D84–D90 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 15 November 2011*
doi:10.1093/nar/gkr991

Ensembl 2012

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Susan Fairley[2], Stephen Fitzgerald[1], Laurent Gil[1], Leo Gordon[1], Maurice Hendrix[2], Thibaut Hourlier[2], Nathan Johnson[1], Andreas K. Kähäri[1], Damian Keefe[1], Stephen Keenan[1], Rhoda Kinsella[1], Monika Komorowska[1], Gautier Koscielny[1], Eugene Kulesha[1], Pontus Larsson[1], Ian Longden[1], William McLaren[1], Matthieu Muffato[1], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Harpreet Singh Riat[2], Graham R. S. Ritchie[1], Magali Ruffier[2], Michael Schuster[1], Daniel Sobral[1], Y. Amy Tang[2], Kieron Taylor[1], Stephen Trevanion[2], Jana Vandrovcova[1], Simon White[2], Mark Wilson[2], Steven P. Wilder[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Ian Dunham[1], Richard Durbin[2], Xosé M. Fernández-Suarez[1], Jennifer Harrow[2], Javier Herrero[1], Tim J. P. Hubbard[2], Anne Parker[2], Glenn Proctor[1], Giulietta Spudich[1], Jan Vogel[2], Andy Yates[1], Amonida Zadissa[2] and Stephen M. J. Searle[2]

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

Received October 10, 2011; Accepted October 17, 2011

ABSTRACT

The Ensembl project (http://www.ensembl.org) provides genome resources for chordate genomes with a particular focus on human genome data as well as data for key model organisms such as mouse, rat and zebrafish. Five additional species were added in the last year including gibbon (*Nomascus leucogenys*) and Tasmanian devil (*Sarcophilus harrisii*) bringing the total number of supported species to 61 as of Ensembl release 64 (September 2011). Of these, 55 species appear on the main Ensembl website and six species are provided on the Ensembl preview site (Pre!Ensembl; http://pre.ensembl.org) with preliminary support. The past year has also seen improvements across the project.

INTRODUCTION

The Ensembl project provides a genome browser at http://www.ensembl.org as well as integrated genome resources. The depth of genome information varies across supported species with the most comprehensive information provided for human, mouse, rat and zebrafish, which are also the most highly accessed genomes. For all species on the main site, we provide comprehensive, evidence-based gene annotations and comparative genomics resources including alignments and homology, orthology and paralogy relationships based on Ensembl GeneTrees (1). We integrate these annotations with a large number of external data sources including InterPro (2), UniProt (3) and Pfam (4). Eighteen of our most popular species also include dedicated variation resources (5) derived from dbSNP (6), DGVa (7) and other sources. The Ensembl regulatory build provides regulatory annotation on the human and mouse genomes and incorporates data from the ENCODE (8) and Roadmap Epigenomics Program (9).

In addition to the data available through the Ensembl website, we provide open access to the Ensembl API (10) and all supporting Ensembl databases to enable flexible, programmatic interaction with our data for use in genomic analysis. Data can also be accessed through the Ensembl BioMart (11,12). We support those who use multiple web-based genome bioinformatics sites by providing links to the UCSC Genome Browser (13) and NCBI's MapViewer (14) on all of our LocationView pages. We also support user data upload and visualization using BAM, BigWig, VCF and other common data formats (see http://www.ensembl.org/info/website/upload/index.html for further information and the most current list of supported upload formats).

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

Here we highlight some of Ensembl's new features and new data released in the last year. As with previous updates (15,16), we can only describe a subset of the information provided. Further information is available in the documentation section of the Ensembl website, from the Ensembl blog (http://www.ensembl.info) or by contacting helpdesk@ensembl.org.

RESULTS

Ensembl produces approximately five releases each year. Releases are numbered sequentially (September 2011 was release 64) and include newly supported species, new assemblies and new or updated annotations of already supported species. The Ensembl genome browser, code base and other genomic information described in this report are also updated with each release.

