Ensembl Genomes 2016: more genomes, more complexity

D574–D580 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
doi: 10.1093/nar/gkv1209

Ensembl Genomes 2016: more genomes, more complexity

Paul Julian Kersey[1,*], James E. Allen[1], Irina Armean[1], Sanjay Boddu[1], Bruce J. Bolt[1], Denise Carvalho-Silva[1], Mikkel Christensen[1], Paul Davis[1], Lee J. Falin[1], Christoph Grabmueller[1], Jay Humphrey[1], Arnaud Kerhornou[1], Julia Khobova[1], Naveen K. Aranganathan[1], Nicholas Langridge[1], Ernesto Lowy[1], Mark D. McDowall[1], Uma Maheswari[1], Michael Nuhn[1], Chuang Kee Ong[1], Bert Overduin[1], Michael Paulini[1], Helder Pedro[1], Emily Perry[1], Giulietta Spudich[1], Electra Tapanari[1], Brandon Walts[1], Gareth Williams[1], Marcela Tello–Ruiz[1], Joshua Stein[2], Sharon Wei[2], Doreen Ware[2,3], Daniel M. Bolser[1], Kevin L. Howe[1], Eugene Kulesha[1], Daniel Lawson[1], Gareth Maslen[1] and Daniel M. Staines[1]

[1]The European Molecular Biology Laboratory, The European Bioinformatics Institute, The Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, [2]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA and [3]USDA-ARS NAA Plant, Soil and Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY 14853, USA

Received October 8, 2015; Accepted October 27, 2015

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 context of 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 reference sequence, gene models, transcriptional data, genetic variation and comparative analysis. This paper provides an update to the previous publications about the resource, with a focus on recent developments. These include the development of new analyses and views to represent polyploid genomes (of which bread wheat is the primary exemplar); and the continued up-scaling of the resource, which now includes over 23 000 bacterial genomes, 400 fungal genomes and 100 protist genomes, in addition to 55 genomes from invertebrate metazoa and 39 genomes from plants. This dramatic increase in the number of included genomes is one part of a broader effort to automate the integration of archival data (genome sequence, but also associated RNA sequence data and variant calls) within the context of reference genomes and make it available through the Ensembl user interfaces.

OVERVIEW AND ACCESS

Ensembl Genomes (http://www.ensemblgenomes.org) is organized 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 (http://www.ensembl.org) (1); Ensembl Genomes provides a complementary set of interfaces for non-vertebrate species. Core data available for all species includes genome sequence and annotations of protein-coding and non-coding genes; additional data includes transcriptional data, genetic variation and comparative analysis. Interactive access 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. sequences and external sequence features mapped onto the genome) supporting the

*To whom correspondence should be addressed. Tel: +44 0 1223 494601; Fax: +44 0 1223 494468; Email: pkersey@ebi.ac.uk
Present addresses:
L. Falin, Department of Computer Science & Electrical Engineering, Brigham Young University, ID, USA.
A. Kerhornou, Swiss Institute of Bioinformatics, CMU, Geneva 1211, Switzerland.
B. Overduin, Edinburgh Genomics, The University of Edinburgh, Edinburgh EH9 3FL,UK.

primary annotations. Functional information is provided through direct curation, import from the UniProt Knowledgebase (2) or imputation from protein sequence (using the classification tool InterProScan (3)). We provide much of the data available on each page in a variety of formats for download, and tools that process and visualize various types of user-generated data in the context of the reference sequence and annotation. DNA and protein-based sequence search are also available. Fully referenced documentation of the analytical approaches taken is available online, and an online helpdesk (helpdesk@ensemblgenomes.org) provides a rapid response to users' questions.

The data are stored in a set of MySQL databases using the same schemas as those in use for the Ensembl project. Direct access to these is provided through 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. Database dumps and common datasets (e.g. DNA, RNA and protein sequence sets, and sequence alignments) can be directly downloaded in bulk via FTP (ftp://ftp.ensemblgenomes.org). Ensembl source code is available from GitHub (https://github.com/Ensembl) under an open-source licence.

