D754–D761 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 16 November 2017*
doi: 10.1093/nar/gkx1098

Ensembl 2018

Daniel R. Zerbino[1], Premanand Achuthan[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Jyothish Bhai[1], Konstantinos Billis[1], Carla Cummins[1], Astrid Gall[1], Carlos García Girón[1], Laurent Gil[1], Leo Gordon[1], Leanne Haggerty[1], Erin Haskell[1], Thibaut Hourlier[1], Osagie G. Izuogu[1], Sophie H. Janacek[1], Thomas Juettemann[1], Jimmy Kiang To[1], Matthew R. Laird[1], Ilias Lavidas[1], Zhicheng Liu[1], Jane E. Loveland[1], Thomas Maurel[1], William McLaren[1], Benjamin Moore[1], Jonathan Mudge[1], Daniel N. Murphy[1], Victoria Newman[1], Michael Nuhn[1], Denye Ogeh[1], Chuang Kee Ong[1], Anne Parker[1], Mateus Patricio[1], Harpreet Singh Riat[1], Helen Schuilenburg[1], Dan Sheppard[1], Helen Sparrow[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Brandon Walts[1], Amonida Zadissa[1], Adam Frankish[1], Sarah E. Hunt[1], Myrto Kostadima[1], Nicholas Langridge[1], Fergal J. Martin[1], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Dan M. Staines[1], Stephen J. Trevanion[1], Bronwen L. Aken[1], Fiona Cunningham[1], Andrew Yates[1] and Paul Flicek[1,3,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Eagle Genomics Ltd., Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK and [3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received September 22, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 21, 2017

ABSTRACT

The Ensembl project has been aggregating, processing, integrating and redistributing genomic datasets since the initial releases of the draft human genome, with the aim of accelerating genomics research through rapid open distribution of public data. Large amounts of raw data are thus transformed into knowledge, which is made available via a multitude of channels, in particular our browser (http://www.ensembl.org). Over time, we have expanded in multiple directions. First, our resources describe multiple fields of genomics, in particular gene annotation, comparative genomics, genetics and epigenomics. Second, we cover a growing number of genome assemblies; Ensembl Release 90 contains exactly 100. Third, our databases feed simultaneously into an array of services designed around different use cases, ranging from quick browsing to genome-wide bioinformatic analysis. We present here the latest developments of the Ensembl project, with a focus on managing an increasing number of assemblies, supporting efforts in genome interpretation and improving our browser.

INTRODUCTION

Ensembl's purpose is to accelerate genomic research worldwide and amplify the impact of new discoveries by providing an openly-accessible window into the wealth of data produced by the scientific community. Genomic and epigenomic datasets are selected from public archives such as INSDC (1), ENA (2), dbSNP (3) or EVA (https://www.ebi.ac.uk/eva), downloaded, then processed by our multiple automated analysis methods. Their results are finally stored into an integrated array of databases and tailored storage solutions that are read by various APIs and utilities. Our web-based genome browser (http://www.ensembl.org) is in effect the visible tip of a very large underlying infrastructure.

The Ensembl project has grown with the field of genomics since the first releases of the draft human genome (4), when we launched our initial visualisation of the genomic sequence and the location of the genes within it (5). As the field expanded, so did Ensembl. Starting with mouse as the second sequenced vertebrate genome, we developed comparative genomic resources that now encompass multiple whole genome alignments and gene-level phylogenetic trees (6). Data from large variation discovery projects such as HapMap (7) and the 1000 Genome Project (8) were incorporated into our variation storage and annotation resources (9). International epigenomic surveys, including ENCODE

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

(10) and Blueprint (11), provided supporting evidence for our genome-wide annotation of regulatory elements (12). This expansion has led us to collaborate directly with major bioinformatics databases and resources such as UniProt (13), GENCODE (14), UCSC and NCBI (15).

In parallel, Ensembl is used in increasingly sophisticated ways. Our infrastructure now underlies an array of services for many different use cases. For quick queries, our web browser is likely the tool of choice. For genome-wide analyses, a few lines of code are sufficient to connect any computer to our databases via the Ensembl application programming interfaces (APIs, 16,17). Finally, for common data analysis workflows, we support dedicated tools, such as BioMart (18) and the Ensembl Variant Effect Predictor (VEP, (19)).

