D710–D716 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 19 December 2015*
doi: 10.1093/nar/gkv1157

Ensembl 2016

Andrew Yates[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Konstantinos Billis[1], Denise Carvalho-Silva[1], Carla Cummins[1], Peter Clapham[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[1], Leo Gordon[1], Thibaut Hourlier[1], Sarah E. Hunt[1], Sophie H. Janacek[1], Nathan Johnson[1], Thomas Juettemann[1], Stephen Keenan[1], Ilias Lavidas[1], Fergal J. Martin[1], Thomas Maurel[1], William McLaren[1], Daniel N. Murphy[1], Rishi Nag[1], Michael Nuhn[1], Anne Parker[1], Mateus Patricio[1], Miguel Pignatelli[1], Matthew Rahtz[2], Harpreet Singh Riat[1], Daniel Sheppard[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Steven P. Wilder[1], Amonida Zadissa[1], Ewan Birney[1], Jennifer Harrow[2], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Giulietta Spudich[1], Stephen J. Trevanion[1], Fiona Cunningham[1], Bronwen L. Aken[1], Daniel R. Zerbino[1] and Paul Flicek[1,2,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Received September 19, 2015; Revised October 19, 2015; Accepted October 19, 2015

ABSTRACT

The Ensembl project (http://www.ensembl.org) is a system for genome annotation, analysis, storage and dissemination designed to facilitate the access of genomic annotation from chordates and key model organisms. It provides access to data from 87 species across our main and early access Pre! websites. This year we introduced three newly annotated species and released numerous updates across our supported species with a concentration on data for the latest genome assemblies of human, mouse, zebrafish and rat. We also provided two data updates for the previous human assembly, GRCh37, through a dedicated website (http://grch37.ensembl.org). Our tools, in particular the VEP, have been improved significantly through integration of additional third party data. REST is now capable of larger-scale analysis and our regulatory data BioMart can deliver faster results. The website is now capable of displaying long-range interactions such as those found in *cis*-regulated datasets. Finally we have launched a website optimized for mobile devices providing views of genes, variants and phenotypes. Our data is made available without restriction and all code is available from our GitHub organization site (http://github.com/Ensembl) under an Apache 2.0 license.

INTRODUCTION

Ensembl (http://www.ensembl.org) generates genomic datasets through a system that is designed to analyse, store and distribute data, and which enables interpretation through open data release. While acting as a hub of reference and baseline data similar to the UCSC Genome Browser (1) and RefSeq (2), we also distribute datasets we create and promote standards and interoperability between genomic resources. We engage with the scientific community through an active outreach program and helpdesk. In addition we collaborate with and often play active leadership roles in projects such as ENCODE (3), the Genome Reference Consortium (GRC) (4), the Global Alliance for Genomics and Health (GA4GH) and GENCODE (5). Ensembl is updated four to five times per year with each release representing a data and software freeze. This procedure ensures that all our data are consistent, no matter the method of access. Every release is accompanied by archived versions of our website and BioMart data mining tool with a three year rolling retention policy. All public data releases regardless of age are available from our FTP site, MySQL servers and public Git repositories. In addition a REST API provides program language agnostic access to the current data release.

Our analysis methods construct annotation through the processing and summarization of experimental evidence. Gene annotation relies on the alignment of cDNAs and proteins from resources such as RefSeq and UniProt (6) alongside building transcription models from RNA-seq align-

*To whom correspondence should be addressed. Tel: +44 1223 492581; Fax: +44 1223 494494; Email: flicek@ebi.ac.uk

ment data. Our Regulatory Build is based on high quality experimental evidence from projects such as ENCODE and Roadmap Epigenomics (7) and is capable of annotating a diverse set of features across many distinct cell types (8). All gene and regulatory annotation is accessioned and versioned between releases enabling downstream analysis to accurately refer back to these annotations. We also produce comparative genomics resources, which build on top of this gene annotation to calculate gene evolution and orthology information and use genomic DNA to build whole genome pairwise and multiple sequence alignments. Finally our variation resources integrate disparate data sources (including dbSNP (9), HGMD (10), ClinVar (11)) and present them through a consistent integrated interface. Variant consequences are calculated with reference to our gene and regulatory annotation and quantified by standard protein consequence analyses.

