the MC1R-001 *transcript* (Fig. 7A).



555147.1

cortin 1 receptor [Source:HG

, CMM5, MSH-R, SHEP2

Location: Chromosome 16: 89,917,879-89,920

About this transcript: This transcript has 1 exon, is annotat … s to 30 oligo probes.

Gene: This transcript is a product of gene E

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
MC1R-001	ENST00000555147.1	3099	317aa	Protein coding	CCDS56011	Q01726 Q1JUL4	NM_002386 NP_002377	TSL:NA GENCODE basic APPRIS P1
MC1R-002	ENST00000555427.1	3637	382aa	Protein coding	-	G3V4F0	-	TSL:5 GENCODE basic
MC1R-003	ENST00000539976.1	2464	No protein	Retained intron	-	-	-	TSL:5

Statistics: Exons: 1, Coding exons: 1, Transcript length: 3,099 bps, Translation length: 317 residues

CCDS: This transcript is a member of the Human CCDS set: CCDS56011

Uniprot: This transcript corresponds to the following Uniprot identifiers: Q01726

Transcript Support Level (TSL): TSL:NA

Version: ENST00000555147.1

Type: Known protein coding

Annotation Method: Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.

Alternative transcripts: This transcript corresponds to the following database identifiers:
Havana transcript: OTTHUMT00000412014

GENCODE basic gene: This transcript is a member of the Gencode basic gene set.

Fig. 6 The Ensembl transcript tab. The transcript tab contains all views for looking at a transcript and its associated protein, where relevant. (A) The left-hand menu on the transcript tab lists all the pages for looking at transcripts and proteins, and differs subtly from the gene tab menu. It has three different sequence views, allowing you to view the exon and intron sequences in a table, an alignment of the cDNA, CDS and peptide sequences, and the protein sequence only. As you open different features, such as genes, transcripts, and variants, tabs appear in the top bar, allowing easy navigation between the different features you've been looking at. (B) The Supporting evidence page shows which cDNA and protein evidence was used to annotate the transcript model

15. Filter the table to view missense variants between amino acid coordinates 150–160.

    (a) Filter the table for missense variants by clicking "Consequence" in the Filter section, then "Turn All Off" and "Missense variant."

    (b) Filter the table to view variants at a specific amino acid coordinate within the translated sequence of the transcript by clicking on "Filter Other Columns," then "AA Coord." Use the sliders to restrict the area for which variants are shown to 150–160.

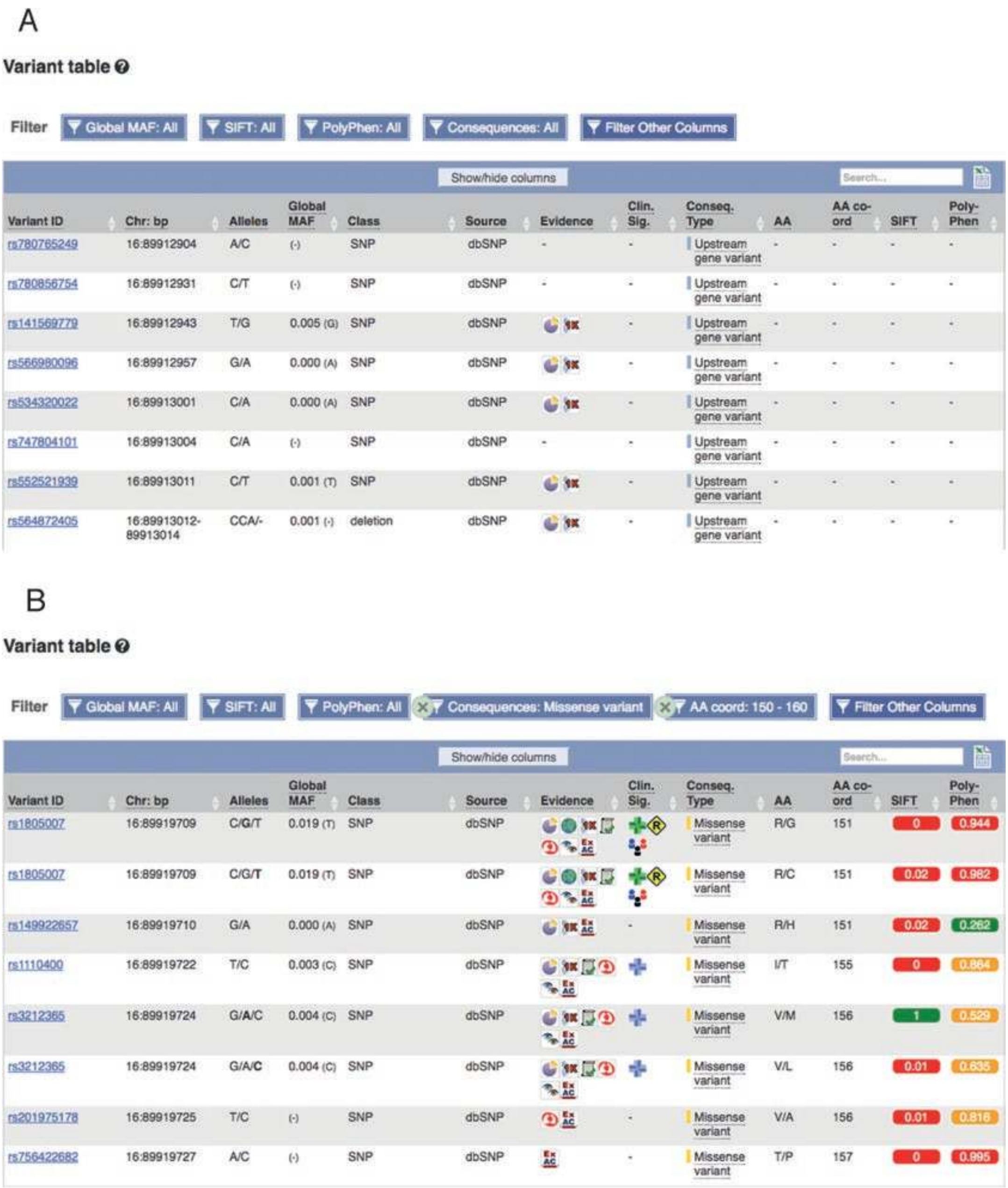
    You can filter this table in numerous ways, including by consequence, source, and genomic or amino-acid coordinates (Fig. 7B). *For missense variants, there are also options to filter by predicted pathogenicity score, as determined by SIFT* [39] *and/or PolyPhen* [40] *(PolyPhen calculations are available only for human variants). SIFT and PolyPhen pathogenicity predictions have been calculated for rs1805007 and the amino acid substitution is considered deleterious (An additional variant, rs149922657, has been observed at the same position of* MC1R, *but has not been associated with any phenotype.).*

16. Click "Haplotypes" in the left-hand menu.

    This page allows you to view linked variants that tend to be coinherited. As a default, the amino acid identities and coordinates of each haplotype are shown, along with their frequencies in different 1000 Genomes Project populations [26]*; however, clicking "switch to CDS view" at the top of the table will show nucleotide sequences instead* (Fig. 8A). *The fifth haplotype listed in the protein-haplotype table represents our variant of interest. The frequency of this haplotype is, as already seen, higher in the European subgroup. Lower in the table can be found the 151R>H haplotype corresponding to rs149922657, the other variant observed at position 151; this variant was recovered in only two 1000 Genomes Project participants.*

    Clicking on any haplotype will load a table indicating its frequencies in different 1000 Genomes populations in more detail (Fig. 8B), *as well as a sequence view highlighting the nucleotide and amino-acid positions altered, if applicable* (Fig. 8C).