Our work has always been led by the firm conviction that scientific progress can only be accelerated by making data freely available as early as possible. This philosophy led to us adopting the FAIR principles of Findable, Accessible, Interoperable and Reusable (20) long before these were formalised. We define globally unique and persistent identifiers for our genes and other genomic features in our databases, our data are freely and programmatically available, we adhere to international naming standards and ontologies, and we track and credit the provenance of all our annotations. In addition to serving the community, these practices have the pragmatic effect of supporting our aim to develop a sustainable ecosystem of data services that can be automatically combined. Keeping up with the fast pace of genomics has required that our resources be integrated and inter-compatible.

Now that next-generation sequencing (NGS) is commonplace in many laboratories and that efficient bioinformatics toolkits have been developed, knowledge extraction is the bottleneck of genomics (21). Genomes are no longer sequenced one by one, rather in batches, and already Ensembl can display clades, as illustrated by our addition in 2016 of a collection of laboratory mouse strains (22). Furthermore, NGS machines are making their way into clinical laboratories and personal genotyping is becoming routine. To better support genome interpretation, we regularly enrich our annotation of variants. Finally, we are constantly improving our usability, for example by regularly refreshing our web interface with interactive selection tools to help guide visitors through the many available options and datasets.

COVERING MORE GENOMES

Ensembl release 90 (August 2017) included 15 new and updated annotated rodent genomes including two assemblies of the Chinese hamster ovary (CHO) cell line, male and female genome assemblies for naked mole-rat (*Heterocephalus glaber*), and three chromosome-level assemblies (*Mus pahari*, *Mus caroli* and *Microtus ochrogaster*). We generated these annotations using a combination of annotation mapping (via whole genome alignment) from the *Mus musculus* GENCODE gene set (14), alignments of a targeted subset of UniProt (13) vertebrate protein sequences and, where available, RNA-seq data. For *Mus caroli* and *Mus pahari* we imported the annotations generated by the Mouse Genomes Project (23). These new and updated rodent genomes join the 16 mouse strains whose annotations were imported in Ensembl release 86 (October 2016) (22).

We also annotated the newest pig reference assembly (Sus scrofa 11.1) using a combination of Illumina data from 28 tissues and PacBio IsoSeq data from nine tissues. With this wealth of transcriptome data, we annotated over 25 000 genes with almost 50 000 transcript isoforms, a significant increase when compared to the previous assembly. The annotation generated from individual tissues for both the Illumina and PacBio data are viewable as tracks in the browser and accessible via the API. We plan to update the gene count in subsequent releases by including manual annotation mapped from the previous assembly and more long non-coding RNA (lncRNA) annotation.

We updated the annotation of the GRCz10 zebrafish assembly using the extensive set of RNA-seq data aligned to the genome since the original annotation was completed, including the transcriptomes of 18 different developmental stages. These data added new UTRs, transcript isoforms and genes, including over 2000 lncRNAs. In addition, this final update to the GRCz10 gene set included a track of primary miRNA transcripts, assembled de novo using RNA-seq data and then mapped to the genome (https://www.biorxiv.org/content/early/2017/02/20/107631). These transcripts are viewable in the browser and can be compared to the existing zebrafish annotation across the genome.

We regularly updated the GENCODE gene set (14) for both mouse and human over the past year. Mouse, which is currently the subject of GENCODE's intensive manual annotation effort, has been updated every release. Because a first pass of manual annotation has been completed for human, updates are currently applied every other release only.

We annotated two different assemblies of the Chinese hamster ovary cell line CHOK1: CriGri_1.0 (GCA_000223135.1) and CHOK1GS_HDv1 (GCA_900186095.1). Both annotations were produced with the same pipeline but different inputs. First, the CriGri_1.0 assembly is older (released in 2011) than CHOK1GS_HDv1 (2017). Second, on CriGri_1.0 we used a selection of transcriptomic data from the European Nucleotide Archive, whereas for CHOK1GS_HDv1 we used transcriptomic data specifically produced for this annotation by the assembly provider and available at http://www.ebi.ac.uk/ena/data/view/PRJEB14303. Despite these differences in inputs, the two annotations contained similar numbers of genes in each category (except for lncRNA genes which will be added to CHOK1GS_HDv1 in a future release of Ensembl). Nonetheless, the CHOK1GS_HDv1 assembly had slightly more genes overall, likely reflecting the use of more recent genomic and transcriptomic data.