Annotation is made available through a set of mature Perl Application Programming Interfaces (APIs), which broker data from our databases. These same APIs are used to build our website and analysis methods and they are available externally for others to use in building their own methods and tools. Our infrastructure, whilst originally developed to operate on chordates and core model organisms, has been successfully deployed for a wide range of taxa as shown by our sister project Ensembl Genomes (12). Tools such as the Ensembl Variant Effect Predictor (VEP) can be applied to any genome due to our common programming interface and support for standard data formats (13). All Ensembl software is available under an Apache 2.0 license and is free for all to use.

Release 82 (September 2015) makes 69 species available from our main website, 18 species from our pre-release website (http://pre.ensembl.org) and GRCh37 annotation served from a dedicated website (http://grch37.ensembl.org). All chordates have been analysed by our gene annotation methods. Variation data is available for 22 species and regulatory data is available for human and mouse. All three websites provide sequence search. Additional tools are available on our main and GRCh37 websites, including the VEP and an assembly coordinate conversion tool based on CrossMap (14). The main website is hosted from four locations based in the UK, Singapore, US East and West coasts; the final three are deployed on Amazon Web Services. We also provide the ability to attach standard bioinformatic data formats including BigBED, BigWig (15), VCF (16) and BAM (17) to visualize external data in the context of our own data and support the UCSC track hub format to orchestrate track configuration (18).

This report focuses on new data and important technological changes to the project. We explain how these improvements enhance Ensembl and aid the analysis and interpretation of genomic data.

GENOME ANNOTATION

Protein coding and non-protein coding gene annotation

Over the past year, we have concentrated on supporting our most accessed species and annotating a selection of new genomes. As a member of the GENCODE consortium, we have followed its recent decision to improve mouse gene annotation to a similar level to human and have adopted a new gene annotation release cycle. Computationally annotated mouse gene annotation is currently merged every release with manual annotation from the HAVANA project (19). The human genome receives an update every other release with zebrafish or rat receiving gene annotation updates in those releases when human does not. We have incorporated several minor assembly updates (three for human and two for mouse) including GRCh38.p3 and GRCm38.p4. Both were released in Ensembl 81 (July 2015). The human and mouse gene annotations are supplemented by three methods to help quantify transcript support and provide subsets of the GENCODE dataset. Transcript support levels (TSL) are an expression of how well mRNA and EST libraries align to transcripts across splice junctions. Transcripts are assigned a numeric value from one to five indicating the level of support. APPRIS is used to identify principle transcript isoforms of genes from proteomic datasets (20). Finally the GENCODE Basic representative transcript set prioritizes full-length protein coding transcripts over partial or non-protein coding transcripts and is based on rules agreed by the GENCODE consortium (21).

In addition, we have annotated two major assembly updates for rat (Rnor_6.0) and zebrafish (GRCz10). Both gene sets include manual annotation from the HAVANA project. We have also recently updated our lincRNA annotation methods to using candidate transcript models built from RNA-seq data. These are tested for protein coding potential by searching for Pfam domains and alignments to the UniProt database. Models that show no protein coding potential are labelled as lincRNA. We have applied this method to generate lincRNAs for rat and sheep and aim to extend the method to other species over the coming year. We have also produced preliminary transcript models for Crab-eating macaque (*Macaca fascicularis*) and sperm whale (*Physeter macrocephalus*) by aligning experimental data and homologous proteins; these are available on our Pre! website.

To compare our annotation to external gene sets we have imported both Consensus CDS (CCDS) and RefSeq transcripts for human, mouse and selected other species, into our infrastructure (22). We annotate RefSeq transcripts when their mRNA sequence does not exactly match the underlying genomic sequence and provide details on where the mismatch or insertion-deletion occurs e.g. 5′ UTR, CDS, 3′ UTR. RefSeq transcripts that exactly match an overlapping Ensembl-generated model, with respect to the entire model or just coding exons, are annotated accordingly. All annotation is available to downstream analysis tools including the VEP and via our unified APIs.