17. *Exporting Ensembl variation data*: Data can be exported from Ensembl at multiple scales. A link to the BioMart tool, which permits the download of customized datasets at intermediate scale, can be found in the navigation bar at the top of all Ensembl pages (Fig. 2) [12]. In the BioMart interface, select the Dataset "Ensembl Variation" (this will also include the release number, which is 88 at the time of writing), then "Human Short Variants (SNPs and indels excluding flagged variants)." To download all variants of ≤50 bp lying within



Variant table

Filter: Global MAF: All | SIFT: All | PolyPhen: All | Consequences: All | Filter Other Columns

Show/hide columns

Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA co-ord	SIFT	Poly-Phen
rs780765249	16:89912904	A/C	(-)	SNP	dbSNP	-	-	Upstream gene variant	-	-	-	-
rs780856754	16:89912931	C/T	(-)	SNP	dbSNP	-	-	Upstream gene variant	-	-	-	-
rs141569779	16:89912943	T/G	0.005 (G)	SNP	dbSNP		-	Upstream gene variant	-	-	-	-
rs566980096	16:89912957	G/A	0.000 (A)	SNP	dbSNP		-	Upstream gene variant	-	-	-	-
rs534320022	16:89913001	C/A	0.000 (A)	SNP	dbSNP		-	Upstream gene variant	-	-	-	-
rs747804101	16:89913004	C/A	(-)	SNP	dbSNP	-	-	Upstream gene variant	-	-	-	-
rs552521939	16:89913011	C/T	0.001 (T)	SNP	dbSNP		-	Upstream gene variant	-	-	-	-
rs564872405	16:89913012-89913014	CCA/-	0.001 (-)	deletion	dbSNP		-	Upstream gene variant	-	-	-	-

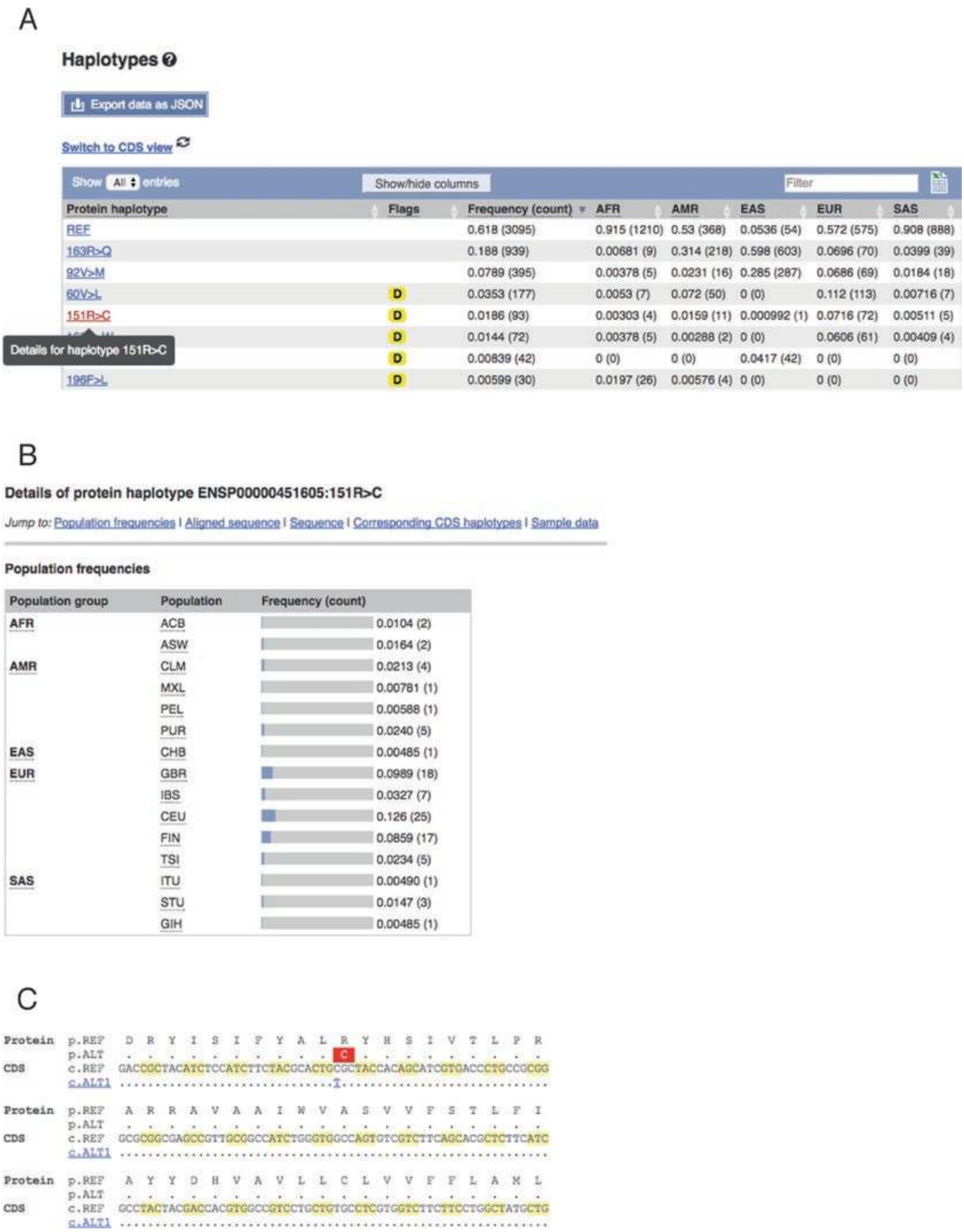
Variant table

Filter: Global MAF: All | SIFT: All | PolyPhen: All | Consequences: Missense variant | AA coord: 150 - 160 | Filter Other Columns

Show/hide columns

Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA co-ord	SIFT	Poly-Phen
rs1805007	16:89919709	C/G/T	0.019 (T)	SNP	dbSNP	ExAC		Missense variant	R/G	151	0	0.944
rs1805007	16:89919709	C/G/T	0.019 (T)	SNP	dbSNP	ExAC		Missense variant	R/C	151	0.02	0.982
rs149922657	16:89919710	G/A	0.000 (A)	SNP	dbSNP	ExAC	-	Missense variant	R/H	151	0.02	0.262
rs1110400	16:89919722	T/C	0.003 (C)	SNP	dbSNP	ExAC		Missense variant	I/T	155	0	0.864
rs3212365	16:89919724	G/A/C	0.004 (C)	SNP	dbSNP	ExAC		Missense variant	V/M	156	1	0.529
rs3212365	16:89919724	G/A/C	0.004 (C)	SNP	dbSNP	ExAC		Missense variant	V/L	156	0.01	0.635
rs201975178	16:89919725	T/C	(-)	SNP	dbSNP	ExAC	-	Missense variant	V/A	156	0.01	0.816
rs756422682	16:89919727	A/C	(-)	SNP	dbSNP	ExAC	-	Missense variant	T/P	157	0	0.995

Fig. 7 Table of short variants found within a transcript. The variant table lists all the variants found within a transcript. A similar page can be found in the gene tab listing all the variants in a gene. The table lists the variants, which are links to the variant tab, with their positions, alleles, SO consequences, and predicted protein effects. Buttons above the table allow you to filter to table to only show variants of interest. (A) The unfiltered table for *MC1R-001*. (B) The same table, filtered to only show missense variants between residues 150 and 160. The applied filters are shown above the table and can be easily removed