Gene annotation and supported species

Over the past year, five new species have been added to Ensembl. As of release 64, two of these species, white-cheeked gibbon and Tasmanian devil, are fully supported on our main site. Tasmanian devil is noteworthy because it was the second species (after zebrafish) in Ensembl to include RNAseq-based gene models. The Atlantic cod (*Gadus morhua*) was fully annotated and released on our preview site in conjunction with the genome article (17) and will be released on the main site in late 2011 as part of Ensembl release 65. Genome articles have been published for orang-utan (*Pongo abelii*) (18), anole lizard (*Anolis carolinensis*) (19) and tamar wallaby (*Macropus eugenii*) (20). The Nile tilapia (*Oreochromis niloticus*) and domestic ferret (*Mustela putorius furo*) are also available on our preview site. In addition to the new species, we released annotation on assemblies for sea lamprey (*Petromyzon marinus*), Western clawed frog (*Xenopus tropicalis*), cow (*Bos taurus*) and microbat (*Myotis lucifugus*). The microbat assembly has also moved from low to high coverage. An updated chimpanzee assembly, Pan_troglodytes-2.1.3, was released on our Pre! site. All new annotation projects are now released with a genebuild summary document on the species homepage, providing the user with detailed methods.

Gene annotation on our most popular species—human, mouse and zebrafish—has been fully updated this year. These annotations are a merged gene set consisting of the output of the Ensembl evidence-based automatic pipeline (21) and the manual annotation from the Havana project (22). For human, the major update to the gene set was a part of release 62 (April 2011), and the Ensembl/Havana merged gene set continues to be equivalent to the GENCODE gene set, the reference gene set for the ENCODE project (23). The Ensembl RNAseq annotation pipeline, first developed for zebrafish, was used to generate gene models for 16 human tissues using the Illumina Human BodyMap 2.0 data, produced on HiSeq 2000 instruments in 2010. The gene models based on this annotation have been made available to the user community in a dedicated RNAseq database and can also be viewed on the Ensembl website alongside intron supporting features that indicate the number of intron-spanning reads that have mapped to the transcript model. The zebrafish and mouse gene sets were updated, respectively, in release 60 (November 2010) and release 61 (February 2011), with zebrafish moving to the new Zv9 assembly.

The methods used to merge the Ensembl and Havana annotation are updated regularly as the input data evolves, for example genes with biotype 'lincRNA' were introduced into Havana this year. We continue to be part of the Consensus Coding Sequence (CCDS) project (24) for human and mouse, and all current CCDS models are included in our gene sets.

Ensembl release 64 (September 2011) displays the full human GRCh37.p5 assembly produced in June 2011 by the Genome Reference Consortium (25). This patch includes 105 regions of which 40 are of the fix type and the rest are novel patches. By default, the primary assembly (GRCh37 chromosomes 1–22, X, Y) is displayed on the Ensembl website and these sequences are identical to those found on other genome browsers. Users can choose to display alternate loci (haplotypes, fix patches and novel patches) by selecting them in LocationView. We provide annotation on the GRC assembly patches and have developed a dedicated pipeline for this purpose where alignment-based annotation is combined with projected annotation from the primary assembly to provide the most complete annotation coverage of each patch. We are the only genome browser to integrate these alternate sequences into the primary assembly and to allow for visualization of the alternate sequences alongside the surrounding primary assembly (Figure 1).

Comparative genomics

The gene set from each supported species is included in the Ensembl GeneTrees. As new species are added, we can better resolve the phylogenetic history of the genes. For instance, with the addition of the turkey gene set, we can now detect that 15% of the gene duplications that appeared to be specific to the chicken genome happened in the Phasianidae lineage (26).

To cater for the increase in the number of species, the orthologues view now includes a summary table that shows the distribution of one-to-one, one-to-many and many-to-many relationships among species. This table serves both as a summary and as a way to filter the results to, for example, all fishes.

In addition to the five-way fish Enredo-Pecan-Ortheus (EPO), the set of whole-genome multiple alignments provided by Ensembl now includes a three-way avian EPO alignment (27,28) incorporating chicken, turkey and zebra finch. We use each of these alignments and GERP (29) to estimate a per-base sequence conservation score and to annotate regions of evolutionary constraint.

Ancestral alleles for the human genome are inferred from the six-way primate EPO alignments. In the EPO pipeline, Ortheus uses Pecan alignments to reconstruct the per-base history of the sequences. We use the most recent ancestor to call the ancestral allele for the whole human genome. These data are used in the 1000 Genome Project (30) as well as in dbSNP to supplement their

