Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1011

Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species

Paul Julian Kersey[1,*], James E. Allen[1], Alexis Allot[1], Matthieu Barba[1], Sanjay Boddu[1], Bruce J. Bolt[1], Denise Carvalho-Silva[1], Mikkel Christensen[1], Paul Davis[1], Christoph Grabmueller[1], Navin Kumar[1], Zicheng Liu[1], Thomas Maurel[1], Ben Moore[1], Mark D. McDowall[1], Uma Maheswari[1], Guy Naamati[1], Victoria Newman[1], Chuang Kee Ong[1], Michael Paulini[1], Helder Pedro[1], Emily Perry[1], Matthew Russell[1], Helen Sparrow[1], Electra Tapanari[1], Kieron Taylor[1], Alessandro Vullo[1], Gareth Williams[1], Amonida Zadissia[1], Andrew Olson[2], Joshua Stein[2], Sharon Wei[2], Marcela Tello-Ruiz[2], Doreen Ware[2,3], Aurelien Luciani[1], Simon Potter[1], Robert D. Finn[1], Martin Urban[4], Kim E. Hammond-Kosack[4], Dan M. Bolser[1], Nishadi De Silva[1], Kevin L. Howe[1], Nicholas Langridge[1], Gareth Maslen[1], Daniel Michael Staines[1] and Andrew Yates[1]

[1]The European Molecular Biology Laboratory, The European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [2]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, USA, [3]USDA-ARS NAA Plant, Soil and Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY 14853, USA and [4]Rothamsted Research, Department of Biointeractions and Crop Protection, Harpenden, Hertfordshire, AL5 2JQ, UK

Received September 18, 2017; Revised October 06, 2017; Editorial Decision October 12, 2017; Accepted October 24, 2017

ABSTRACT

Ensembl Genomes (http://www.ensemblgenomes.org) is an integrating resource for genome-scale data from non-vertebrate species, complementing the resources for vertebrate genomics developed in the Ensembl project (http://www.ensembl.org). Together, the two resources provide a consistent set of programmatic and interactive interfaces to a rich range of data including genome sequence, gene models, transcript sequence, genetic variation, and comparative analysis. This paper provides an update to the previous publications about the resource, with a focus on recent developments and expansions. These include the incorporation of almost 20 000 additional genome sequences and over 35 000 tracks of RNA-Seq data, which have been aligned to genomic sequence and made available for visualization. Other advances since 2015 include the release of the database in Resource Description Framework (RDF) format, a large increase in community-derived curation, a new high-performance protein sequence search, additional cross-references, improved annotation of non-protein-coding genes, and the launch of pre-release and archival sites. Collectively, these changes are part of a continuing response to the increasing quantity of publicly-available genome-scale data, and the consequent need to archive, integrate, annotate and disseminate these using automated, scalable methods.

OVERVIEW AND ACCESS

Ensembl Genomes (http://www.ensemblgenomes.org) is organised as five sites, each focused on one of the traditional kingdoms of life: bacteria, protists, fungi, plants and (invertebrate) metazoa. Vertebrate metazoa are the focus of the Ensembl project (1); Ensembl Genomes provides a complementary set of interfaces for non-vertebrate species. Our goals are to provide high-quality reference genome sequence and annotation for every species for which these are available; to represent genomic diversity for all species of major research interest; to link out to phenotypic data and resources containing biological material; and to provide a set of tools that allows users to interrogate these data in con-

*To whom correspondence should be addressed. Tel: +44 1223 494601; Fax: 44 1223 494468; Email: pkersey@ebi.ac.uk
Present addresses:
B. Bolt, The Pirbright Institute, Ash Road, Pirbright, Woking GU24 0NF, UK.
C. Grabmueller, Geospock Limited, Saint Andrew's House, St Andrew's Rd, Cambridge, Cambridge CB4 1DL, UK.