Variation annotation

Our variation resources integrate essentially all publicly available germline and somatic variant data for 20 vertebrate species. Over the past year the number of SNPs, indels and structural variants in the databases has almost doubled to 468 million variants. We have seen a dramatic increase in genotypes for human (206 000 million), cow (160 million) and sheep (6300 million). This led us to develop a new VCF-

based genotype layer to reduce storage, processing and access time for these data. In addition, we have redesigned the variation database schema to model individuals with multiple samples.

Alongside variation data we also bring in phenotype, trait and disease annotations for 14 species totalling 2.8 million annotations of genes, short variants, structural variants and QTLs. For human these data span over 15 000 phenotypes, traits and diseases from 17 sources including ClinVar, OMIM (Online Mendelian Inheritance in Man) (23) and the NHGRI-EBI GWAS Catalog (24). Eight sources are incorporated for other species including RGD (25), OMIA (26), AnimalQTL (27), ZFIN (28).

In addition to the above work we supplement our data with available citations, pass variants though our quality-control procedures and predict the consequence of variants on our gene sets and regulatory regions. For every possible amino acid change in our 10 most popular species, we run SIFT with enhanced quality information (29) and PolyPhen-2 (30) (human only). We also compute Human Genome Variation Society (HGVS) nomenclature for every variant and recently moved to 3′ shifting of indels to conform to the HGVS specification.

Regulatory annotation

Ensembl Regulation annotation describes the functional role of non-genic genomic elements, in particular enhancers, promoters and insulators, using biochemical assays such as ChIP-Seq or DNAse1 hypersensitivity. The re-designed method, which was deployed last year on 18 human cell lines, has been extended to mouse and covers 8 cell lines and tissues (8).

In anticipation of high-level *cis*-regulatory datasets, which will link regulatory elements to neighbouring genes, we can now render interaction data on our graphical genome location view. Interaction elements are described by the existing WashU Epigenome Browser formats and then loaded onto our website via user upload or by specifying a HTTP URL (31). The data are then visualized as arcs spanning across the region in question, as illustrated in Figure 1.

Comparative annotation

Ensembl's comparative analysis integrates the genome sequences and gene annotations of all available species into a single comprehensive resource. We have updated our whole-genome alignments due to updates to both rat and zebrafish assemblies. The zebrafish assembly update resulted in re-computing 20 pairwise whole genome alignments and our fish Enredo Pecan Ortheus (EPO) multiple alignments (32). We have also retired our fish-specific EPO method resulting in a single EPO production pipeline applicable to fish, mammal and sauropsid multiple sequence alignments.

Major development work is on going to move towards a new protein clustering and classification system. Our current method is based on clustering blastp distances using hcluster_sg (33). It will be replaced with a more straightforward HMM classification based upon PANTHER (34). Moving to an HMM classification will enable analysis and clustering of proteins in linear time compared to our previous approach. We are also developing two methods of gene tree reconstruction. The first constructs gene trees *de novo* whilst the second enables gene tree modification by removing genes or inserting new genes into the tree. Our new methods are being benchmarked via the Quest for Orthologues service (http://orthology.benchmarkservice.org), which tests a range of metrics including tree-consistency approaches, gene ontology and enzyme classification tests (35). These developments will address the increasing numbers of species available for comparative analysis in Ensembl and Ensembl Genomes and improve the stability of the predicted gene-trees and orthologies.

GRCh37 human assembly support

Our GRCh37 website, supporting the previous human assembly, has received two major releases over the past year. Our first (March 2015) included new data variant from the 1000 Genomes Project phase 3 (36), dbSNP, COSMIC v71 (37) and HGMD. The second update (October 2015) incorporated variants from the Exome Aggregation Consortium (ExAC) (38) and NHLBI Exome Sequencing Project (39). A full selection of resources is available for GRCh37 including BioMart and public MySQL access. Our FTP site (ftp://ftp.ensembl.org/pub/grch37) provides VEP cache and FASTA updates. We also maintain a GRCh37 REST API hosted at http://grch37.rest.ensembl.org.