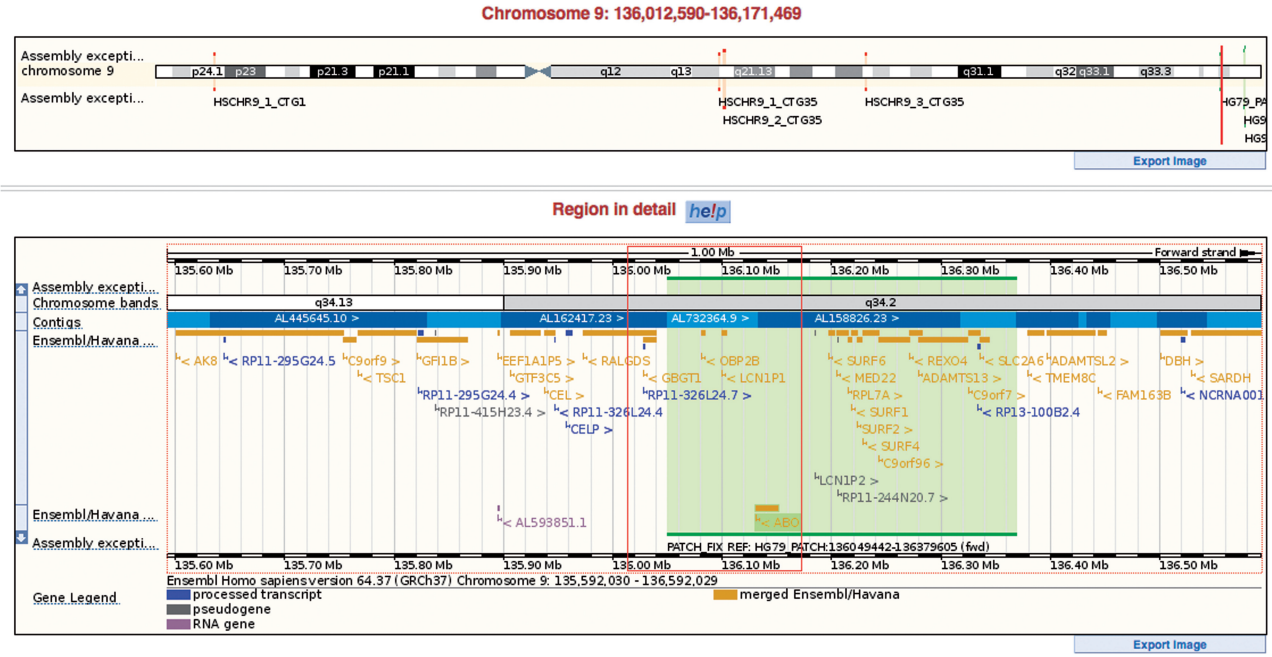

Figure 1. A region of human chromosome 9 from the GRCh37.p5 assembly showing the fix patch applied in the ABO locus and displayed in genomic context (green region on the lower panel). In the upper panel, the full chromosome display shows the locations of a number of other fix and novel patches that are a part of the GRCh37.p5 assembly.

previous ancestral allele calls (31). The same approach was used to study the recent evolution in the primate lineage (32).

Regulation

Over the past year, we have continued to incorporate additional regulatory information into Ensembl from high-throughput sequence assays of chromatin samples. The data are processed by an integrated mapping and processing pipeline using the eHive system (33). As of Ensembl release 64, the Ensembl Regulation database contained 369 ChIP-seq and DNase-seq data sets from 10 human and 5 mouse cell lines. These included the genomic locations of binding regions for 74 different transcription factors (TFs) as well as the locations of sites for 40 modified histones, and an additional 26 data sets that identify regions of open chromatin or DNase I hypersensitivity. Twenty-five of the TFs have binding matrices available through the JASPAR database (34), and we provide the positions of high probability TF-binding sites within the binding regions based upon these matrices. From release 64, we have run separate peak calling methods tuned for either sharp, punctate signals (TFs or punctate histone modification ChIP-seq samples) or broad region signals (e.g. H3K36me3). These data are used as the basis for the Ensembl Regulatory Build that integrates peak calls with histone modifications on a cell-specific basis to provide a regulatory annotation of the genome. In human, the combined regulatory build across 10 cell lines annotates 228 Mb of genome in 442 258 regulatory features. The addition of data for TFs involved in RNA Polymerase 3 (Pol3) gene transcription has allowed broadening of the scope of annotation of regulatory features to include Pol3 gene associated features. Furthermore, we have recently added Distributed Annotation System (DAS) access to reduced representation bisulphite sequencing data generated by the ENCODE project annotating the extent of DNA methylation at hundreds of thousands of CpG dinucleotides across the genome in 44 human cell lines. The Ensembl Regulation database continues to provide mapping of probe sets for all the common microarray platforms including new arrays such as the Illumina Infinium methylation designs.