Table 1. Growth in Ensembl Genomes 2015–2017

Release version	Date	Number of genomes				
		Number in brackets indicates genomes directly imported from INSDC.				
		Ensembl Bacteria	Ensembl Protists	Ensembl Fungi	Ensembl Plants	Ensembl Metazoa
28	August 2015	23 001 (23,001)	133 (101)	407 (359)	41 (8)	55 (1)
37	September 2017	44 048 (44,048)	189 (157)	811 (760)	45 (12)	68 (2)
Increase		21 047 (21,047)	56 (56)	404 (401)	4 (4)	13 (1)

junction with their own. This paper describes the current state of the resource and ongoing progress towards these aims.

For all species included in the resource, we currently provide access to genome sequence and annotations of protein-coding and non-coding genes. Transcriptional, genetic variation, and comparative analysis data are additionally available for many species. For most species, these data are automatically imported using standard pipelines from the archives of the International Nucleotide Sequence Database Consortium (INSDC), i.e. the European Nucleotide Archive (2), GenBank (3), and the DNA Database of Japan (4), the European Variation Archive (http://www.ebi.ac.uk/eva), Wikipedia, and other open access sources. For a few species of particular research or socio-economic importance, additional high-value data sets are identified and manually imported. A link to 'information and statistics' for each genome provides information including the methods of assembly, references to publications and archival submission numbers, and the date the assembly was first incorporated into the resource.

Interactive access to all data is provided through a web interface providing genome browsing capabilities: users can scroll through a graphical representation of a DNA molecule at various levels of resolution, seeing the relative locations of features—including conceptual annotations (e.g. genes, SNP loci), sequence patterns (e.g. repeats) and experimental data (e.g. expressed RNA sequences mapped onto the genome) that supports the primary annotations. Functional information is provided through direct curation, import from the UniProt Knowledgebase (5), or imputation from protein sequence (using the classification tool InterProScan (6)). Various tools for text and sequence search, data upload and data analysis are available, allowing researchers to examine their own data in the context of the reference sequence and annotation.

Ensembl data have traditionally been stored in a set of MySQL databases which can be directly accessed via a public MySQL server (host: mysql.ebi.ac.uk port: 4157 username: anonymous) and additionally through well-developed Perl and RESTful APIs that provide an object-oriented framework for working with genomic data. Increasingly the Ensembl web application directly utilizes data files stored in archival resources (such as the European Nucleotide archive), avoiding the need for database builds and improving the speed of response. All data in the resource is open-access, and both database dumps and common data sets (e.g. DNA, RNA and protein sequence sets and sequence alignments) can be directly downloaded in bulk via FTP (ftp://ftp.ensemblgenomes.org).

Ensembl Genomes data is also organised in additional databases, constructed using the BioMart data warehousing system (7), optimised around common gene- and variant-centric queries. The BioMart framework provides web-based query building tools, and a variety of other interfaces for interactive and programmatic access. BioMarts are not currently available for Ensembl Bacteria.

Ensembl Genomes is updated 4–5 times a year in synchrony with updates to Ensembl, utilising the same software as the corresponding Ensembl release. The overall suite of Ensembl Genomes interfaces mirrors those provided for vertebrate genomes in Ensembl, allowing users to access genomic data from across the tree of life in a consistent manner. In addition, Ensembl Genomes contributes to collaborative database projects focused on various domains of life, including Gramene (http://www.gramene.org) (8) for plants, PhytoPath (http://phytopathdb.org) (9) for plant pathogens, VectorBase (http://www.vectorbase.org) (10) for invertebrate vectors of human pathogens, and WormBase (http://www.wormbase.org) (11) for helminths. In these projects, we work with our partners to develop common datasets, which are made available through both Ensembl Genomes and additional project-specific interfaces.

NEW AND IMPROVED GENOME ASSEMBLIES

Ensembl Genomes has continued to grow in 2016 and 2017 (see Table 1). The resource contains all annotated assemblies from fungal and protist species that are present in the INSDC, and it is planned to extend this approach to plants and metazoa within the next year. However, owing to the very large number of bacterial genome assemblies now available, a filter has been applied from release 35 onwards to exclude new genome assemblies which fail to add significant diversity to the overall collection. This approach mirrors that already adopted by the UniProt Knowledgebase for filtering data from bacterial genomes (12).