WEBSITE, TOOLS AND INFRASTRUCTURE

VEP

This year we have improved the VEP's ability to report transcript attributes such as a transcript's existence in the GENCODE Basic set and a transcript's TSL support level (both subject to availability). We also record, in the VEP's output, if phenotype/disease data are available. Additionally, it is possible to report predictions on RefSeq transcripts and the VEP will now indicate if the transcript matches a model annotated by Ensembl. Finally the VEP now supports selenocysteine modifications.

The standalone VEP has been enhanced by several new plugins. It is now possible to retrieve ExAC allele frequencies from downloaded VCFs, to query for splice site predictions from dbscSNV (40) and finally to retrieve gene expression levels from Gene Expression Atlas data via their web service API (41). Other plugins allow the VEP to locate the nearest gene to a variant and indicate if a variant has been shifted in HGVS notation. Our online tool interface has also been updated to provide an immediate overview of a single variant's consequences. The 'Instant VEP' tool queries our live REST API to return consequence data in less than a second.

Extensive development work on the VEP has resulted in significant reductions in runtime. For example, analysis of NA12878 from Illumina's Platinum Genome dataset (annotating 4 498 138 variants) using the GRCh38 assembly and release 81 data took 113 min to complete using four compute cores (42). The same analysis performed using release 77 data and analysis, October 2014, took 199 min to complete. To improve the installation procedure we now quality