Ensembl's comparative genomics resources have all been updated to include the new genomes and updated assemblies, as illustrated on Figure 1. This represents a 25% increase of the number of pairwise-alignments available, as all genomes are aligned to human, and all rodents to mouse. The multiple alignments have also been expanded, increasing the size of the Eutherian mammals EPO-LOW-COVERAGE multiple-alignment from 40 genomes to 55. We added both naked mole-rat genomes and both CHO genomes to the main set of trees, allowing direct orthology

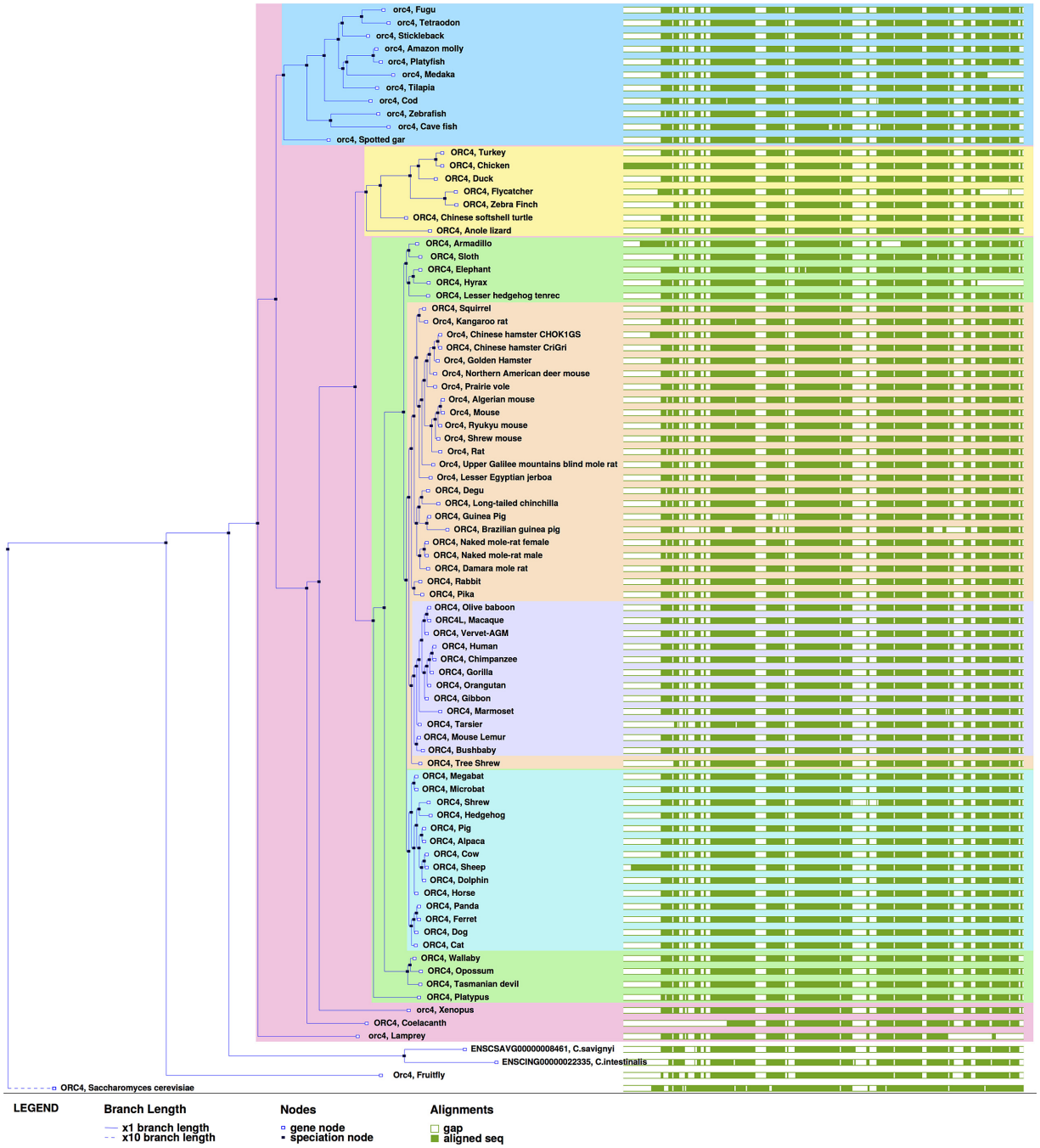

Figure 1. *ORC4* gene tree across 84 species.

calls to mouse, human and other key species, and we added *Mus caroli* and *Mus pahari* to the Murinae gene-trees and orthologues set.