Ensembl Genomes data is also made available through a series of data warehouses, optimized around common (gene- and variant-centric) queries, using the BioMart data warehousing system (4). The BioMart framework provides a series of interfaces, including web-based query building tools, accessible at each of the Ensembl Genomes eukaryotic portals and a variety of other interfaces for interactive and programmatic access. BioMarts are not currently available for Ensembl Bacteria.

Ensembl Genomes is released 4–5 times a year, in synchrony with releases of Ensembl, utilizing the same software as the corresponding Ensembl release. The overall suite of Ensembl Genomes interfaces mirrors the interfaces provided for vertebrate genomes in Ensembl, and allows users access to genomic data from across the tree of life in a consistent manner.

INVERTEBRATES AND PLANTS

Ensembl Genomes has continued to grow in 2014 and 2015. In the last two years, six species have been added to Ensembl Metazoa, bringing the total number of species included to 55, and 11 species to Ensembl Plants, bringing the total number of included species to 39. The new invertebrate species are the mountain pine beetle (*Dendroctonus ponderosae*) (5), the Glanville fritillary butterfly (*Melitaea cinxia*) (6), the warty comb jelly (*Mnemiopsis leidyi*) (7), a parasitic nematode (*Onchocerca volvulus,* the causative agent of river blindness), the red fire ant (*Solenopsis invicta*) (8) and the Nevada dampwood termite (*Zootermopsis nevadensis*) (9). The new plant species comprise a primitive flowering shrub (*Amborella trichopda*) (10), cabbage (*Brassica oleracea*) (11), cocoa (*Theobroma cacao*) (12), a wild grass (*Leersia perrieri*) (13), six species of rice (*Oryza barthii, Oryza glumaepatula*, *Oryza meridionalis*, *Oryza nivara*, *Oryza punctata* and *Oryza rufipogon*) (13) and bread wheat (*Triticum aestivum*) (14–16), bringing the total number of species represented to 39.

In addition, an ongoing process of data update continues for all genomes included in the database. In the same period, eight metazoan and eight plant genome assembly updates have occurred; and additionally, 14 new metazoan gene sets and two new plant gene sets have been released, annotated on existing assemblies. The plant databases are maintained jointly with the Gramene resource (http://www.gramene.org) (17) and can be accessed from either site.

COMRPEHENSIVE COVERAGE OF MICRO-ORGANISMS

In Ensembl Genomes, genome sequence and annotation are taken directly from experts or databases recognized as authorities in their communities, where such resources exist; otherwise, the raw data is imported from the appropriate sequence archives and derived data are calculated as part of the Ensembl Genomes release process. Revisions to genome assemblies require re-alignment of any sequences that have been assigned a location on the genome by computation and a re-call of features (gene calls, variant calls, synteny blocks) that have been derived from such alignments. Updates to gene models require the assignment of new functional annotation even when the underlying assembly has not changed; and moreover, changes in just one species require the recalculation of all downstream comparative analyses. Genomes that are major foci of scientific research have mostly already been included in the resource; but genome sequence is increasingly available for a much larger number of species of interest to only small research communities, or which are of interest primarily in the context of comparative analysis. However, the cost (in terms of human and computer time) of importing, updating and calculating derived data (especially comparative data) has hitherto limited our the rate of growth of the resource and our ability to serve smaller communities.

Previously, we reported (18) the introduction of a new procedure to automatically update Ensembl Bacteria with all annotated genome sequence present in the archives of the International Nucleotide Sequence Database Collaboration (19). In addition to the data imported, basic functional annotation is added and a selection of species included in a broad-range comparative analysis. Since this report, we have continued to operate this pipeline and the number of bacterial species represented in the database has increased from ∼9000 to over 29 000. The same pipeline has since been applied to Ensembl Fungi and Ensembl Protists, increasing the number of represented fungal genomes to 408 (an eight-fold increase over the number previously included) and protist genomes to 133 (a fourfold increase). Associated revisions to the Ensembl interfaces and API have been introduced to support navigation and selection of genomes (following the model previously established for Ensembl Bacteria). With each release, gene models are automatically updated with new functional annotation: protein domains and gene functions defined using InterProScan (3) and the Gene Ontology (20). Additionally, one representative genome from every species (i.e. 273 fungal genomes and 89 protist genomes) are included in a comparative analysis

(the Compara Gene Tree analysis (21)) with other genomes from the same kingdom. This generates evolutionary histories of every gene family and infers true orthologues by reconciliation with the species tree, and is updated with each release as new data becomes available.