Haplotypes

Protein haplotype	Flags	Frequency (count)	AFR	AMR	EAS	EUR	SAS
REF		0.618 (3095)	0.915 (1210)	0.53 (368)	0.0536 (54)	0.572 (575)	0.908 (888)
163R>Q		0.188 (939)	0.00681 (9)	0.314 (218)	0.598 (603)	0.0696 (70)	0.0399 (39)
92V>M		0.0789 (395)	0.00378 (5)	0.0231 (16)	0.285 (287)	0.0686 (69)	0.0184 (18)
60V>L	D	0.0353 (177)	0.0053 (7)	0.072 (50)	0 (0)	0.112 (113)	0.00716 (7)
151R>C	D	0.0186 (93)	0.00303 (4)	0.0159 (11)	0.000992 (1)	0.0716 (72)	0.00511 (5)
[illegible]	D	0.0144 (72)	0.00378 (5)	0.00288 (2)	0 (0)	0.0606 (61)	0.00409 (4)
[illegible]	D	0.00839 (42)	0 (0)	0 (0)	0.0417 (42)	0 (0)	0 (0)
196F>L	D	0.00599 (30)	0.0197 (26)	0.00576 (4)	0 (0)	0 (0)	0 (0)

Details of protein haplotype ENSP00000451605:151R>C

Jump to: Population frequencies | Aligned sequence | Sequence | Corresponding CDS haplotypes | Sample data

Population frequencies

Population group	Population	Frequency (count)
AFR	ACB	0.0104 (2)
	ASW	0.0164 (2)
AMR	CLM	0.0213 (4)
	MXL	0.00781 (1)
	PEL	0.00588 (1)
	PUR	0.0240 (5)
EAS	CHB	0.00485 (1)
EUR	GBR	0.0989 (18)
	IBS	0.0327 (7)
	CEU	0.126 (25)
	FIN	0.0859 (17)
	TSI	0.0234 (5)
SAS	ITU	0.00490 (1)
	STU	0.0147 (3)
	GIH	0.00485 (1)

```
Protein  p.REF   D  R  Y  I  S  I  F  Y  A  L  R  Y  H  S  I  V  T  L  P  R
         p.ALT   .  .  .  .  .  .  .  .  .  .  C  .  .  .  .  .  .  .  .  .
CDS      c.REF  GACCGCTACATCTCCATCTTCTACGCACTGCGCTACCACAGCATCGTGACCCTGCCGCGG
         c.ALT1 ..............................T.............................

Protein  p.REF   A  R  R  A  V  A  A  I  W  V  A  S  V  V  F  S  T  L  F  I
         p.ALT   .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
CDS      c.REF  GCGCGGCGAGCCGTTGCGGCCATCTGGGTGGCCAGTGTCGTCTTCAGCACGCTCTTCATC
         c.ALT1 ............................................................

Protein  p.REF   A  Y  Y  D  H  V  A  V  L  L  C  L  V  V  F  F  L  A  M  L
         p.ALT   .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
CDS      c.REF  GCCTACTACGACCACGTGGCCGTCCTGCTGTGCCTCGTGGTCTTCTTCCTGGCTATGCTG
         c.ALT1 ............................................................
```

Fig. 8 Representation of protein haplotypes found in 1000 Genomes individuals. For each of the individuals in the 1000 Genomes population, the complete protein and CDS sequences were calculated. Sets of cosegregating variants were defined as protein and transcript haplotypes, their frequencies determined and listed in the Transcript haplotype page. (A) The table lists all the haplotypes found by the amino acid change. Click on the haplotype for more details (shown in panels B and C). By default, the page shows the protein haplotypes, but can be switched to show the CDS haplotypes. (B) The frequency of the selected haplotype across 1000 Genomes subpopulations. (C) An alignment of the reference and haplotype protein and CDS sequences

MC1R, as well as 5 kb upstream and downstream of the gene, filter by "Gene-associated Variant Filters," selecting "Gene stable IDs" and inputting "ENSG00000258839," the stable ID for the *MC1R* gene. You can choose attributes of interest under "Variant" or "Flanking sequences"—for example, the variant name, source, consequence, start and end coordinates, and pathogenicity predictions—which will be listed next to each variant in the output table. Click the "Results" button to view and download the results table (Fig. 9).

3.2 WF2: Gene-Based Searches and Identification of Regulatory Features in a Genomic Region

The following workflow describes a gene-based search and indicates some of the data and annotations collated in the Gene tab and Regulation tab.

The *POU5F1* gene, formerly known as *OCT4*, encodes one of the so-called "Yamanaka factors" implicated in cellular de-differentiation and induction of pluripotency [41, 42]. We can search Ensembl to view the *POU5F1* gene model and associated annotation, including predicted regulatory features.

1. *Getting started*: Type "POU5F1" into the search box on the homepage, www.ensembl.org, or in the upper right corner of any browser page, and click "Go." This will generate a search-results page with "POU5F1 (Human Gene)" as the top hit. Click the title link to navigate directly to the *POU5F1* Gene tab.

   The gene "Summary" containing a graphical representation of the gene model loads by default following navigation from the search results.

2. *Downloading gene sequences*: The sequence of the gene and flanking regions can be downloaded from the Gene tab in two ways.

   (a) To download the sequence in FASTA format for processing in an external tool, simply click the "Export data" button below the left-hand menu.

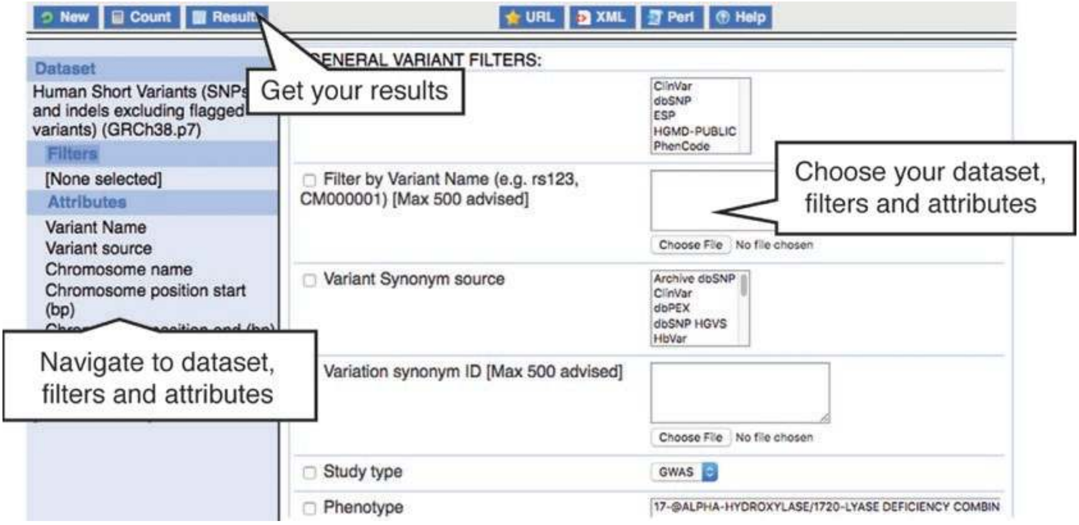
   This will open a pop-up window that presents customization options.

   (b) To view *POU5F1* sequence in the browser, click "Sequence" in the left-hand menu.