Species of particular societal, research or taxonomic interest that have been recently incorporated include *Bombus impatiens* (the common bumblebee) (13), *Octopus bimaculoides* (the California two-spot octopus) (14), *Sarcoptes scabiei* (the itch mite, the cause of scabies) (15), *Beta vulgaris* (sugar beet) (16) and *Brassica napus* (rapeseed) (17). Several existing assemblies have also been upgraded, and a number of previously highly fragmented genomes have now been incorporated in more contiguous forms. Cereal genomes are of particular interest, owing to their large size (at 16 Gb the polyploid bread wheat genome is the largest genome currently represented in the resource) and complex repeat structure, which have historically made them difficult to assemble. However, recent advances in technology

Table 2. RNA-seq alignment tracks by division

Division	Tracks	Experiments	Species
Protists	71	36	3
Fungi	6384	4822	24
Plants	29 836	1418	43
Metazoa	198	105	34

are now yielding dramatically improved genome assemblies even for cereals. For example, the latest assembly of the barley genome (18) has been added to the resource and comprises just 6,347 scaffolds with an N50 of 1.9 Mb (cf. the previous assembly, which contained 376 261 unscaffolded contigs of over 1 Kb in length with an N50 of just 1.4 Kb). While this might not yet be a complete molecular assembly, it is closer to a finished state than many smaller genomes in the resource that were sequenced and assembled using previous technologies. Similarly, the bread wheat genome is also undergoing rapid improvement. A significantly improved new assembly, the TGAC1.0 assembly (19), has already been incorporated in the resource, and we are currently working on a further upgrade to incorporate the IWGSC RefSeq v1.0 assembly (currently available at https://www.wheatgenome.org).

INCREASED DATA FROM COMMUNITY ANNOTATORS

Through our involvement in the VectorBase project, we are able to provide community-provided gene models (as modifications or extensions to the previous genome-wide annotation) for 26 genomes. Community members can access an instance of Apollo (20), an online genome editing tool, to assess evidence, and submit proposed changes, which are quickly visible in the browser and which are subsequently assessed for inclusion in a revised gene set. We have subsequently expanded our support for community annotation to enable the complete re-annotation of two fungal phytopathogen species, *Botrytis cinerea* (21) and *Blumeria graminis*, by members of their respective communities, and have incorporated the revised gene sets within Ensembl Fungi. We are currently working with the *Zymoseptoria tritici* community in a similar initiative, and are exploring ways of providing generic access to Apollo for all species in future.

INTEGRATION OF RNA-SEQ DATA

Relatively recently, transcriptional evidence for gene models was scarce for many non-model species. Today, data from many thousands of RNA-Seq experiments are present in the nucleotide sequence archives; however, the raw read sequence is not immediately useful. We have therefore developed a pipeline to automatically identify sequence read data in the INSDC archives and align them to the corresponding genomic sequence. These alignments are stored in Compressed Read Alignment Map (CRAM) format (22) and are resubmitted to the ENA for persistent archiving. Data from technical replicates are merged by default. To make these thousands of tracks accessible in Ensembl Genomes, alignments derived from a single experiment ('Study' in the ENA data model) are organised in track hubs (23), a convenient format that can group sets of related positional data prior to their visualisation as tracks in a genome browser. To date, alignments have been generated for plants, invertebrate vectors and plant pathogens, and will shortly be produced for other fungi, protists, and metazoan species. A summary of currently available alignments is shown in Table 2.

Track hubs are stored and indexed in a dedicated registry (http://trackhubregistry.org), and a search interface over this registry has been implemented in the Ensembl browser. Users of the browser can directly identify hubs containing data located on the genome they are currently browsing, filter the list to select only those hubs whose annotated meta data matches a given search term, and then select tracks from within the chosen hub for visualization. This process is illustrated in Figure 1. Researchers can also submit their own track hubs directly to the registry, and thereby expose their data through Ensembl Genomes and other track-hub compliant browsers.

FAST PROTEIN SEQUENCE SEARCH WITH HMMER

A new fast, accurate protein sequence search has been introduced, utilising the HMMER3 tool (24), which uses Hidden Markov Models to find matching sequences. The search has been implemented by indexing Ensembl Genomes protein sequences within an existing public HMMER3 server (25), and connecting this server to pages for the entry of query sequence and the visualisation of results within the Ensembl Genomes site. After a search has completed, users are shown a taxonomic breakdown of significant hits, a presentation of the alignments of query and target sequences, and a view of the domain architecture of the top hit (see Figure 2). BLAST search (26) of both protein and nucleotide sequences continues to be available.