ALIGNMENTS AND VARIANTS

Closely related eukaryotic species are identified as suitable subjects for pairwise whole genome alignment, normally carried out using the lastZ (22) alignment tool followed by chaining and netting (23). For some groups of species, a well annotated genome is used as the point of reference for related species; in other cases, particularly where the genome is smaller, an all-versus-all approach is used. The number of pairwise alignments present in the database has increased to 1205 over the past two years. In addition, variation data are available for 24 species: new data incorporated since 2013 includes data from a new SNP-chip recently developed for the mosquito *Aedes aegytpi* (24), various datasets for barley (25–27) and wheat (see below), and resequencing data from 84 varieties of the tomato *Solanum lycopersicum* (28). Finally, alignments of gene expression (EST and RNA-seq) data are available for a total of 82 species. Users can additionally upload any positional data of their own or visualize data held locally in most common file formats (BAM, VCF, GFF, (Big)Wig, (Big)BED, etc.).

FROM DIPLOIDY TO POLYPLOIDY

The recent release of genome sequence for the hexaploid bread wheat *Triticum aestivum* has been accommodated in Ensembl Plants with extensions to the analysis pipelines and user interfaces presented. The bread wheat genome is over five times larger than a human genome and while the best genome assembly is still fragmented, rapid incremental improvements have been released in recent years: Ensembl Plants has successively incorporated the data of Brenchley *el al*. (29); the International Wheat Genome Sequence Consortium's Chromosome Survey Sequence (14); and currently, an improved version of the latter enhanced by improved genetic mapping data (15) and a higher-quality assembly of the 3B chromosome (16). The large size of the wheat genome is partly due to its allohexaploidy, as it comprises of diploid genomes derived from three closely-related precursor species. Genome assemblies for two of these three precursors (*Triticum uratu*, the precursor of the bread wheat 'A' genome and *Aegilops tauschii,* the precursor of the 'D' genome), are also available in Ensembl Plants (the closest ancestor of the 'B' genome has not yet been unambiguously determined). To present the hexaploid in Ensembl Plants, alignments of the A, B and D genomes against each other have been generated, and which can be visualized in a pre-configured view. Additionally, in the gene tree analysis, the three wheat genomes are treated as separate species, allowing the evolutionary relationship of the genes from the different component genomes to be determined. The gene tree view has been linked into the genome alignment view via a new page, specifically presenting the 'homoeologues' (orthologues within the same species; see Figure 1) and the supporting evidence for the assessment (see Figure 2). Whole genome alignments between the bread wheat genomes and their diploid precursors, and also alignments to the genomes of other related species such as barely, *Brachypodium distachyon*, and rice, are also available.

Bread wheat belongs to the *Pooideae* subfamily of the *Poaceae* (the true grasses) and many important crop plants belong to this particular section of the taxonomy, which has evolved over a relatively short interval of 4 million years. We have prioritized *Pooideae* data for inclusion and the sub-family is now represented in the database by 21 distinct genomes (counting the A, B and D genomes of the hexaploid bread wheat separately), all which are included in the gene tree analysis for plants. Even though some of the presently available assemblies are still in a highly fragmented state, the assembly and annotation of the coding regions is reasonably complete and consistent; and a total of 918 gene families have been computationally identified with a single orthologue in every species and an inferred gene history exactly conformant with the taxonomy (Figure 1B). As more genomes are sequenced, and as the quality of available genome assemblies improves, it is to be expected that gene trees will offer increasingly accurate and comprehensive representations of evolutionary history, and that departures from the taxonomy will likely represent actual gene duplication or loss events and not artifacts of misannotation or misassembly.