   This opens a display in FASTA format; buttons to download and to BLAST the sequence are shown on this page, and download customization options are similarly available (Fig. 10).

3. *Exploring regulatory features*: Select "Summary" in the left-hand menu. Scroll down to the graphical view of the gene model and locate the Regulatory Build track.

   The Regulatory Build depicts regulatory features that have been annotated based on epigenome-scale data imported from sources such as ENCODE [43], *Roadmap Epigenomics* [44] *and*


New
Count
Results
URL
XML
Perl
Help
Get your results
Choose your dataset, filters and attributes
Navigate to dataset, filters and attributes
Choose File
No file chosen

Dataset

Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p7)

Filters

[None selected]

Attributes

Variant Name
Variant source
Chromosome name
Chromosome position start (bp)
Chromosome position end (bp)

GENERAL VARIANT FILTERS:

ClinVar
dbSNP
ESP
HGMD-PUBLIC
PhenCode

□ Filter by Variant Name (e.g. rs123, CM000001) [Max 500 advised]

□ Variant Synonym source

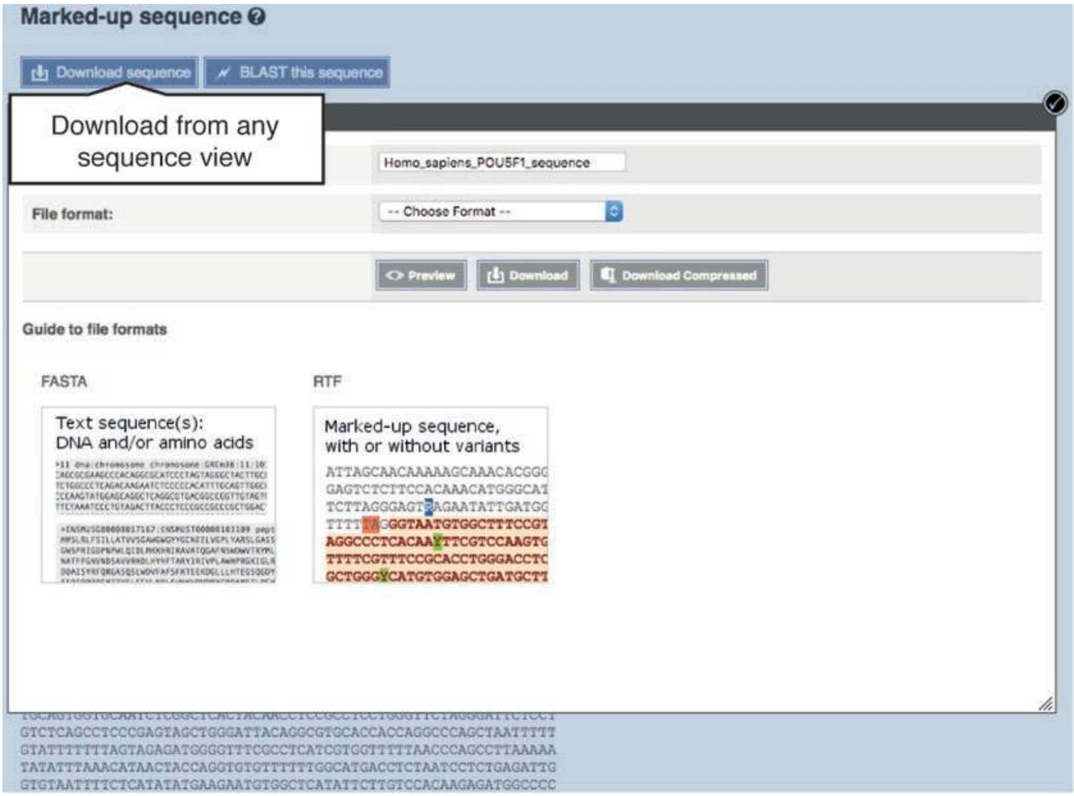
Archive dbSNP
ClinVar
dbPEX
dbSNP HGVS
HbVar

Variation synonym ID [Max 500 advised]

□ Study type GWAS

□ Phenotype 17-@ALPHA-HYDROXYLASE/1720-LYASE DEFICIENCY COMBIN

Fig. 9 The BioMart interface. BioMart allows easy export of tables of gene, variant, or regulatory feature data. A video tutorial for BioMart is available at https://www.youtube.com/watch?v=QvGT2G0-hYA&ab_channel=EnsemblHelpdesk



Marked-up sequence

Homo_sapiens_POU5F1_sequence

File format: -- Choose Format --

Guide to file formats

FASTA

RTF

```
ATTAGCAACAAAAAGCAAACACGGG
GAGTCTCTTCCACAAACATGGGCAT
TCTTAGGGAGTRAGAATATTGATGG
TTTTTAGGGTAATGTGGCTTTCCGT
AGGCCCTCACAAYTTCGTCCAAGTG
TTTTCGTTTCCGCACCTGGGACCTC
GCTGGGYCATGTGGAGCTGATGCTT
```

```
GTCTCAGCCTCCCGAGTAGCTGGGATTACAGGCGTGCACCACCAGGCCCAGCTAATTTTT
GTATTTTTTTAGTAGAGATGGGGTTTCGCCTCATCGTGGTTTTTAACCCAGCCTTAAAAA
TATATTTAAACATAACTACCAGGTGTGTTTTTTGGCATGACCTCTAATCCTCTGAGATTG
GTGTAATTTTCTCATATATGAAGAATGTGGCTCATATTCTTGTCCACAAGAGATGGCCCC
```

Fig. 10 Exporting gene sequence from Ensembl. All sequence views in Ensembl allow download in either plain FASTA or annotated rich text format (RTF)

Blueprint [45]. *These motifs are color-coded according to the predicted function of the element* (Fig. 11A).

4. Click on the red promoter overlapping the 5′ end of the longest transcript of *POU5F1*, *POU5F1-004*, to open a pop-up box with the stable ID ("ENSR00000195510"), type ("Promoter"), and genomic coordinates of the core element and flanking sequences. Click the stable ID to open the Regulation tab.

   Note: POU5F1 *is transcribed from the reverse strand, and thus the 5' sequences containing the promoter are located to the right of the gene.*

   The Regulation tab displays a graphical representation of the genomic region surrounding the element and a table of the 68 cell types with regulation data currently in Ensembl, organized by activity state. In addition to the Regulatory Build, several tracks are shown by default; these include CRISPR/Cas9 genome-editing sites predicted by the Wellcome Trust Sanger Institute (WTSI) [30], *transcription start sites identified by FANTOM5* [46], *miRNA binding sites imported from Tarbase* [31], *and enhancers identified by VISTA* [29]. *Tracks with no data in the immediate region of the feature are not shown* (Fig. 11B) *(The term "track" refers to a data type that can be plotted against the genome.).*