The addition of multiple genomes of the same species in the gene trees is still under development. In the initial implementation described here, we have decided not to unilaterally promote one assembly for a given species as representative of all the others available. Instead, we inserted all gene sequences into the trees independently although this leads to a simplified representation of evolution. Specifically, intra-species evolution is generally marked by recombination meaning that the history of the genes is correctly represented as a directed acyclic graph known as an Ancestral Recombination Graph (ARG) (24). As an intermediate step, we will soon update the gene trees in the relevant species so that relationship between the separate assemblies are appropriately labelled.

In our TreeFam gene homology resource, we increased the sensitivity of our orthology-calling methods, especially for short sequences. The proportion of proteins shorter than 50 amino-acids with at least one homologue rose from 43% to 50%, and the proportion of proteins shorter than 20

amino-acids with at least one homologue rose from 1.3% to 25%. All our protein-families and gene gain/loss trees can now be retrieved from our public REST API, which expands the available programmatic options beyond our Perl API.

We regularly update the links to external references for all 97 chordate species in Ensembl. For the newly added mouse strains, where little strain-specific external data is available, these links have been inferred from the reference mouse. Genes in a strain which had a one-to-one ortholog in the mouse GRCm38 reference assembly thus inherited links relevant to that gene. We also mapped murine microarray probes to all the different strain genomes.

SUPPORT FOR GENOME INTERPRETATION

Our updates to the Ensembl REST server can be more frequent than the time taken to complete long-running research projects or applications that require consistent analysis against the same Ensembl release. For these reasons, we now maintain archives of the REST server starting with Ensembl release 87 (e.g. http://mar2017.rest.ensembl.org/). Archives will be available for at least five years from their initial release to enable consistent and reproducible analysis for publications and other genomic workflows.

A number of large scale sequencing projects, such as the Genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org/), UK10K (25) and NHLBI Trans-Omics for Precision Medicine (TOPMed, https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed) projects, have made allele frequency data available this year in addition to the 1000 Genomes Phase 3 data (8). We have extended our API to efficiently produce frequency data from these cohorts. To help filter out common variants when performing association studies, we report the highest minor allele frequency observed in any population in the 1000 Genomes, gnomAD and TOPMed projects, both on our variant pages (see Figure 2A) and via our Perl API. We now also make linkage disequilibrium (LD) plots available for insertions and deletions on our website.

This year we significantly updated the VEP code to improve its robustness and functionality. In particular we enhanced our analysis of RefSeq human transcripts. Predicting the functional consequences of variants on RefSeq transcripts that differ from the reference assembly can be a challenge. To deliver more accurate results, the VEP now uses NCBI's alignments of these transcripts onto the genome to expose any differences. Additionally, the VEP now predicts the impact of missense variants on the protein function of RefSeq transcripts using SIFT (26) and PolyPhen 2 (27). New plugins support more detailed descriptions of variants located near splice sites, loss of function intolerant scores for genes (28), and additional measures of variant deleteriousness (https://www.biorxiv.org/content/early/2017/06/12/148353).

We import phenotype associations from many different sources into Ensembl. Often, we encounter the same disease or trait in different databases under different labels (e.g. Type 2 diabetes and 'diabetes, type II'). We now map these descriptions to ontology terms, thus bringing together records describing the same disease under different names, as well as different subtypes of traits and diseases sharing the same phenotypes. This improves the ability to query results aggregated across many sources. It also improves the legibility of our phenotype tables, which are now grouped by ontology term, as shown on Figure 2B. As these tables may contain hundreds of records, we added filters to display only selected results based on attributes such as locus type (e.g. genes or variants) or data source. For non-human species, in particular mouse and a number of livestock species, we now map to the Mammalian (https://github.com/obophenotype/mammalian-phenotype-ontology) and Clinical Measurement (29) ontologies respectively. New REST endpoints have been created to allow programmatic access to these mapped results across all species.

Similarly, variants can appear under a number of identifiers including dbSNP RefSNP identifiers, ClinVar (30) accessions and the Human Genome Variation Society (HGVS, 31) nomenclature at genomic, transcript or protein level. Identifiers used in past publications are often made obsolete over time, making it difficult to link them to current knowledge. To help address this problem, we have implemented a REST endpoint that returns all currently known identifiers for a given variant name. Many types of malformed and redundant HGVS descriptions are correctly interpreted and all possible variant identifiers returned.