The identification of homoeologues has in turn allowed the identification of inter-homoeologous variants—single nucleotide (and larger) variations between the A, B and D genomes. These are not necessarily polymorphisms as they may have become fixed since the ancestor species diverged. These data have been identified from the whole genome alignments in regions of 1:1 homoeology, and can be visualized alongside the inter-varietal polymorphisms also contained in the resource, which are imported from CerealsDB (30) and the wheat HapMap project (31).

Although bread wheat is the first polyploid species in Ensembl, common crop varieties of the *Brassica* genus are similarly allotetraploid, and two diploid precursors of the tetraploid species are already included in Ensembl Plants. It is therefore likely that the data structure and visualization interface developed for wheat will be deployed for further species in the near future.

COMMUNITY AND COLLABORATION

Direct data curation by the scientific community has several potential benefits: people are likely to volunteer where the data is relevant to their own speciality, and thus in areas where their expertise is high and a research programme is active. Ensembl Genomes is working to encourage community-led curation in the context of our partnerships with WormBase (32), VectorBase (33), PhytoPath (Pedro *et al.*, in press) /PHI-base (34) and PomBase (35), providing tools such as Web Apollo (36) and Canto (37) to allow the remote submission of structural and functional annotation. Through these collaborations, we have accommodated substantial community annotation of gene models for the parasitic worm *Brugia malayi*, seven species of invertebrate vectors and are currently collecting data from three fungal species; while community-derived functional anno-

Figure 1. Comparative genomics of bread wheat, as visualized in Ensembl Plants. Panel **A** shows the alignments of two homoeologous genes at the level of protein sequence. The selected gene is highlighted in red. Panel **B** shows these genes in the wider context of a gene tree, showing 1:1 orthology over 21 grass genomes including the 3 bread wheat genomes and the two sequenced diploid precursors.