Over the year, we have made a number of changes to the database and interface to provide improved performance and utility. Browser response time of the 'multi-wiggle' track type displays that were first introduced in 2010 (16) has been considerably advanced by moving the signal data into compressed binary files stored outside of the main database, resulting in more than 50-fold improvement in data load times. Control of the content and display of the signal and peak tracks for ChIP-seq and DNase-seq data in the Location and Regulatory Feature views has been refined by an enhanced matrix style interface that centralizes all display options (Figure 2). From release 64 onwards, we also store explicit links to raw data in the European Nucleotide Archive (ENA) with API support to provide better documentation of experimental samples. Finally, the Ensembl Regulation BioMart has been restructured to improve usability as well as to include more data.

Variation

Ensembl's variation data was updated with the major human data release from dbSNP 132. We also support

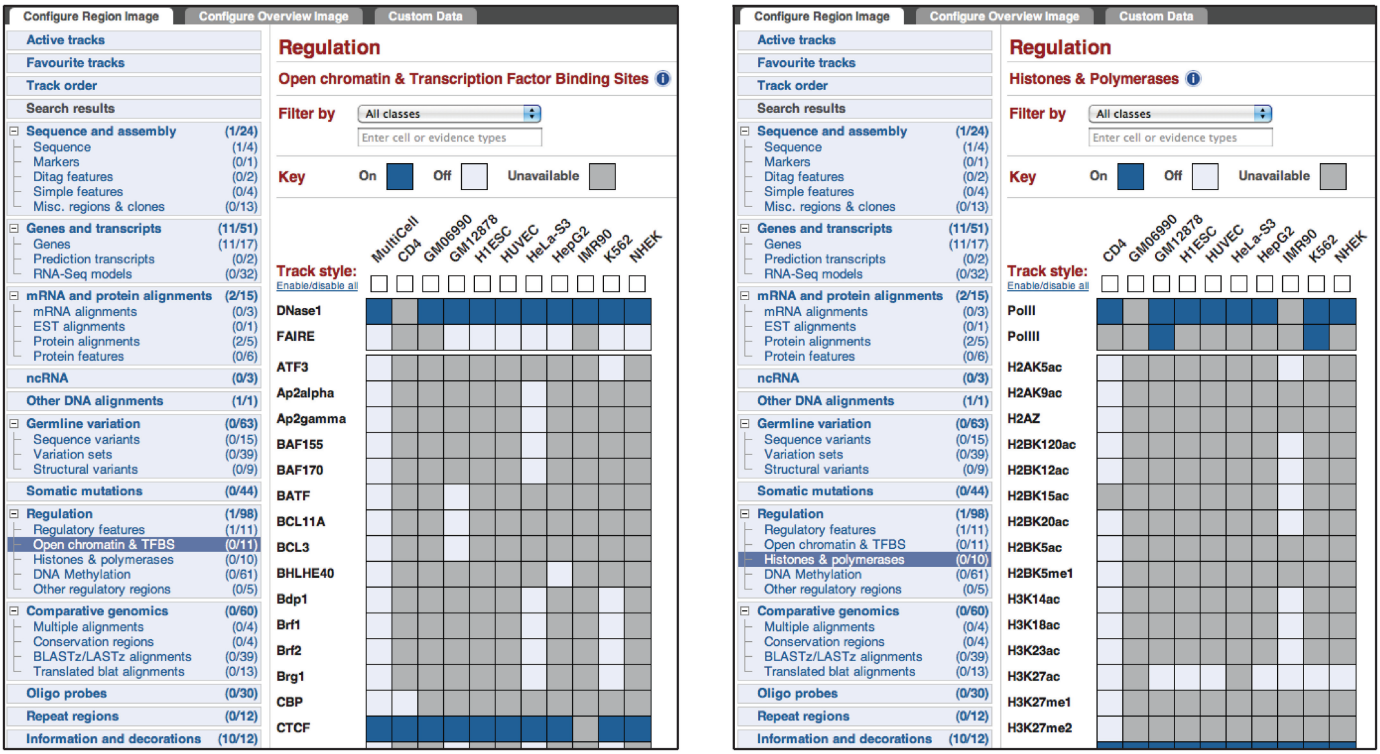

Figure 2. Matrix style configuration panels for Open chromatin & Transcription Factor Binding Sites (on right) and Histones and Polymerases (on left). Both panels are accessible from the relevant portions of the Regulation section of the 'Configure this page' link found on the left menu of all location view pages.

three new species with the most recently released variation data from dbSNP: cat (dbSNP 131), opossum (dbSNP 131) and pig (dbSNP 128). In addition, the variation data for zebrafinch, Tetraodon, horse, rat, zebrafish, cow and mouse were updated in the last year. Structural variation data are imported with each Ensembl release from DGVa (7) and now include the full structural variation data from the 1000 Genomes Pilot Project (30,35) in addition to structural variants from mouse, pig and dog. To facilitate browsing, the website separates data for 1000 Genomes Project variants and variation data with phenotypes in different tracks. New structural variation BioMart data sets for mouse, dog and pig have also been added.