In collaboration with RNAcentral and University College London, we added GO term annotations for some for non-coding transcripts as of Ensembl release 89 (May 2017). These supplement the protein-coding transcript GO terms Ensembl has included for many years from UniProt.

As the number of Ensembl's tools and services grows, we recognise the need for our infrastructure to be quickly and easily deployed locally. This supports use cases such as independent annotation (32) or clinical genetics applications that cannot send queries to our servers for privacy concerns. We now provide an automated deployment tool using Ansible (https://github.com/Ensembl/ensembl-rest-deploy) to go from a fresh VM to a ready-to-deploy Ensembl REST service. Similarly, the external dependencies of the Ensembl analysis methods can be installed very rapidly on any system using our Homebrew recipes (https://github.com/Ensembl/homebrew-ensembl). Both are already successfully used within the project to set up our analysis tools, deploy REST for a new release as well as for the VM available from the FTP site. These tools supplement parallel efforts such as GenomeHubs, which support rapid local deployment of the Ensembl databases and web server (33).

Finally, we now distribute intermediary results of our epigenomic processing pipeline, which are particularly useful for the high-throughput genome-wide reanalysis of non-coding variants. These include BigWig files of all the consistently mapped ChIP-seq datasets as well as segmentation BigBed files of all the epigenomes included in Ensembl.

AN INTUITIVE BROWSING EXPERIENCE

A new gallery portal showcases the wealth of visual interfaces available in Ensembl (http://www.ensembl.org/info/website/gallery.html). Searches of the gallery return thumbnail images for all relevant views associated with specific