PRE-RELEASE SITES AND ARCHIVE SITES

On publication of an updated version of a genome assembly of an important species, we use pre(-release) sites to make the initial data quickly available before we have had time to recompute the full range of analyses performed on previous version. The previous version remains available in the normal site until the newer version is fully described. Genomes are removed from the pre-site when analysis is complete and the new assembly is ready to migrate to the main site. Pre-sites are accessed at URLs such as http://pre.metazoa.ensembl.org, and are advertised prominently within the main site when available. Assemblies recently made available on pre-sites include new assemblies for *Beauveria bassiana* (a parasite of arthropods), *Fusarium graminarium* and *Fusarium culmorum* (both plant pathogens), *Hordeum vulgare* (barley), *Triticum aestivum* (bread wheat),

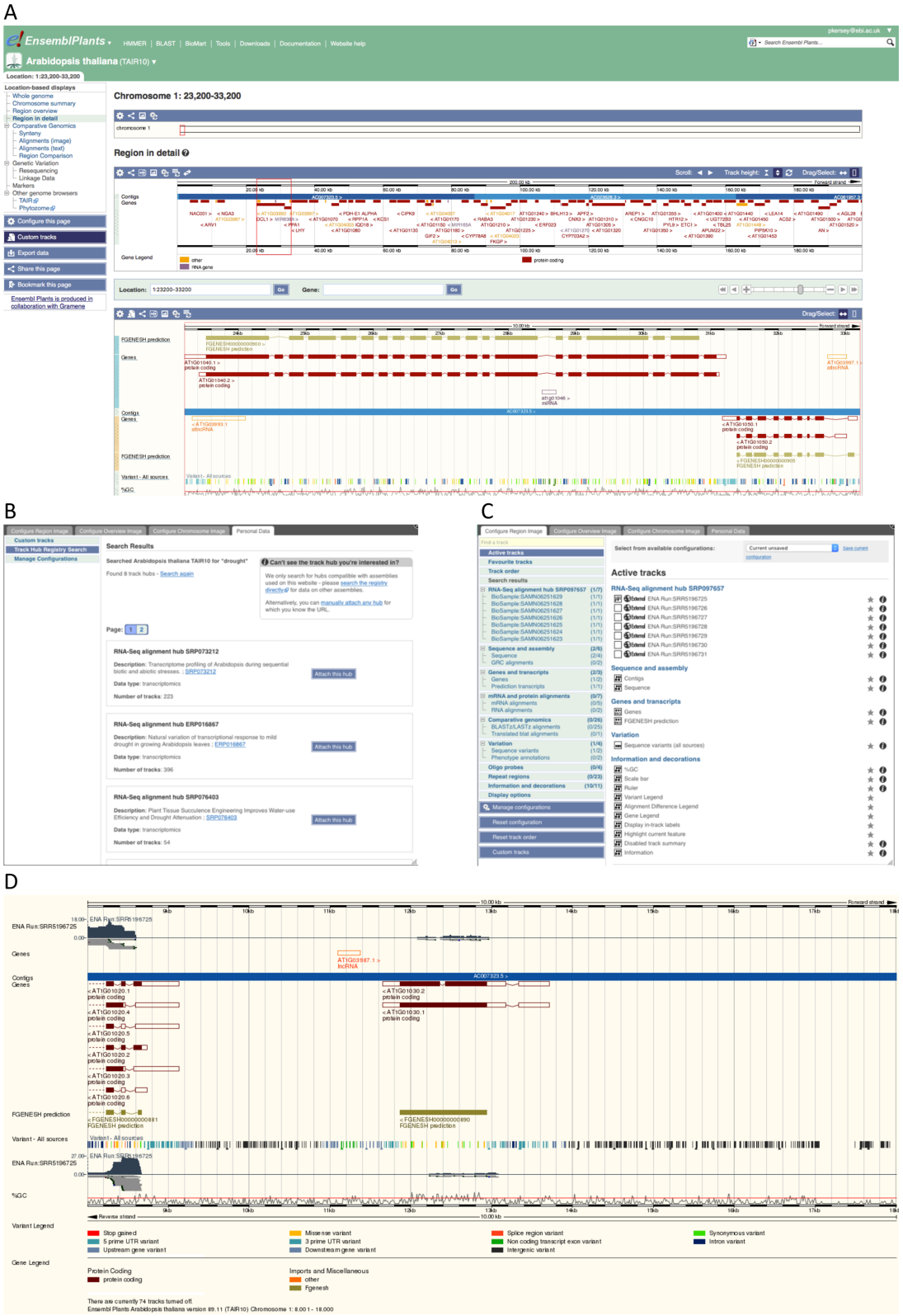