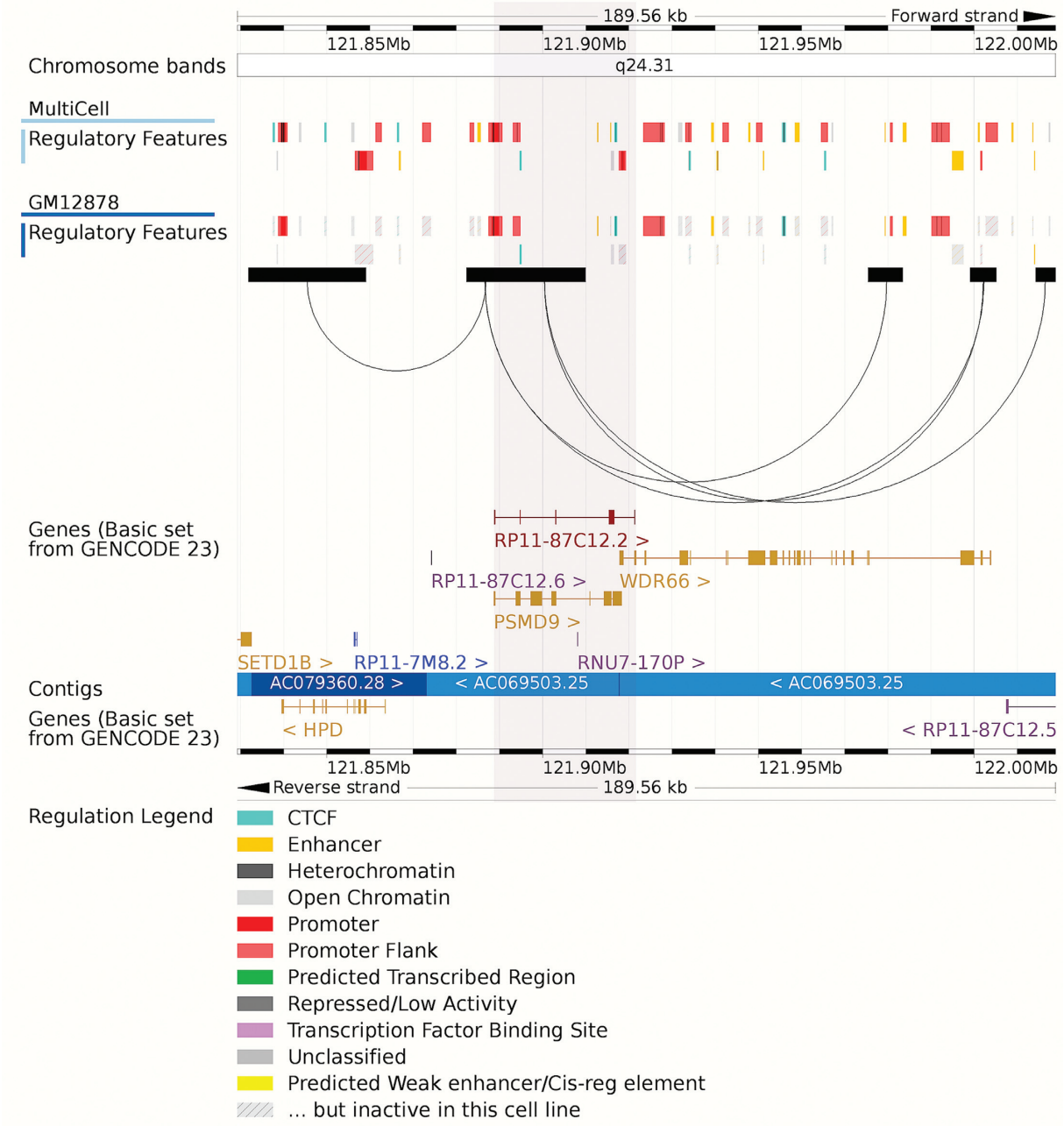

Figure 1. Ensembl's location view showing the drawing of long-range interaction arcs and new region marking tool. The grey boxes indicate HindIII fragments and the arcs represent selected significant interactions between promoters and their distal interacting elements, measured using high resolution Capture Hi-C in the GM12878 lymphoblastoid cell line and displayed for the GRCh38 assembly (43). The summary Ensembl Regulatory Build and the GM12878 specific regulatory activities are also shown. Our region marking tool is shown as a light grey box surrounding the transcripts PSMD9 and RP11–87C12.2.

control all downloads using an automatic checksum computation. We have also improved the installation script's warning handling. In addition version information has been added to the cache to improve debugging and data source tracking.

Web

This year has seen a number of incremental improvements to our website. Data export has been re-engineered to simplify extracting data from our resource and now enables the download of sequences, pairwise alignments, multiple alignments, orthologues and gene-trees. We also provide improved high-resolution images for publication and high contrast images for use in presentations. User data import has been enhanced through better support for UCSC Track Hubs and we now accept hubs with data from multiple species. Additionally we support composite tracks and have improved track labelling. Finally our website supports track

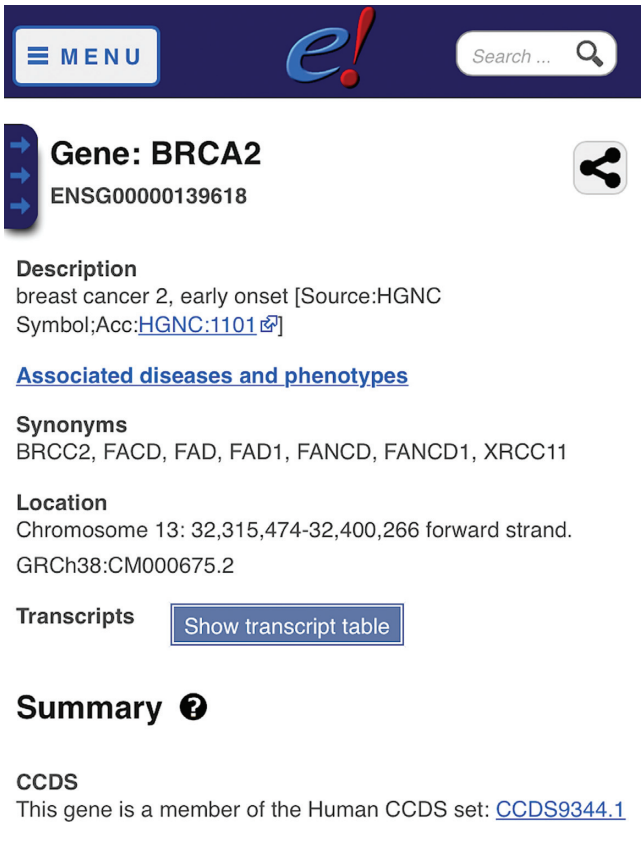

Figure 2. The new Ensembl mobile site showing BRCA2 and detailing available synonyms, genomic location and links to external resources such as CCDS. Search is available in the top right corner on all mobile site pages and each page has the ability to be shared over social media and email.

visibility settings allowing a hub to have a number of tracks enabled by default.