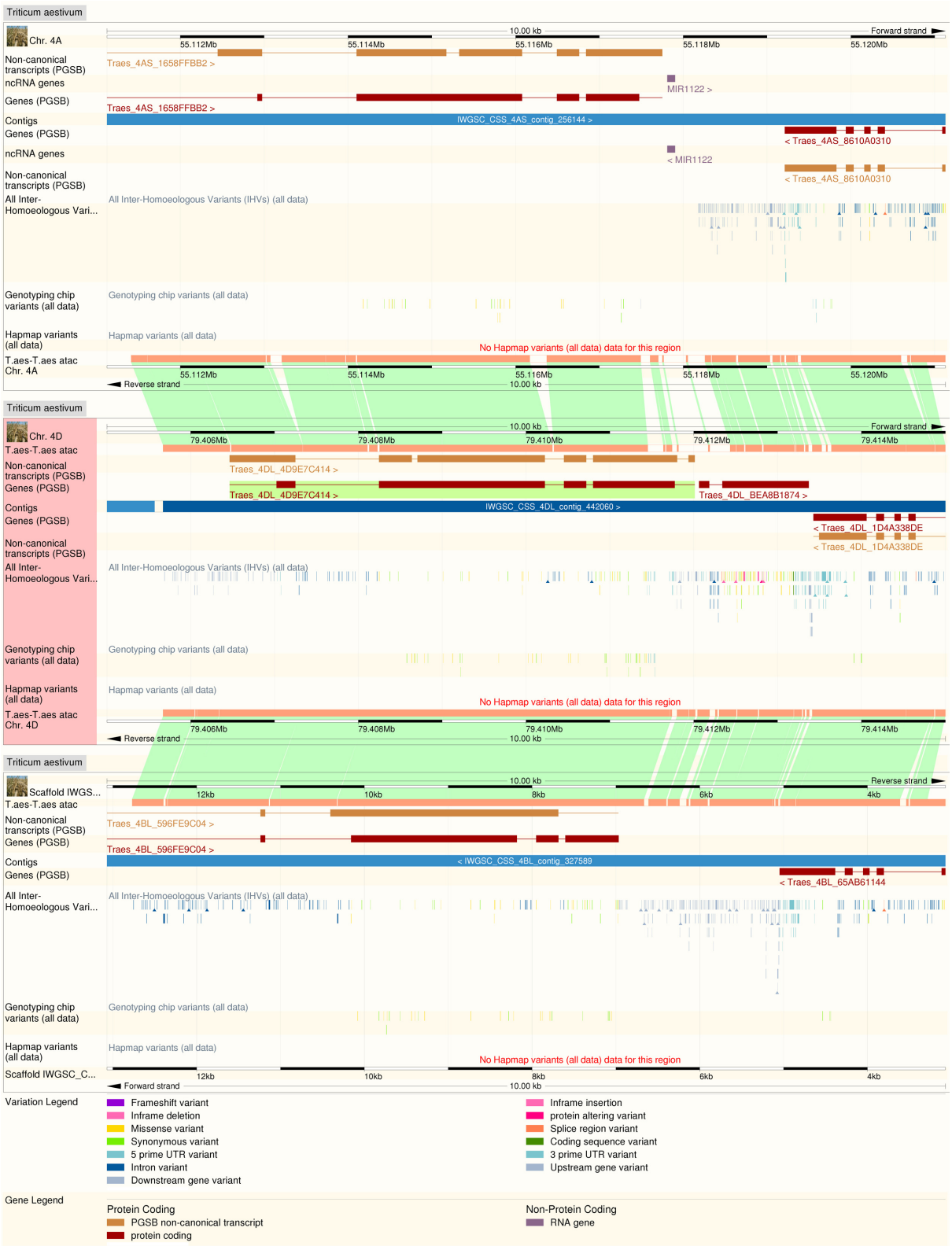

Figure 2. Whole genome alignment between the three bread wheat component genomes at a set of homoeologous loci. Inter-homoeologous variant calls and inter-variety polymorphisms are visible on tracks on each genome.

tations have been collected for *Schizosaccharomyces pombe* and numerous fungal phytopathogen species. New gene models curated through Web Apollo can be immediately visualized as a track in the Ensembl Genomes browsers, and are incrementally imported into the primary gene set. An automatic quality control process is applied which compares new community-supplied annotations to their predecessors, after which they are either accepted or (in the case of major discordance) sent for prior manual inspection before incorporation. The procedure also allows for the re-application of earlier manual curation following automatic re-annotation, ensuring that expert-supplied knowledge is not lost following subsequent analyses.

Collaboration with WormBase has also resulted in a new sister project, WormBase ParaSite (http://parasite.wormbase.org/) which provides access to 99 genomes from parasitic helminths through a compatible set of Ensembl interfaces (including web browser, BioMart and RESTful API). This model, of specialized sites with a focus on specific domains linked to Ensembl and Ensembl Genomes through integrated search and comparative genomics, is likely to become more common as certain domains of life are subject to increasingly comprehensive sequencing.

FUTURE PERSPECTIVES: AUTOMATED ACCESS TO ARCHIVAL DATA

With the increasing volumes of available data in future, it is unlikely that most genome sequences will be subject to manual curation or quality control. For genomes where an insufficiently large community exists to sustain such activities, the Ensembl framework can still play a useful role, organizing both primary data and derived annotations through a standard interfaces in the context of reference genome sequence. A large amount of such data (reads, alignments, feature calls) has already been manually identified and made visible through Ensembl Genomes; but we are developing new pipelines to automatically identify (and, where necessary, align) RNA-seq and variant call data from the relevant archives (e.g. European Nucleotide Archive (38), European Variant Archive (http://www.ebi.ac.uk/eva)) and make these automatically accessible through Ensembl. Doing this successfully will require standards and support for the submission of appropriate meta data (sample and experimental descriptions) and the development of new interfaces within Ensembl to help users identify and select data for inclusion (based on the meta data attached). It is likely that track hubs (39) (a data format proposed by the UCSC Genome Browser and now implemented in Ensembl) will be used as the vehicle to deliver (potentially complex) data sets into the browser on demand; programatic retrieval of specified data sets will also be of growing importance as the number of genomes and alignments grow

ACKNOWLEDGEMENT

We would also like to acknowledge Andrew Yates for reading the manuscript; the contributions of our numerous collaborators; and of all colleagues working on the Ensembl project.