As the amount of variation data has increased, so has the requirement to provide annotation of those variants subject to data problems or other concerns. For all data imported into the variation database, we continually update our methods for conducting quality and sanity checks and those data failing one of these checks are flagged with the reasons for concern. As of release 64, we flag as suspect a variant with any of the following characteristics: no genotypes with no alleles; four or more different alleles; no allele that matches the reference allele; alleles with ambiguity codes; mapped position incompatible with the reported alleles; no associated sequence; genotype frequencies which do not add up to 1; data that either do not map to the genome or have no genotypes; non-nucleotide alleles. Each variant is now checked by ssID so that a single problem submission can be effectively filtered from the database.

We continue to provide extensive data resources for disease and phenotype annotations for both germline variants and somatic mutations. As of release 64 (September 2011), a total of 174 681 distinct variants in Ensembl have phenotype annotation. These data include 43 272 somatic mutations from COSMIC (36) and 61 793 mutations from the public portion of the HGMD database (37). We also have more than 5000 phenotype–variant associations from the NHGRI GWAS catalog (38) and over 14 000 from OMIM.

We have made significant improvements to the variation consequence pipeline, which now supports allele-specific consequences, complete SIFT (39) and PolyPHEN (40) predictions for the human proteome and new regulatory region consequences. All variation consequences are reported with standard sequence ontology terms (41) to ensure consistent definitions across browsers and facilitate comparison with other resources.

The Variant Effect Predictor (42) has also been extended to include these features and as an option, to run as a stand-alone software program that does not require a network connection or Ensembl database. The integration of regulatory features with DNA variation data is particularly powerful. Overlap of variants with regulatory features is reported, as well as overlap with TF-binding sites and high information content bases within the binding site. This provides an additional level of annotation of variants identified in GWAS and other studies.

Ensembl website and software infrastructure

This year saw the release of a third mirror of the Ensembl website in the Asia-Pacific region located at http://asia.ensembl.org. As with our other mirrors at http://useast.ensembl.org and http://uswest.ensembl.org, the Asia mirror uses Amazon Web Services (AWS) to provide the infrastructure (the USWest mirror was migrated to AWS Northern California data centre in 2011). By consolidating all of the supported Ensembl mirrors in AWS, we are able to provide consistent support and increased performance for users around the world. All users visiting the Ensembl website are automatically redirected to their nearest mirror, ensuring the best possible performance. For users accessing Ensembl data via our API or direct MySQL queries, we have also launched a second database server at useastdb.ensembl.org.

The last major update to the Ensembl web interface was release 51 (November 2008) (43,44). Over the past year, we have continued to focus on small but significant improvements to the web interface and on-going updates to the underlying software infrastructure. The latter included the deployment of Lucene, an open-source search engine that we also mirrored in AWS.

Display of the user's own data has been extended to include attachment of large indexed formats such as BAM, BigWig and VCF. For file types such as BED or GFF, the file upload process has been improved to give a count of features parsed from the file and to provide a link to sample coordinates where data can be viewed. Improvements to the discoverability of content have also been made, with a redesigned masthead containing links to popular content and the rewriting of some text-heavy pages to provide clear links enhanced by graphics. When users log in to their personal Ensembl accounts, each tab in the masthead also has dropdown menus giving quick access to other species and recently visited locations, genes, variations and regulatory features.

Finally, design of the pop-up configuration panel has been improved with the aid of user input. Favorite tracks can be saved, and a new graphical matrix allows quick and intuitive selection of cell/tissue data for regulatory tracks.

Ensembl user support

Ensembl continues to support users through our outreach activities, in addition to targeting new users and getting first-hand feedback about our resources from scientists. Our trainers have held workshops in 46 countries, and the 90 workshops given in 2011 included both Japan and Eastern Europe.

Documentation for the Ensembl API was updated this year to properly describe our inheritance model. This required the deployment of a new documentation system based on Doxygen (http://www.doxygen.org). The full API documentation is available at http://www.ensembl.org/info/docs/api/index.html.