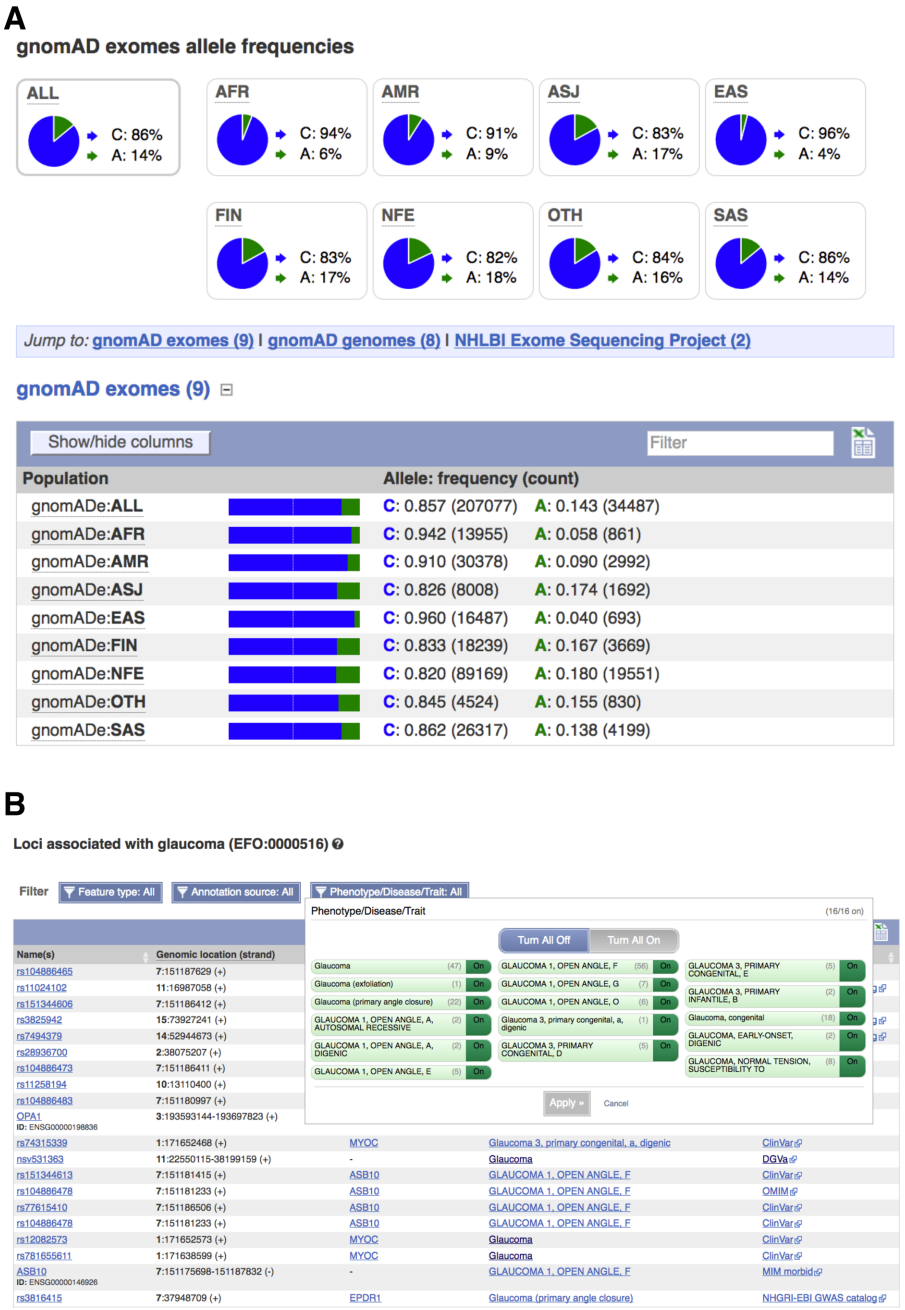

Figure 2. Detailed variant reports. (**A**) Variant minor allele frequencies of a given variant across gnomAD populations. (**B**) Variant association table, as returned by the ontology-aware search, that contains results closely related to the query, in this case 'glaucoma'.

genes, variants or genomic locations (see Figure 3B). Because of the number of views available, they are grouped by themes. For example, for a given SNP identifier, there are views relating to its region, overlapping transcripts, overlapping genes, overlapping protein sequences, associated phenotypes and population genetics. The gallery portal is designed to help newcomers discover unknown resources in Ensembl, as well as more experienced users jump directly to specific services.

A new interface to search for and select species within our tools and comparative views has been released. The new design (see Figure 3A) allows us to present the available species in a hierarchy of clades and we hope this will prove indispensable as the number of species increases over the next few years.

As we continue to deliver training courses around the world, we developed a new training website (http://training.ensembl.org) to distribute training materials from these courses, including slides, screenshot-by-screenshot walkthroughs of the website and hands-on exercises with answers. Materials from previous courses are available in perpetuity, allowing course participants to access them both during and after the course. The materials carry a Creative Commons BY license, allowing other trainers to use and