Gene Expression Atlas (GXA) baseline expression data is integrated into our website using a GXA provided JavaScript widget. Gene expression baseline levels are available for a number of studies including FANTOM5 (44) and GTEx (45). Our location view has been enhanced with two new interaction modes. In 'Select' mode, clicking and dragging with the mouse will select a resizable region and present a menu with options to either zoom or mark a region. The marked area is shown as a grey box and will remain marked until actively removed. This mark persists into our image exports as demonstrated in Figure 1. In 'Drag' mode, clicking on the image enables rudimentary scrolling navigation dragging the image to the left or right. Finally we released a new mobile optimized website, as shown in Figure 2, which can be accessed at http://m.ensembl.org or via optional redirection from our main website. The website is optimized for reduced display sizes and offers targeted views of genes, variants and phenotypes. Mobile users can opt to return to the full site when they require more advanced functionality.

REST service

The last year has seen substantial growth in both data and usage for our programming language agnostic REST API (46). All DNA from the previous five human assemblies (versions NCBI34 to GRCh38) are available by specifying the desired assembly version. Omitting a version assumes the latest assembly. Our VEP endpoint now supports annotation using HGVS variant nomenclature (e.g. AGT:c.803T>C) and querying for variants from a protein has been significantly improved. Building on our release in 2014, eight endpoints now support batch querying via the HTTP POST method including our sequence, identifier lookup and archive endpoints. We also support a number of GA4GH methods for retrieving sample genotype calls, variant calls on a reference sequence and for discovering available variant datasets and are actively working on a GA4GH variant annotation prototype.

eHive

eHive is the pipeline management system that powers a significant proportion of our compute, over 300 CPU years of compute per year and is versioned outside of the Ensembl release cycle (47). This year we have released version 2.3, which now supports a generic guest language interface to facilitate the writing of runnables in languages other than Perl through a standardized interprocess communication protocol. We have written a reference implementation in Python allowing Python and Perl code to be executed in the same workflow. Support for Java is under active development. Finally version 2.3 enhances our standard modules by improving their ability to capture and respond to erroneous system commands alongside improved security when interacting with databases. All development is continuously tested on Travis CI (a public continuous integration service) with 70% of the code tree covered by tests.

BioMart

Our BioMart databases continue to be updated every release in order to provide the latest annotation and imported data (48). We have made protein domain coordinates, transcript length and the previously described GENCODE Basic, TSL and APPRIS datasets available. Our relationship with biomaRt and Bioconductor/R has resulted in the release of dedicated subroutines to allow R developers to easily query our live, GRCh37 and archive BioMart services (49). Finally, in release 79 (March 2015), we redesigned our regulation BioMart to improve query performance and to meet the projected demands in data volumes. We have created seven datasets targeting distinct classes of data such as regulatory features as annotated by our methods, as well as binding motifs and miRNA targets. Consequently querying for data restricted by genomic location can be retrieved six times faster than using our previous BioMart.

OUTREACH AND TRAINING

External user support is provided by means of face-to-face training courses, online training materials, social media and email help channels. Annually we deliver roughly 100 workshops at research institutes and conferences around the world in person and through live webinars. Online training covering five Ensembl courses is available through the EMBL-EBI Train Online interface (http://www.ebi.ac.

uk/training/online/subjects/11), while our YouTube channel contains 35 training videos (https://www.youtube.com/user/EnsemblHelpdesk). Training material is also available via our help pages and workshops can be requested via our helpdesk.

Queries about working with Ensembl data, interfaces and APIs can be directed to our helpdesk (helpdesk@ensembl.org) or our public developers mailing list (dev@ensembl.org). We are active on social media channels such as Twitter (https://twitter.com/ensembl), Facebook (https://www.facebook.com/Ensembl.org) and our blog (http://www.ensembl.info/). For example, we regularly use #citedEnsembl hashtag on Twitter to highlight published research that has used Ensembl resources.

ACKNOWLEDGEMENTS

Thank you to Steve Moss for his previous work on the Ensembl REST API. We also thank Roy Storey for his eHive pull request, which triggered the development and deployment of eHive's test suite. We also thank Mikhail Spivakov for providing the sample interaction data shown in Figure 1.