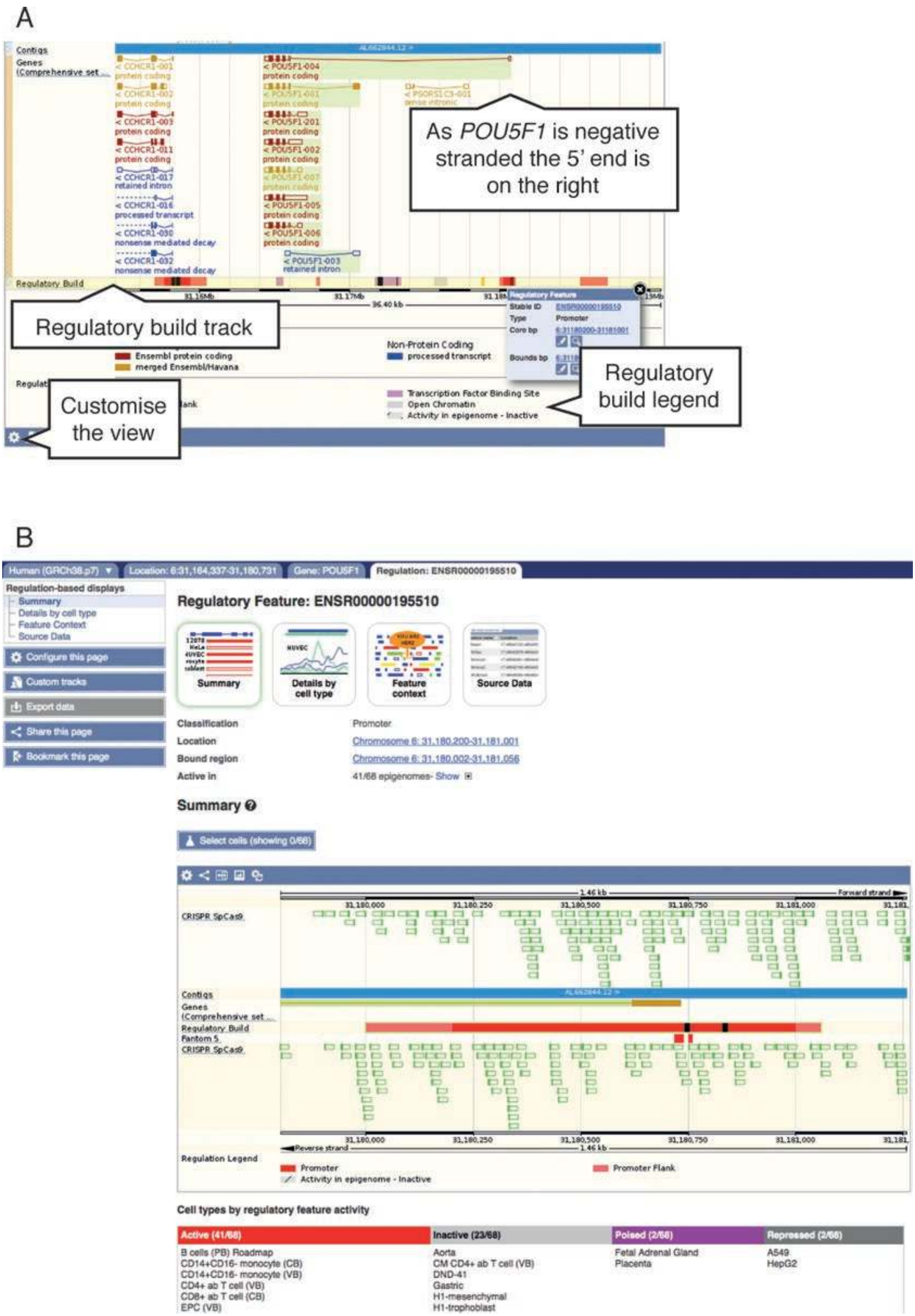
   Feature activity by cell type can be viewed in graphical form by clicking the "Select cells" button and, in the resulting pop-up, choosing "All on" or selecting individual cell types.

5. To view the element's activation state in individual cell types, click the "Details by cell type" button at the top of the Regulation tab or the link in the left-hand menu. Click the "Select cells" button and then choose "A549" (repressed), "Placenta" (poised), "Pancreas" (inactive), "GM12878" (active). Next, click "Select evidence," then "All on," to load the experimental data available for the cell types of interest.

   You are now viewing data from cell types in which the element is active, inactive, poised, and repressed (Fig. 12). *These activation states are determined on the basis of the histone modifications observed in the region, along with transcription factor and RNA polymerase II or III binding, as well as areas of DNase I hypersensitivity indicating open chromatin* [4].

   Additional tracks can be accessed by clicking the "Configure this page" button, at left, or the cogwheel at the top of the image. These include the evidence underlying the Regulatory Build, as well as comparative genomics analyses and variation data that may provide additional context for the annotated feature.

6. Ensure that both "Peaks" and "Signal" buttons are selected.



A
Contigs
Genes (Comprehensive set...
< CCHCR1-001 protein coding
< POU5F1-004 protein coding
< POU5F1-001 protein coding
< PSORS1C3-001 sense intronic
< POU5F1-201 protein coding
< POU5F1-002 protein coding
< POU5F1-005 protein coding
< POU5F1-006 protein coding
< POU5F1-003 retained intron
< CCHCR1-017 retained intron
< CCHCR1-016 processed transcript
< CCHCR1-030 nonsense mediated decay
< CCHCR1-032 nonsense mediated decay
Regulatory Build
31.16Mb
31.17Mb
36.40 kb
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript
Transcription Factor Binding Site
Open Chromatin
Activity in epigenome - Inactive
As *POU5F1* is negative stranded the 5′ end is on the right
Regulatory build track
Customise the view
Regulatory build legend
B
Human (GRCh38.p7)
Location: 6:31,164,337-31,180,731
Gene: POU5F1
Regulation: ENSR00000195510
Regulation-based displays
Summary
Details by cell type
Feature Context
Source Data
Configure this page
Custom tracks
Export data
Share this page
Bookmark this page
Summary
Details by cell type
Feature context
Source Data
Select cells (showing 0/68)
CRISPR SpCas9
Contigs
Genes (Comprehensive set...
Regulatory Build
Fantom 5
1.46 kb
Forward strand
Reverse strand
31,180,000
31,180,250
31,180,500
31,180,750
31,181,000
Regulation Legend
Promoter
Promoter Flank
Activity in epigenome - Inactive

Regulatory Feature

Stable ID	ENSR00000195510
Type	Promoter
Core bp	6:31180200-31181001
Bounds bp	6:3118

Regulatory Feature: ENSR00000195510

Classification	Promoter
Location	Chromosome 6: 31,180,200-31,181,001
Bound region	Chromosome 6: 31,180,002-31,181,056
Active in	41/68 epigenomes- Show

Summary

Cell types by regulatory feature activity

Active (41/68)	Inactive (23/68)	Poised (2/68)	Repressed (2/68)
B cells (PB) Roadmap	Aorta	Fetal Adrenal Gland	A549
CD14+CD16- monocyte (CB)	CM CD4+ ab T cell (VB)	Placenta	HepG2
CD14+CD16- monocyte (VB)	DND-41		
CD4+ ab T cell (VB)	Gastric		
CD8+ ab T cell (CB)	H1-mesenchymal		
EPC (VB)	H1-trophoblast		

Fig. 11 The Ensembl regulatory build and regulatory features. (A) The regulatory build is shown as a track on the gene image. Clicking on a feature in the track opens a pop-up menu, with a link to the regulatory feature tab. Some transcripts have been removed from this image for size. (B) The summary page of the regulatory feature tab contains a table listing activity in different cell types. The graphic shows the feature in context, along with genes, CRISPR-Cas9 sites and FANTOM5 annotation

This will display a summary of the aligned reads (signal) as well as the peaks for each assay. Annotated features are clickable; for example, clicking on a predicted promoter will indicate any transcription factors known to bind it, along with links to the JASPAR database [47], *where further information on motifs is presented. For other elements, the position of the apex is indicated with black arrowheads* (Fig. 12).

7. To view regulatory features across a larger genomic region, navigate to the Location tab, available to the left of the Gene tab in the navigation bar.