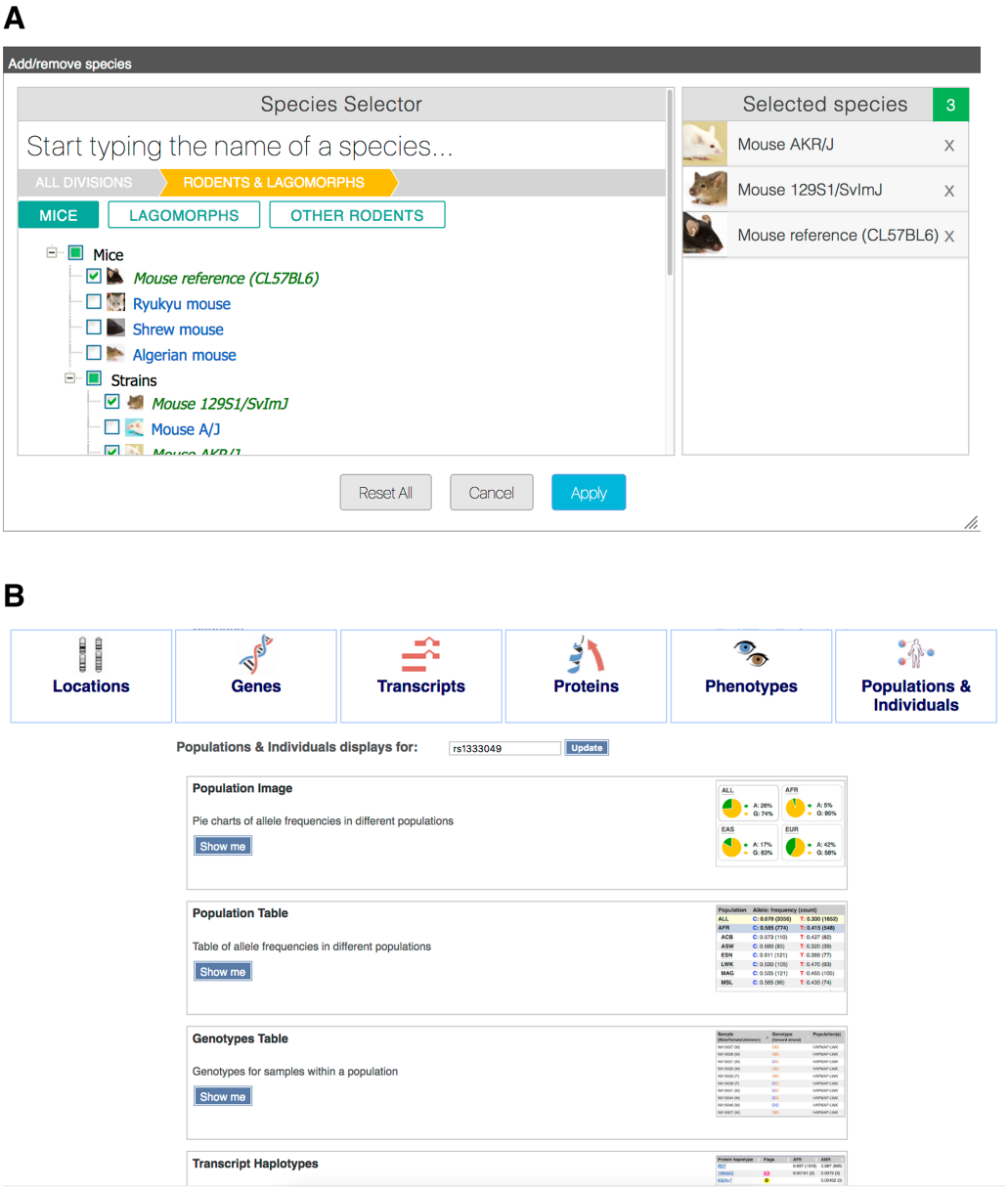

Figure 3. Quick selection menus. (**A**) The species selection tool can be used to quickly search for species by clade. (**B**) The plot gallery produces direct links to all Ensembl views and resources regarding a gene, variant or locus of interest.

adapt these materials for their own training. The site also includes information for hosts wanting to invite Ensembl for their own course, and links to courses that participants can register for.

CONCLUSION

The Ensembl infrastructure is keeping up with the pace of cutting-edge genomics research, in scale, breadth and complexity. Thanks to robust engineering, we are developing new applications from our existing resources, and regularly upgrading the latter when needed. Our current priorities are scaling up to more species, delivering useful services for genome interpretation and improving the web interface. Given the unrelenting pace of genomics, we expect to be pursuing these efforts for years to come.

AVAILABILITY

The Ensembl website (http://www.ensembl.org) provides access to all of our services and documentation, including the REST API (http://rest.ensembl.org) and BioMart (http://www.ensembl.org/biomart/). Ensembl imposes no restrictions on access to, or use of, the data provided and the software used to analyse and present it. All Ensembl code is available on Github (http://www.github.com/Ensembl/) under the Apache 2.0 licence.

Queries about hosting Ensembl workshops and any other questions about Ensembl can be directed to our helpdesk

(helpdesk@ensembl.org). We can also be contacted informally via social media platforms, including Twitter (@ensembl) and Facebook (Ensembl.org). Our blog posts include detailed descriptions of every Ensembl release and other information (http://www.ensembl.info).

ACKNOWLEDGEMENTS

We thank all our users for their regular questions and feedback. We thank the Enright group at EMBL-EBI for creating and mapping the primary miRNA transcripts for the GRCz10 assembly. We thank Abel Ureta-Vidal for assistance with the CHOK1GS_HDv1 annotation. For their assistance with the Ensembl infrastructure and hardware migration we acknowledge Simone Badoer, Andy Bryant, Liz Beresford, Andy Cafferkey, Andrea Cristofori, Jonathan Barker, Pete Jokinen, Rodrigo Lopez, Manuela Menchi, Sundeep Nanawa, Steven Newhouse and Jordi Vallis from the EMBL-EBI Technical Services Cluster as well as Peter Clapham, Paul Bevan, Luke Burling, Martin Burton, Mark Flint and Jonathan Nicholson from the Wellcome Trust Sanger Institute.

FUNDING

Ensembl receives majority funding from the Wellcome Trust [WT108749/Z/15/Z] with additional funding for specific project components from the National Human Genome Research Institute of the National Institutes of Health [U41HG007234, U41HG007823, U41HG007823-S1, 2U41HG007234]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Specific project components are also funded by the Biotechnology and Biological Sciences Research Council [BB/L024225/1, BB/M011615/1, BB/M020398/1]; Open Targets; the Wellcome Trust [WT104947/Z/14/Z, WT200990/Z/16/Z, WT201535/Z/16/Z]; European Molecular Biology Laboratory. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. [634143] (MedBioinformatics). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement [733161] (MultipleMS). This project has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement [115582], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme [FP7/2007–2013] and EFPIA companies' in kind contribution. We also receive funding from the 'Save the Tasmanian Devil Program'.