Figure 1. Data discovery and visualisation of alignment data in Ensembl Plants. While browsing the genome, a user can follow the 'Custom tracks' option in the left-hand menu (panel **A**). This gives access to a menu that searches the track hub registry for hubs that are anchored on this genome and match a specified metadata search term (panel **B**). Having selected a hub, the user is led through to a second menu where the tracks contained in the hub can be figured for display (panel **C**). When the configuration is complete, the selected tracks appear in the browser as selected (panel **D**).

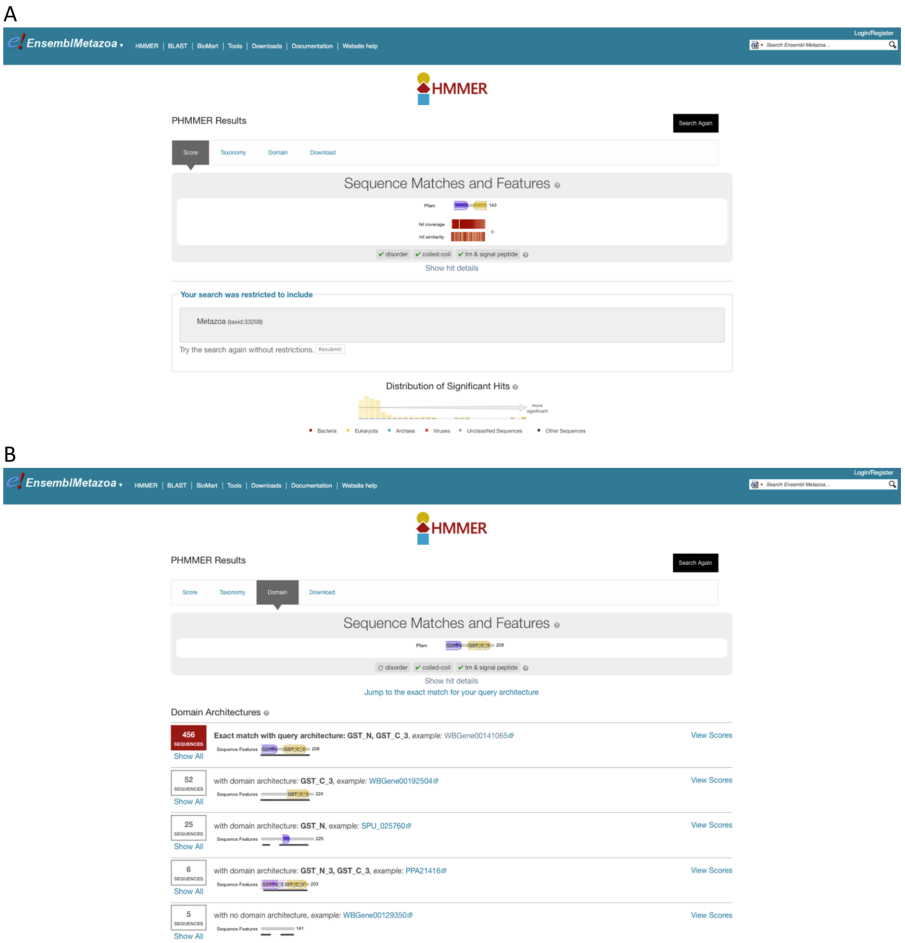

Figure 2. Hidden Markov Model search integrated into Ensembl Genomes. Results from a protein search using Hidden Markov Models, implemented using HMMer3, in Ensembl Genomes. Various options are available in a tabbed display including (i) a description of the domain architecture of the query sequence, and a graphical summary of the distribution of the significance scores of matches against the library (panel **A**) (ii) a breakdown of matching library sequences ordered by domain architecture (panel **B**).

and the diatom *Phaeodactylum tricornutum*, all of which have subsequently migrated to the live site.