   The Location tab displays three images: a global view of the chromosome of interest, an intermediate-scale view providing an overview of the region flanking the relevant genomic locus (in this case that of POU5F1*), and a final view that presents gene-annotation, comparative genomics and variation tracks by default, along with the Regulatory Build.*

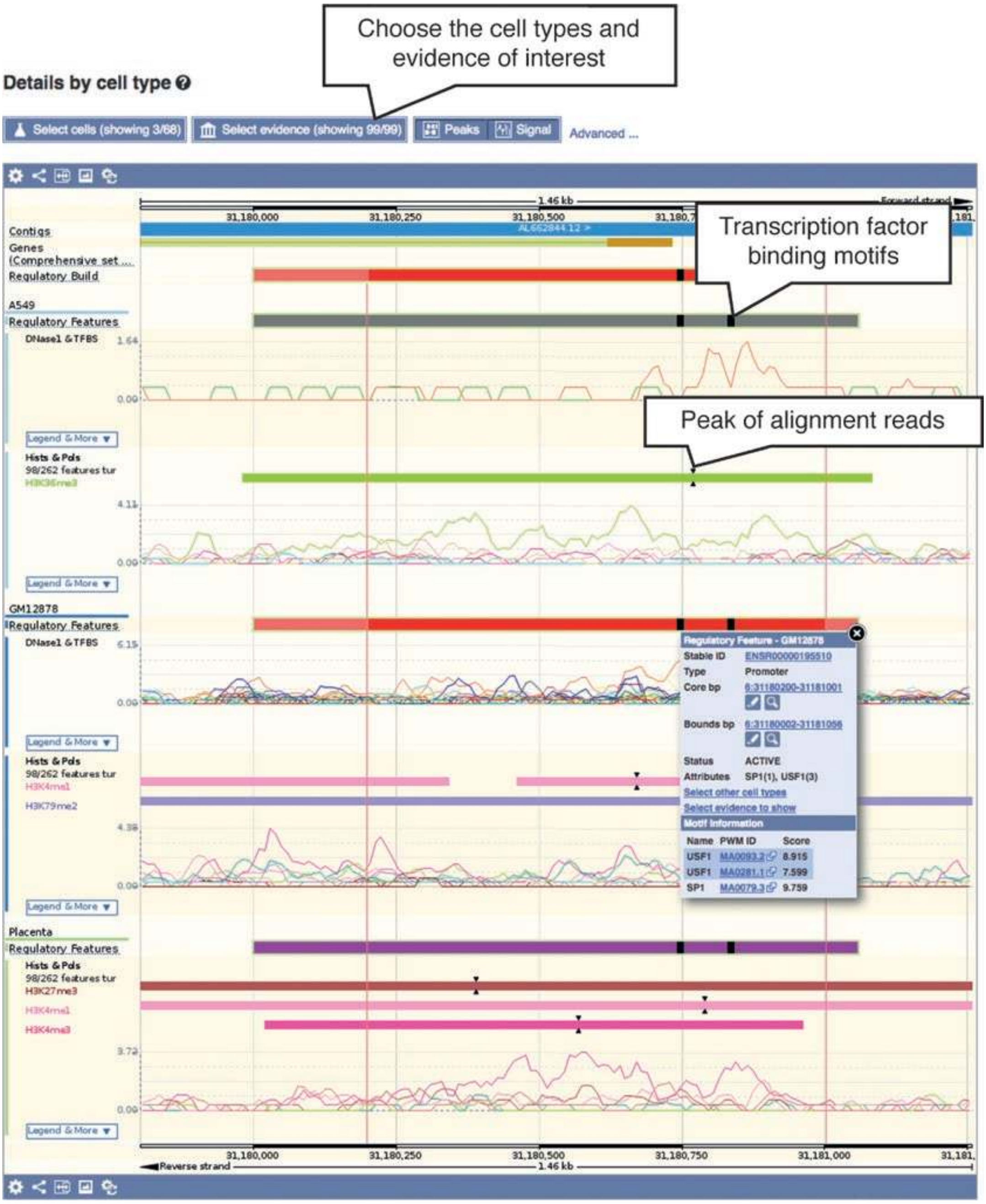
   It is possible to configure the page to view the activity of local regulatory features by cell type, along with the evidence underlying these determinations. As in the Regulation tab, tracks depicting other Ensembl annotations can be added to provide context to the elements shown.

8. Click on the blue "Configure this page" button to add regulatory data tracks for the same cell types: A549, placenta, pancreas, and GM12878.

   This opens a menu listing the many possible tracks available to display on the genome. Categories of tracks are listed on the left. Tracks can be turned on and off by clicking on the box alongside them. To see the activity of regulatory features in different cell types, turn them on within the "Regulatory features" section. In the "Histones & polymerases" and "Open chromatin & TFBS" sections, you will find that tracks are displayed as a matrix, with cell types along the top and evidence to the side (Fig. 13).

9. *Exporting regulatory features with BioMart*: A list of regulatory features, by type, in a genomic region can also be exported via BioMart. Navigate to BioMart, then select "Ensembl Regulation" > "Human Regulatory Features" (it may be necessary to refresh the window by clicking "New" if you have performed a previous query). For features within 5 kb up- and downstream of *POU5F1*, filter for Chromosome 6, Base pair start: 31159337, Base pair end: 31185731. As defaults, "Chromosome Name," "Start (bp)," "End (bp)," and "Feature Type" are selected as Attributes. Add "Regulatory Stable ID" and generate your results.

   Nine features are returned for this genomic region, including the promoter we explored, ENSR00000195510.



Choose the cell types and evidence of interest

Transcription factor binding motifs

Peak of alignment reads

Regulatory Feature - GM12878

Stable ID ENSR00000195510

Type Promoter

Core bp 6:31180200-31181001

Bounds bp 6:31180002-31181056

Status ACTIVE

Attributes SP1(1), USF1(3)

Select other cell types

Select evidence to show

Motif Information

Name	PWM ID	Score
USF1	MA0093.2	8.915
USF1	MA0281.1	7.599
SP1	MA0079.3	9.759

Fig. 12 Evidence for and activity of regulatory features in different cell types. The Details by cell type page in the regulatory feature tab can be manipulated to show the activity of the feature in cell types of interest using the buttons at the top. For each cell type, the feature is shown colour-coded to indicate its activity, with the evidence shown below. The evidence is ChIP-seq and DNase-seq data, and is shown as peaks of significant activity and as signal giving the number of reads. The top of the peak is indicated in the peak bar by pairs of black arrows. Black blocks in the regulatory features indicate the position of transcription factor binding motifs, which are listed in a pop-up when clicked on