FUNDING

Ensembl receives majority funding from the Wellcome Trust (grant numbers WT095908 and WT098051) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007234, 1R01HD074078, and U41HG007823), the Biotechnology and Biological Sciences Research Council (BB/I025506/1, BB/I025360/2, BB/K009524/1, BB/L024225/1, BB/M018458/1 and BB/M020398/1), the Centre for Therapeutic Target Validation (CTTV) and the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union's Seventh Framework Programme [FP7/2007-2013] under grant agreement [HEALTH-F4-2010-241504] (EURATRANS). The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 282510 (BLUEPRINT). The research leading to these results has received funding from the European Union's Seventh Framework Capacities Specific Programme under grant agreement n° 284209 (BioMedBridges). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics). Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
2. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
3. ENCODE,Project Consortium. (2012) An integrated encyclopaedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
4. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.-C., Agarwala,R., McLaren,W.M., Ritchie,G.R.S. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
5. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
6. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
7. Roadmap Epigenomics Consortium. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
8. Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T. and Flicek,P.R. (2015) The Ensembl Regulatory Build. *Genome Biol.*, **16**, 56.
9. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
10. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S.T., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
11. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
12. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D.S.T., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
13. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
14. Zhao,H., Sun,Z., Wang,J., Huang,H., Kocher,J.-P. and Wang,L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.
15. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
16. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R., Lunter,G., Marth,G., Sherry,S.T. *et al.* (2011) The Variant Call Format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
17. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
18. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
19. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2007) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
20. Rodriguez,J.M., Carro,A., Valencia,A. and Tress,M.L. (2015) APPRIS WebServer and WebServices. *Nucleic Acids Res.*, **43**, W455–W459.
21. Frankish,A., Uszczynska,B., Ritchie,G.R., Gonzalez,J.M., Pervouchine,D., Petryszak,R., Mudge,J.M., Fonseca,N., Brazma,A., Guigo,R. *et al.* (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, **16**, S2.
22. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *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**, 1316–1323.
23. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in

Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
24. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
25. Shimoyama,M., Pons,J., Hayman,G.T., Laulederkind,S.J.F., Liu,W., Nigam,R., Petri,V., Smith,J.R., Tutaj,M., Wang,S.-J. *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
26. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599–D601.
27. Hu,Z.-L., Park,C.A., Wu,X.-L. and Reecy,J.M. (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res.*, **41**, D871–D879.
28. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
29. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
30. Ivan Adzhubei,D.M.J. (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0720s76.
31. Zhou,X., Lowdon,R.F., Li,D., Lawson,H.A., Madden,P.A.F., Costello,J.F. and Wang,T. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods*, **10**, 375–376.
32. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
33. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J.M., Guo,Y., Hériché,J.-K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
34. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
35. Sonnhammer,E.L.L., Gabaldón,T., Sousa da Silva,A.W., Martin,M., Robinson-Rechavi,M., Boeckmann,B., Thomas,P.D., Dessimoz,C. and Quest for Orthologs consortium. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics*, **30**, 2993–2998.
36. The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1, 092 human genomes. *Nature*, **491**, 56–65.
37. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
38. Exome Aggregation Consortium. (2015) Cambridge, http://exac.broadinstitute.org.
39. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
40. Liu,X., Jian,X. and Boerwinkle,E. (2013) dbNSFP v2.0: a Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Hum. Mutat.*, **34**, E2393–E2402.
41. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
42. Illumina, Inc. (2015) Illumina Platinum Genomes. http://www.illumina.com/platinumgenomes/.
43. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
44. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
45. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
46. Yates,A., Beal,K., Keenan,S., McLaren,W., Pignatelli,M., Ritchie,G.R.S., Ruffier,M., Taylor,K., Vullo,A. and Flicek,P. (2015) The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, **31**, 143–145.
47. Severin,J., Beal,K., Vilella,A.J., Fitzgerald,S., Schuster,M., Gordon,L., Ureta-Vidal,A., Flicek,P. and Herrero,J. (2010) eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics*, **11**, 240.
48. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, **2011**, bar030.
49. Durinck,S., Spellman,P.T., Birney,E., Bolstad,B., Dettling,M., Dudoit,S. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.