While improved assemblies are obviously desirable, they can be problematic for researchers currently attempting to complete a lengthy data analysis, and the loss of previous versions from the website also makes it harder for scientists to check on results previously published. Moreover, genome alignment tracks lose utility if the reference sequence used is no longer available to view in a browser. Ideally, older assemblies would remain available in the browser, even after new versions have been created. As a first step towards achieving this, we have made available an archived version of release 32 of Ensembl Plants, alongside the live version. It is planned to shortly deploy archival versions of all Ensembl Genomes sites containing the release 37 data set, and thereafter to supplement these with approximately annual updates.

OTHER IMPROVEMENTS

Annotation of non-coding RNA genes is often poor or non-existent in archival submissions. An updated pipeline has been written for the identification of non-coding RNA genes, using updated versions of Rfam (27) and tRNAscan-SE (28) and improved filtering of the results, and has been applied to 162 eukaryotic genomes, resulting in an additional 213 717 gene annotations (an average of ∼1300 per genome). These are accessible alongside protein-coding annotations in the database downloads and browser.

We have improved our integration with PHI-base (29), a database of genes involved in plant pathogenesis, using sequence similarity to locate genes not linked to the genome. The number of cross-referenced genes has increased from 1491 to 2756, which comprises 98.9% of the potentially mappable genes. Plant genes have been linked to pathways in the Plant Reactome (http://plantreactome.gramene.org) (30) database.

Finally, Ensembl Genomes is now available for download in RDF format.

FUTURE PERSPECTIVES: FISHING IN THE DATA DELUGE

For some years, as genome sequencing technology has continued to improve, it has been forecast every organism of interest would soon have a completed genome sequence. Yet

while the quantity of published sequence has steadily increased, the best assembly available for many species has continued to be highly fragmented (and indeed, many recent genome assemblies have been more fragmented than those produced with earlier technologies). However, the availability of new assemblies for wheat and barley, and the increasing availability of unbroken whole chromosome assemblies for smaller genomes (e.g. many fungal species), indicates that the era of universal reference genome sequences is finally dawning. Since Ensembl Genomes organises data around contiguous sequences, the challenge of data presentation is simplified as assemblies become more complete; in addition, more contiguous assemblies are likely to better represent repeat structure, heterozygosity, and other phenomena that can lead to a mis-interpretation of the true genomic content of an organism.

Nonetheless, Ensembl Genomes faces various challenges as the total quantity of available data continues to rise. Firstly, it becomes increasingly important that access to data is provided computationally as well as via interactive interfaces. Ensembl and Ensembl Genomes have always provided a variety of data downloads and APIs for this purpose, and the availability of data in RDF format represents a further offering in this respect. Secondly, data processing pipelines need to be sufficiently automatic and performant to be able to process the available volume of data. The implementation of procedures for the automatic import of reference genomes from the public archives (whose use will be expended within the next year to cover invertebrate metazoa and plant species), and for the automatic generation of tracks from alignment data, have already enabled a massive increase in the quantity of data contained within the resource. A priority for the near-future is the establishment of a pipeline to allow for the automatic representation of any variant call data represented in the European Variation Archive (http://www.ebi.ac.uk/eva) within the framework of reference annotation/interpretation through Ensembl interfaces. This model is dependent, of course, on data producers continuing to subscribe to long-established norms about submitting assembly and annotation data to the INSDC databases, and other data types to appropriate broad-scope repositories. If data is archived in universal archives, it becomes easier for resources such as Ensembl Genomes to integrate and interpret them; the more dispersed data is, the higher the overheads of re-use. In our opinion, it is important that the norms of archival submission are maintained, and we try to practice what we preach: when Ensembl Genomes generates alignment data, these are submitted back into the ENA and advertised through the Track Hub registry, and thus made available in any compliant browser outside of the Ensembl infrastructure. A culture of data sharing improves all resources, and thereby empowers researchers.