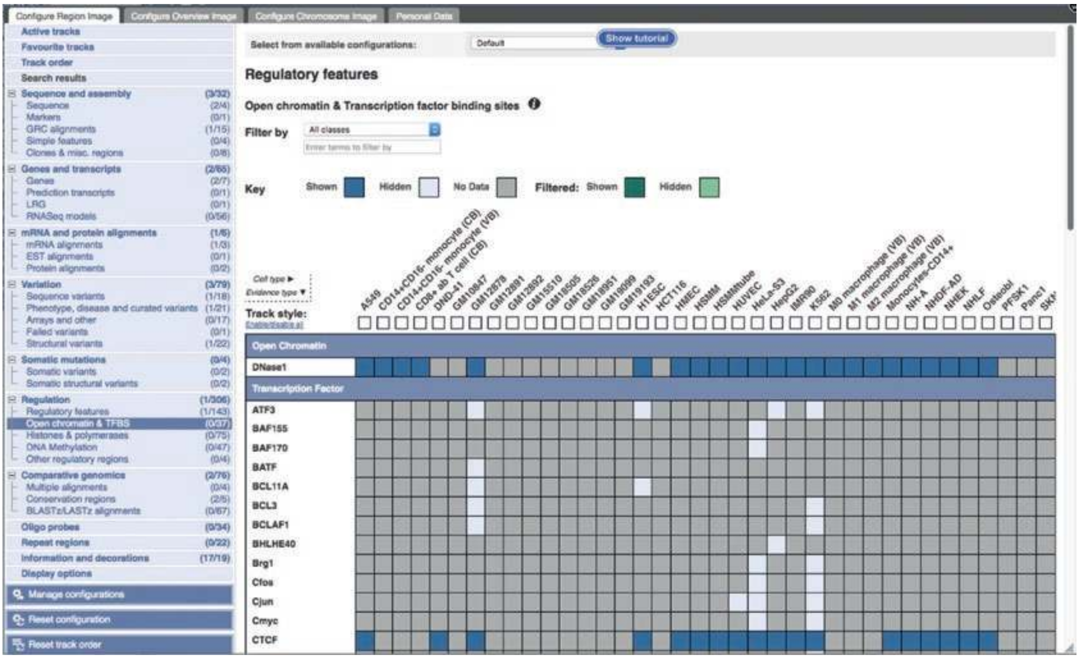

Configure Region Image
Configure Overview Image
Configure Chromosome Image
Personal Data
Active tracks
Favourite tracks
Track order
Search results
Sequence and assembly (3/32)
Sequence (2/4)
Markers (0/1)
GRC alignments (1/15)
Simple features (0/4)
Clones & misc. regions (0/8)
Genes and transcripts (2/65)
Genes (2/7)
Prediction transcripts (0/1)
LRG (0/1)
RNASeq models (0/56)
mRNA and protein alignments (1/6)
mRNA alignments (1/3)
EST alignments (0/1)
Protein alignments (0/2)
Variation (3/79)
Sequence variants (1/18)
Phenotype, disease and curated variants (1/21)
Arrays and other (0/17)
Failed variants (0/1)
Structural variants (1/22)
Somatic mutations (0/4)
Somatic variants (0/2)
Somatic structural variants (0/2)
Regulation (1/306)
Regulatory features (1/143)
Open chromatin & TFBS (0/37)
Histones & polymerases (0/75)
DNA Methylation (0/47)
Other regulatory regions (0/4)
Comparative genomics (2/76)
Multiple alignments (0/4)
Conservation regions (2/5)
BLASTz/LASTz alignments (0/67)
Oligo probes (0/34)
Repeat regions (0/22)
Information and decorations (17/19)
Display options
Manage configurations
Reset configuration
Reset track order
Select from available configurations: Default
Show tutorial
Regulatory features
Open chromatin & Transcription factor binding sites
Filter by All classes
Enter terms to filter by
Key Shown Hidden No Data Filtered: Shown Hidden
Cell type ▸
Evidence type ▾
Track style:
Enable/disable all
A549 CD14+CD16- monocyte (CB) CD14+CD16- monocyte (VB) CD8+ ab T cell (CB) DND-41 GM10847 GM12878 GM12891 GM12892 GM15510 GM18505 GM18526 GM18951 GM19099 GM19193 H1ESC HCT116 HMEC HSMM HSMMtube HUVEC HeLa-S3 HepG2 IMR90 K562 M0 macrophage (VB) M1 macrophage (VB) M2 macrophage (VB) Monocytes-CD14+ NH-A NHDF-AD NHEK NHLF Osteobl PFSK1 Panc1 SK
Open Chromatin
DNase1
Transcription Factor
ATF3
BAF155
BAF170
BATF
BCL11A
BCL3
BCLAF1
BHLHE40
Brg1
Cfos
Cjun
Cmyc
CTCF

Fig. 13 Adding regulation tracks to a region view. The Region in detail view displays a genomic region and can be customized to show tracks of interest using the Configure this page button. This opens a detailed menu listing all the available tracks, using categories on the left, including regulatory features and evidence. Regulatory evidence can be added using a matrix selector, listing the cell type along the top and type of evidence down the left

4 Discussion

Here, we describe methods to navigate variation and regulation data in the Ensembl browser, focusing on human, although the principle of navigation is relevant to queries in all species.

The typical entry point to a query in the browser is the search function. The Ensembl search is versatile and can retrieve information linked to a variety of inputs—including, but not limited to, genomic locations; gene, transcript, protein and regulatory feature IDs; GO terms; variant IDs; and phenotypes. Unless otherwise specified in the query, search results for human will be returned first; filters displayed on the left-hand side of the results page permit the restriction of results by category (e.g., gene, variant) and by organism.

Selecting a search result will open a tab that collates information on the entity: in the two workflows presented above, we present strategies for accessing the Phenotype, Variant, Gene, Transcript, Location, and Regulation tabs following phenotype- and gene-based searches. As you move from tab to tab in a single query, previously accessed tabs will remain open in the blue navigation bar at the top of the page to facilitate seamless data-retrieval in a minimal number of steps; you can reenter a previous tab simply by clicking on the tab header in the navigation bar (Fig. 6A).

By default, tabs will open with a summary of the information available for each entity (e.g., a transcript or variant), although herein we indicate a few cases where other data are loaded: for example, the Gene Variant table is presented immediately upon navigation from the Variant to the Gene tab. Should a view not be as expected, links to all data and annotations available in a tab can be found in the menu on the left; for Location, Gene and Transcript tabs, these links adhere to a similar framework but present annotations at different scales.

Tabs can be customized by clicking on the blue "Configure this page" button below the left-hand menu, or the cogwheel icons that appear in the upper borders of graphical displays (Fig. 13). Customization allows you to add or remove data tracks that may be useful to interpretation or analysis; for example, to view the evidence underlying an activation-state prediction for a regulatory feature in a cell type of interest (in the Location or Regulation tabs). Other examples of customization include, in the Location tab, the addition of tracks containing the data, imported from external repositories, that were used to annotate transcripts in a genomic region (ENA, UniProt, and RefSeq tracks accessible in Location tab) [11]. Public datasets can also be added from the Track Hub Registry [48], and you can import your own data, in multiple formats [49], for examination in the context of the browser.

Data can be exported directly from the browser by clicking the blue "Export data" buttons found below the left-hand menu in most tabs, or the "Download sequence" buttons above FASTA sequences in the Gene and Transcript tabs. In addition, the BioMart tool described in the workflows presented herein can be used to retrieve custom datasets from our Gene, Variation and Regulation databases, and data can be accessed programmatically from our Perl APIs and REST service. Data from all Ensembl releases can also be downloaded en masse from our FTP site.

Species' sequence data and annotations may be updated several times a year. You should therefore be attentive, when querying Ensembl data, to the current browser version, as annotations are subject to change. Data can, however, still be retrieved directly from archived versions of the browser, as well as via BioMart, while the browser web interface remains online. Following the decommissioning of any browser version, the data remain accessible from our FTP site and APIs, as mentioned above.