FUNDING

UK Biosciences and Biotechnology Research Council [BB/H531519/1, BB/F19793/1, BB/J017299/1, BB/J00328X/1, BB/I008071/1, BB/KK020102/1, BB/M018458/1 to P.K.]; Wellcome Trust [090548/B/09/Z to P.K.]; Bill and Melinda Gates Foundation [OPPGD1491 to P.K.]; U.S. National Science Foundation [41686 IPGA Gramene to D.W.]; 7th Framework Programme of the European Union [228421, INFRAVEC; 222886-2, Microme; 284496, transPLANT, 42660, AllBio to P.K.]. Funding for open access charge: 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. The UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucl. Acids Res.*, **43**, D204–D212.
3. Mitchell,A., Chang,H.-Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucliec Acids Res.*, **43**, D213–D221.
4. 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.
5. Keeling,C.I., Yuen,M.M., Liao,N.Y., Docking,T.R., Chan,S.K., Taylor,G.A., Palmquist,D.L., Jackman,S.D., Nguyen,A., Li,M. *et al.* (2013) Draft genome of the mountain pine beetle,Dendroctonus ponderosae Hopkins, a major forest pest. *Genome Biology*, **14**, R27.
6. Ahola,V., Lehtonen,R., Somervuo,P., Salmela,L., Koskinen,P., Rastas,P., Välimäki,N., Paulin,L., Kvist,J., Wahlberg,N. *et al.* (2014) The Glanville fritillary genomeretains an ancient karyotype and reveals selective chromosomal fusions inLepidoptera. *Nat. Commun.*, **5**, 4737.
7. Ryan,J.F., Pang,K., Schnitzler,C.E., Nguyen,A.-D., Moreland,R.T., Simmons,D.K., Koch,B.J., Francis,W.R., Havlak,P., NISC Comparative Sequencing Program *et al.* (2013) The genome of thectenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science*, **342**, 1242592.
8. Wurm,Y., Wang,J., Riba-Grognuz,O., Corona,M., Nygaard,S., Hunt,B.G., Ingram,K.K., Falquet,L., Nipitwattanaphon,M., Gotzek,D. *et al.* (2011) The genome of the fire ant Solenopsis invicta. *Proc.Natl. Acad. Sci. U.S.A.*, **108**, 5679–5684.
9. Terrapon,N., Li,C., Robertson,H.M., Ji,L., Meng,X., Booth,W., Chen,Z., Childers,C.P., Glastad,K.M., Gokhale,K. *et al.* (2014) Molecular traces of alternative social organization in a termite genome. *Nat. Commun.*, **5**, 3636.
10. Project,A.G., Albert,V.A., Barbazuk,W.B., de Pamphilis,C.W., Der,J.P., Leebens-Mack,J., Ma,H., Palmer,J.D., Rounsley,S., Sankoff,D. *et al.* (2013) The Amborella Genome and the Evolution ofFlowering Plants. *Science*, **342**, 1241089.
11. Liu,S., Liu,Y., Yang,X., Tong,C., Edwards,D., Parkin,I.A.P., Zhao,M., Ma,J., Yu,J., Huang,S. *et al.* (2014) TheBrassica oleracea genome reveals the asymmetrical evolution of polyploidgenomes. *Nat. Commun.*, **5**, 3930.
12. Argout,X., Salse,J., Aury,J.-M., Guiltinan,M.J., Droc,G., Gouzy,J., Allegre,M., Chaparro,C., Legavre,T., Maximova,S.N. *et al.* (2011) The genome of Theobroma cacao. *Nat. Genet.*, **43**, 101–108.
13. Jacquemin,J., Bhatia,D., Singh,K., Wing,R.A. *et al.* (2013) The International Oryza Map Alignment Project: development of agenus-wide comparative genomics platform to help solve the 9 billion-peoplequestion. *Current Opinion in Plant Biology*, **16**, 147–156.
14. The International Wheat Genome Sequencing Consortium, Mayer,K.F.X., Rogers,J., Doležel,J., Pozniak,C., Eversole,K., Feuillet,C., Gill,B., Friebe,B., Lukaszewski,A.J. *et al.* (2014) A

chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**, 1251788.
15. Chapman,J.A., Mascher,M., Buluç,A., Barry,K., Georganas,E., Session,A., Strnadova,V., Jenkins,J., Sehgal,S., Oliker,L. *et al.* (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.*, **16**, 26.
16. Choulet,F., Alberti,A., Theil,S., Glover,N., Barbe,V., Daron,J., Pingault,L., Sourdille,P., Couloux,A., Paux,E. *et al.* (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**, 1249721.
17. Monaco,M.K., Stein,J., Naithani,S., Wei,S., Dharmawardhana,P., Kumari,S., Amarasinghe,V., Youens-Clark,K., Thomason,J., Preece,J. *et al.* (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.*, **42**, D1193–D1199.
18. 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.
19. Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41**, D21–D24.
20. The Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
21. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
22. Harris,R. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.
23. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 11484–11489.
24. Evans,B.R., Gloria-Soria,A., Hou,L., McBride,C., Bonizzoni,M., Zhao,H. and Powell,J.R. (2015) A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, aedes aegypti. *G3 (Bethesda)*, **5**, 711–718.
25. International Barley Genome Sequencing Consortium, Mayer,K.F.X., Waugh,R., Brown,J.W.S., Schulman,A., Langridge,P., Platzer,M., Fincher,G.B., Muehlbauer,G.J., Sato,K. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
26. Mascher,M., Muehlbauer,G.J., Rokhsar,D.S., Chapman,J., Schmutz,J., Barry,K., Muñoz-Amatriaín,M., Close,T.J., Wise,R.P., Schulman,A.H. *et al.* (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.*, **76**, 718–727.
27. Comadran,J., Kilian,B., Russell,J., Ramsay,L., Stein,N., Ganal,M., Shaw,P., Bayer,M., Thomas,W., Marshall,D. *et al.* (2012) Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.*, **44**, 1388–1392.
28. 100 Tomato Genome Sequencing Consortium, Aflitos,S., Schijlen,E., de Jong,H., de Ridder,D., Smit,S., Finkers,R., Wang,J., Zhang,G., Li,N. *et al.* (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.*, **80**, 136–148.
29. Brenchley,R., Spannagl,M., Pfeifer,M., Barker,G.L.A., D'Amore,R., Allen,A.M., McKenzie,N., Kramer,M., Kerhornou,A., Bolser,D. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.
30. Wilkinson,P.A., Winfield,M.O., Barker,G.L., Allen,A.M., Burridge,A., Coghill,J.A. and Edwards,K.J. (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics*, **13**, 1–6.
31. Jordan,K.W., Wang,S., Lun,Y., Gardiner,L.-J., MacLachlan,R., Hucl,P., Wiebe,K., Wong,D., Forrest,K.L., Sharpe,A.G. *et al.* (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.*, **16**, 48.
32. Harris,T.W., Baran,J., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., Done,J., Grove,C., Howe,K. *et al.* (2013) WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, **42**, D789–D793.
33. 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.
34. Urban,M., Pant,R., Raghunath,A., Irvine,A.G., Pedro,H. and Hammond-Kosack,K.E., (2014) The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Res.*, **43**, D645–D655.
35. McDowall,M.D., Harris,M.A., Lock,A., Rutherford,K., Staines,D.M., Bähler,J., Kersey,P.J., Oliver,S.G. and Wood,V. (2014) PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res.*, **43**, D656–D661.
36. 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.
37. Rutherford,K.M., Harris,M.A., Lock,A., Oliver,S.G. and Wood,V. (2014) Canto: an online tool for community literature curation. *Bioinformatics*, **30**, 1791–1792.
38. Silvester,N., Alako,B., Amid,C., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Ten Hoopen,P., Kay,S., Leinonen,R. *et al.* (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.*, **43**, D23–D29.
39. 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*, **30**, 1003–1005.