Conflict of interest statement. Paul Flicek is a member of the Scientific Advisory Board of Fabric Genomics, Inc., and Eagle Genomics, Ltd. Daniel Barrell is an employee of Eagle Genomics, Ltd.

REFERENCES

1. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and the International Nucleotide Sequence Database Collaboration (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
2. Toribio,A.L., Alako,B., Amid,C., Cerdeño-Tarrága,A., Clarke,L., Cleland,I., Fairley,S., Gibson,R., Goodgame,N., ten Hoopen,P. *et al.* (2017) European Nucleotide Archive in 2016. *Nucleic. Acids. Res.*, **45**, D32–D36.
3. 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.
4. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic. Acids. Res.*, **30**, 38–41.
5. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.
6. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**, bav096.
7. The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
8. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
9. Chen,Y., Cunningham,F., Rios,D., McLaren,W.M., Smith,J.A., Pritchard,B., Spudich,G.M., Brent,S., Kulesha,E., Marin-Garcia,P. *et al.* (2010) Ensembl variation resources. *BMC Bioinformatics*, **11**, 293.
10. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
11. Adams,D., Altucci,L., Antonarakis,S.E., Ballesteros,J., Beck,S., Bird,A., Bock,C., Boehm,B., Campo,E., Caricasole,A. *et al.* (2012) BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol*, **30**, 224–226.
12. Zerbino,D.R., Johnson,N., Juettemann,T., Sheppard,D., Wilder,S.P., Lavidas,I., Nuhn,M., Perry,E., Raffaillac-Desfosses,Q., Sobral,D. *et al.* (2016) Ensembl regulation resources. *Database*, **2016**, bav119.
13. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic. Acids Res.*, **45**, D158–D169.
14. 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.
15. 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.
16. Ruffier,M., Kähäri,A., Komorowska,M., Keenan,S., Laird,M., Longden,I., Proctor,G., Searle,S., Staines,D., Taylor,K. *et al.* (2017) Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database*, **2017**, bax020.
17. 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. *Bioinform.*, **31**, 143–145.
18. 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*, **2011**, bar030.
19. McLaren,W., Gil,L., Hunt,S.E., Singh Riat,H., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
20. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., Bonino da Silva Santos,L., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.
21. Sboner,A., Mu,X.J., Greenbaum,D., Auerbach,R.K. and Gerstein,M.B. (2011) The real cost of sequencing: higher than you think!. *Genome Biol.*, **12**, 125.
22. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic. Acids Res.*, **45**, D635–D642.

23. Adams,D.J., Doran,A.G., Lilue,J. and Keane,T.M. (2015) The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mamm. Genome*, **26**, 403–412.
24. Griffiths,R.C. and Marjoram,P. (1997) An ancestral recombination graph. In: Donnelly,P and Tavaré,S (eds). *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and Its Applications 87*. Springer-Verlag, Berlin, pp. 257–270.
25. The UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.
26. Sim,N.L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic. Acids Res.*, **40**, W452–W457.
27. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
28. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
29. Shimoyama,M., Nigam,R., Sanders McIntosh,L., Nagarajan,R., Rice,T., Rao,D.C. and Dwinell,M.R. (2012) Three ontologies to define phenotype measurement data. *Front Genet.*, **3**, 87.
30. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitiparalla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic. Acids Res.*, **44**, D862–D868.
31. den Dunnen,J.T., Dalgleish,R., Maglott,D.R., Hart,R.K., Greenblatt,M.S., McGowan-Jordan,M., Roux,A.F., Smith,T., Antonarakis,S.E. and Taschner,P.E.M. (2016) HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mut.*, **37**, 564–569.
32. Eöry,L., Gilbert,M.T.P., Li,C., Li,B., Archibald,A., Aken,B.L., Zhang,G., Jarvis,E., Flicek,P. and Burt,D.W. (2015) Avianbase: a community resource for bird genomics. *Genome Biol.*, **16**, 21.
33. Challis,R.J., Kumar,S., Stevens,L. and Blaxter,M. (2017) GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species. *Database*, **2017**, bax039.