The third challenge, in an environment of data plenitude, is to allow users to discover and select data of interest to visualise or analyse. The grouping of tracks into track hubs, and the provision of interfaces by which hubs can be discovered according to their metadata and selectively imported into the Ensembl framework, is a scalable model for data discovery and selection.

The usefulness of this model is critically dependent on the quality of metadata with which the data has been annotated, including the correct identification of the species and strain to which the data set belongs, and descriptions of the aims of the overall experiment and the differences between individual tracks. However, there are a number of obstacles to the acquisition of such metadata: experiments are diverse and designing standards for describing them are consequently difficult; retro-fitting meta data to independently submitted archival submissions is an innately costly process; the most scalable solutions therefore require that data is annotated with metadata prior to submission to the public archives, but data generators may be poorly incentivised to do so, inexpert in the relevant data standards, and actively hostile to being asked to supply the same information more than once. Finding a solution to these problems requires community acceptance of appropriate standards, the development of helpful tools for data validation and submission, and the automatic re-use of metadata between different resources. We are currently working on a project to further develop existing metadata standards (31) for the plant domain, and to capture submitted metadata to link information in Ensembl Genomes (for example, genotype data for individual crop cultivars) to external repositories holding phenotypic data and/or physical stocks. The BioSamples database (32) will be used to connect different repositories containing data derived from related materials. A similar approach is likely needed across the taxonomic space to ensure that specific archived data can be discovered, visualised and used in Ensembl tools and elsewhere.

FUNDING

UK Biosciences and Biotechnology Research Council [BB/J00328X/1, BB/K020080/1, BB/KK020102/1, BB/L024071/1, BB/M018458/1, BB/P016855/1 to P.K.; BB/J/004383/1, BB/I000488/1, BB/K020056/1, BB/K020056/1 and and BB/P016855/1 to K.H.-K.]; UK Medical Research Council [MR/L001020/1 to P.K.]; Wellcome Trust [to P.K.]; U.S. National Science Foundation [41686 IPGA Gramene to D.W.]; 7th Framework Programme [284496, transPLANT to P.K.]; Horizon20–20 Programme [654248, 731060 to P.K.] of the European Union. Funding for open access charge: Research Councils UK; The European Molecular Biology Laboratory.
Conflict of interest statement. None declared.

REFERENCES

1. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
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. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
4. Mashima,J., Kodama,Y., Fujisawa,T., Katayama,T., Okuda,Y., Kaminuma,E., Ogasawara,O., Okubo,K., Nakamura,Y. and Takagi,T. (2017) DNA Data Bank of Japan. *Nucleic Acids Res.*, **45**, D25–D31.
5. UniProt: the universal protein knowledgebase (2017) *Nucleic Acids Res.*, **45**, D158–D169.

6. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
7. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
8. Tello-Ruiz,M.K., Stein,J., Wei,S., Preece,J., Olson,A., Naithani,S., Amarasinghe,V., Dharmawardhana,P., Jiao,Y., Mulvaney,J. *et al.* (2016) Gramene 2016: comparative plant genomics and pathway resources. *Nucleic Acids Res.*, **44**, D1133–D1140.
9. Pedro,H., Maheswari,U., Urban,M., Irvine,A.G., Cuzick,A., McDowall,M.D., Staines,D.M., Kulesha,E., Hammond-Kosack,K.E. and Kersey,P.J. (2016) PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Res.*, **44**, D688–D693.
10. Giraldo-Calderón,G.I., Emrich,S.J., MacCallum,R.M., Maslen,G., Dialynas,E., Topalis,P., Ho,N., Gesing,S. and VectorBase ConsortiumVectorBase Consortium and Madey,G. *et al.* (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.*, **43**, D707–D713.
11. Howe,K.L., Bolt,B.J., Cain,S., Chan,J., Chen,W.J., Davis,P., Done,J., Down,T., Gao,S., Grove,C. *et al.* (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.*, **44**, D774–D780.
12. Bursteinas,B., Britto,R., Bely,B., Auchincloss,A., Rivoire,C., Redaschi,N., O'Donovan,C. and Martin,M.J. (2016) Minimizing proteome redundancy in the UniProt Knowledgebase. *Database (Oxford)*, **2016**, baw139.
13. Sadd,B.M., Barribeau,S.M., Bloch,G., de Graaf,D.C., Dearden,P., Elsik,C.G., Gadau,J., Grimmelikhuijzen,C.J., Hasselmann,M., Lozier,J.D. *et al.* (2015) The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.*, **16**, 76.
14. Albertin,C.B., Simakov,O., Mitros,T., Wang,Z.Y., Pungor,J.R., Edsinger-Gonzales,E., Brenner,S., Ragsdale,C.W. and Rokhsar,D.S. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–224.
15. Rider,S.D., Morgan,M.S. and Arlian,L.G. (2015) Draft genome of the scabies mite. *Parasites Vectors*, **8**, 585.
16. Dohm,J.C., Minoche,A.E., Holtgräwe,D., Capella-Gutiérrez,S., Zakrzewski,F., Tafer,H., Rupp,O., Sörensen,T.R., Stracke,R., Reinhardt,R. *et al.* (2014) The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*, **505**, 546–549.
17. Chalhoub,B., Denoeud,F., Liu,S., Parkin,I.A.P., Tang,H., Wang,X., Chiquet,J., Belcram,H., Tong,C., Samans,B. *et al.* (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, **345**, 950–953.
18. Mascher,M., Gundlach,H., Himmelbach,A., Beier,S., Twardziok,S.O., Wicker,T., Radchuk,V., Dockter,C., Hedley,P.E., Russell,J. *et al.* (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.
19. Clavijo,B.J., Venturini,L., Schudoma,C., Accinelli,G.G., Kaithakottil,G., Wright,J., Borrill,P., Kettleborough,G., Heavens,H., Chapman,H. *et al.* (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.*, **27**, 885–896.
20. Lee,E., Helt,G.A., Reese,J.T., Munoz-Torres,M.C., Childers,C.P., Buels,R.M., Stein,L., Holmes,I.H., Elsik,C.G. and Lewis,S.E. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.
21. Van Kan,J.A.L., Stassen,J.H.M., Mosbach,A., Van Der Lee,T.A.J., Faino,L., Farmer,A.D., Papasotiriou,D.G., Zhou,S., Seidl,M.F., Cottam,E. *et al.* (2017) A gapless genome sequence of the fungus Botrytis cinerea. *Mol. Plant Pathol.*, **18**, 75–89.
22. Fritz,M.H.-Y., Leinonen,R., Cochrane,G. and Birney,E. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, **21**, 734–740.
23. 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.* (2013) Track Data Hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, doi:10.1093/bioinformatics/btt637.
24. Accelerated Profile HMM Searches.
25. Finn,R.D., Clements,J., Arndt,W., Miller,B.L., Wheeler,T.J., Schreiber,F., Bateman,A. and Eddy,S.R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
28. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
29. Urban,M., Cuzick,A., Rutherford,K., Irvine,A., Pedro,H., Pant,R., Sadanadan,V., Khamari,L., Billal,S., Mohanty,S. *et al.* (2017) PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Res.*, **45**, D604–D610.
30. Naithani,S., Preece,J., D'Eustachio,P., Gupta,P., Amarasinghe,V., Dharmawardhana,P.D., Wu,G., Fabregat,A., Elser,J.L., Weiser,J. *et al.* (2017) Plant Reactome: a resource for plant pathways and comparative analysis. *Nucleic Acids Res.*, **45**, D1029–D1039.
31. Krajewski,P., Chen,D., Ćwiek,H., van Dijk,A.D.J., Fiorani,F., Kersey,P., Klukas,C., Lange,M., Markiewicz,A., Nap,J.P. *et al.* (2015) Towards recommendations for metadata and data handling in plant phenotyping. *J. Exp. Bot.*, **66**, 5417–5427.
32. Faulconbridge,A., Burdett,T., Brandizi,M., Gostev,M., Pereira,R., Vasant,D., Sarkans,U., Brazma,A. and Parkinson,H. (2014) Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res.*, **42**, D50–D52.