A dedicated email helpdesk is available to field any inquiries about Ensembl and we typically reply to messages within two days of receipt. We also hold training workshops upon invitation by research institutes; from 2013 to 2016 we participated in an average of 86 workshops, and trained 2150 students, per year. Our training materials are accessible online [50], along with a number of courses that are available on the Train Online Platform of the European Bioinformatics Institute (EMBL-EBI) [51], and we have published a blogpost outlining the process of hosting your own workshop [52]. Short help videos can be found both on our YouTube channel [53] and, for those who cannot access YouTube, on Youku [54]. We invite the community to contact us via helpdesk@ensembl.org for more information about workshops, with questions regarding the browser, and to suggest features and resources which would assist their work.

Funding/Acknowledgments

Ensembl receives majority funding from the Wellcome Trust (grant number WT108749/Z/15/Z) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007234 and U41HG007823), the Biotechnology and Biological Sciences Research Council (BB/L024225/1, BB/M011615/1, and BB/M020398/1), Open Targets, the Wellcome Trust (WT104947/Z/14/Z, WT200990/Z/16/Z, and WT201535/Z/16/Z) and the European Molecular Biology Laboratory. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 733161 (MultipleMS).

References

1. Aken BL, Ayling S, Barrell D et al (2016) The Ensembl gene annotation system. Database (Oxford) 2016. https://doi.org/10.1093/database/baw093
2. Chen Y, Cunningham F, Rios D et al (2010) Ensembl variation resources. BMC Genomics 11:293. https://doi.org/10.1186/1471-2164-11-293
3. Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. Database (Oxford) 2016. https://doi.org/10.1093/database/baw053
4. Zerbino DR, Johnson N, Juetteman T et al (2016) Ensembl regulation resources. Database (Oxford) 2016. https://doi.org/10.1093/database/bav119
5. Kent WJ, Sugnet CW, Furey TS et al (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006. https://doi.org/10.1101/gr.229102. Article published online before print in May 2002
6. Robinson JT, Thorvaldsdottir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26. https://doi.org/10.1038/nbt.1754
7. Hubbard T, Barker D, Birney E et al (2002) The Ensembl genome database project. Nucleic Acids Res 30(1):38–41
8. The Ensembl Browser. http://www.ensembl.org
9. Kersey PJ, Allen JE, Armean I et al (2016) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 44(D1):D574–D580. https://doi.org/10.1093/nar/gkv1209
10. The Ensembl Genomes Browser. http://www.ensemblgenomes.org
11. Aken BL, Achuthan P, Akanni W et al (2017) Ensembl 2017. Nucleic Acids Res 45(D1):D635–D642. https://doi.org/10.1093/nar/gkw1104

12. Kinsella RJ, Kahari A, Haider S et al (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030. https://doi.org/10.1093/database/bar030
13. Ruffier M, Kahari A, Komorowska M et al (2017) Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. Database (Oxford) 2017(1). https://doi.org/10.1093/database/bax020
14. Yates A, Beal K, Keenan S et al (2015) The Ensembl REST API: Ensembl data for any language. Bioinformatics 31(1):143–145. https://doi.org/10.1093/bioinformatics/btu613
15. The Ensembl FTP site. ftp://ftp.ensembl.org
16. McLaren W, Gil L, Hunt SE et al (2016) The Ensembl variant effect predictor. Genome Biol 17(1):122. https://doi.org/10.1186/s13059-016-0974-4
17. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12(4):656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002
18. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
19. Clarke L, Zheng-Bradley X, Smith R et al (2012) The 1000 genomes project: data management and community access. Nat Methods 9(5):459–462. https://doi.org/10.1038/nmeth.1974
20. Ensembl Publications. http://www.ensembl.org/info/about/publications.html
21. Zhang M, Song F, Liang L et al (2013) Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum Mol Genet 22(14):2948–2959. https://doi.org/10.1093/hmg/ddt142
22. The GWAS catalog. https://www.ebi.ac.uk/gwas/
23. Europe PMC. https://europepmc.org/
24. Sources of Ensembl variation data. http://www.ensembl.org/info/genome/variation/sources_documentation.html
25. Eilbeck K, Lewis SE, Mungall CJ et al (2005) The sequence ontology: a tool for the unification of genome annotations. Genome Biol 6(5):R44. https://doi.org/10.1186/gb-2005-6-5-r44
26. Genomes Project Consortium, Auton A, Brooks LD et al (2015) A global reference for human genetic variation. Nature 526(7571):68–74. https://doi.org/10.1038/nature15393
27. Goldstein DB, Cavalleri GL (2005) Genomics: understanding human diversity. Nature 437(7063):1241–1242. https://doi.org/10.1038/4371241a
28. Exome Variant Server. NHLBI GO Exome Sequencing Project (ESP). http://evs.gs.washington.edu/EVS/
29. Visel A, Minovitsky S, Dubchak I et al (2007) VISTA enhancer browser–a database of tissue-specific human enhancers. Nucleic Acids Res 35(Database issue):D88–D92. https://doi.org/10.1093/nar/gkl822
30. Hodgkins A, Farne A, Perera S et al (2015) WGE: a CRISPR database for genome engineering. Bioinformatics 31(18):3078–3080. https://doi.org/10.1093/bioinformatics/btv308
31. Vlachos IS, Paraskevopoulou MD, Karagkouni D et al (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res 43(Database issue):D153–D159. https://doi.org/10.1093/nar/gku1215
32. Gene Ontology Consortium (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056. https://doi.org/10.1093/nar/gku1179
33. O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189
34. HGNC database of human gene names. http://www.genenames.org/
35. Online Mendelian Inheritance in Man. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). https://www.omim.org/
36. Pruitt KD, Harrow J, Harte RA et al (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19(7):1316–1323. https://doi.org/10.1101/gr.080531.108
37. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169. https://doi.org/10.1093/nar/gkw1099
38. Toribio AL, Alako B, Amid C et al (2017) European nucleotide archive in 2016. Nucleic Acids Res 45(D1):D32–D36. https://doi.org/10.1093/nar/gkw1106

39. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
40. Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. https://doi.org/10.1038/nmeth0410-248
41. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126(4):663–676. https://doi.org/10.1016/j.cell.2006.07.024
42. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. Nature 448(7151):313–317. https://doi.org/10.1038/nature05934
43. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. https://doi.org/10.1038/nature11247
44. Roadmap epigenomics Consortium, Kundaje A, Meuleman W et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317–330. https://doi.org/10.1038/nature14248
45. Fernandez JM, de la Torre V, Richardson D et al (2016) The BLUEPRINT data analysis portal. Cell Syst 3(5):491–495.e495. https://doi.org/10.1016/j.cels.2016.10.021
46. Fantom Consortium, Forrest AR, Kawaji H et al (2014) A promoter-level mammalian expression atlas. Nature 507(7493):462–470. https://doi.org/10.1038/nature13182
47. Bryne JC, Valen E, Tang MH et al (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36(Database issue):D102–D106. https://doi.org/10.1093/nar/gkm955
48. The Track Hub Registry. https://trackhubregistry.org
49. Data formats compatible with Ensembl. http://www.ensembl.org/info/website/upload/index.html - formats
50. The Ensembl Training Site. http://training.ensembl.org
51. EMBL-EBI's Train Online Platform. https://www.ebi.ac.uk/training/online/
52. Hosting an Ensembl Workshop. http://www.ensembl.info/blog/2017/01/05/so-you-want-to-run-an-ensembl-workshop/
53. The Ensembl Helpdesk YouTube channel. https://www.youtube.com/user/EnsemblHelpdesk
54. The Ensembl Helpdesk Youku channel. http://i.youku.com/i/UMzM1NjkzMTI0?spm=a2h0j.8191423.subscription_wrap.DD~A