This year also included new efforts to reach users though the launch of our Facebook site at www.facebook.com/Ensembl.org, where we post popular weekly navigation tips and other information about the project. Our Facebook page joins our blog, Twitter feed and YouTube channel, which features 12 training videos focused on specific uses of the Ensembl website. Finally, we launched a beginner Ensembl browser course as part of the EBI e-learning platform at http://www.ensembl.info/ecourse. The course allows new users to learn the basic navigation of the browser without having to attend a workshop.

ACKNOWLEDGEMENTS

We thank the members of our Scientific Advisory Board and all of our users. We are especially thankful to those who take the time to contact us through our mailing lists and blog. We acknowledge those researchers, organizations and large-scale projects that have provided data to Ensembl prior to publication under the understandings of the Fort Lauderdale meeting discussing Community Resource Projects and the Toronto meeting on prepublication data sharing (45).

FUNDING

The Wellcome Trust provides majority funding for the Ensembl project (WT062023 and WT079643) with additional funding from the National Human Genome Research Institute (U01HG004695, U54HG004563 and U41HG006104) and the European Molecular Biological Laboratory. Additional support as specified: Funded by the European Commission under SLING, grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7 Capacities Specific Programme; The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 222664 ('Quantomics'). This Publication reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein; The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754—the GEN2PHEN project; The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant agreement n° 223210 CISSTEM. Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
3. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

4. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
5. Chen,Y., Cunningham,F., Rios,D., McLaren,W.M., Smith,J., Pritchard,B., Spudich,G.M., Brent,S., Kulesha,E. *et al.* (2010) Ensembl Variation Resources. *BMC Genomics*, **11**, 293.
6. Foelo,M.L. and Sherry,S.T. (2007) NCBI dbSNP Database: content and searching. In: Weiner,M.P., Gabriel,S.B. and Stephens,J.C. (eds), *Genetic Variation: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, pp. 41–61.
7. Church,D.M., Lappalainen,I., Sneddon,T.P., Hinton,J., Maguire,M., Lopez,J., Garner,J., Paschall,J., Dicuccio,M. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.
8. The ENCODE Project Consortium. (2011) A User's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
9. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
10. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
11. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, doi: 10.1093/database/bar030.
12. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
13. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
14. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
15. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
16. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
17. Star,B., Nederbragt,A.J., Jentoft,S., Grimholt,U., Malmstrøm,M., Gregers,T.F., Rounge,T.B., Paulsen,J., Solbakken,M.H. *et al.* (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**, 207–210.
18. Locke,D.P., Hillier,L.W., Warren,W.C., Worley,K.C., Nazareth,L.V., Muzny,D.M., Yang,S.P., Wang,Z., Chinwalla,A.T. *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.
19. Alföldi,J., Di Palma,F., Grabherr,M., Williams,C., Kong,L., Mauceli,E., Russell,P., Lowe,C.B., Glor,R.E. *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**, 587–591.
20. Renfree,M.B., Papenfuss,A.T., Deakin,J.E., Lindsay,J., Heider,T., Belov,K., Rens,W., Waters,P.D., Pharo,E.A. *et al.* (2011) Genome sequence of an Australian kangaroo, Macropus eugenii, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.*, **12**, R81.
21. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
22. Wilming,L.G., Gilbert,J.G., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
23. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7(Suppl. 1)**, S4.1–S4.9.
24. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E. *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.
25. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
26. Dalloul,R.A., Long,J.A., Zimin,A.V., Aslam,L., Beal,K., Ann Blomberg,L., Bouffard,P., Burt,D.W., Crasta,O. *et al.* (2010) Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.*, **8**, e1000475.
27. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
28. Paten,B., Herrero,J., Fitzgerald,S., Beal,K., Flicek,P., Holmes,I. and Birney,E. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, **18**, 1829–1843.
29. Cooper,G.M., Stone,E.A., Asimenos,G., NISC Comparative Sequencing Program. In Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
30. 1000 Genomes Project Consortium., Durbin,R.M., Abecasis,G.R., Altshuler,D.L., Auton,A., Brooks,L.D., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
31. Spencer,C.C., Deloukas,P., Hunt,S., Mullikin,J., Myers,S., Silverman,B., Donnelly,P., Bentley,D. and McVean,G. (2006) The influence of recombination on human genetic diversity. *PLoS Genet.*, **2**, e148.
32. Katzman,S., Capra,J.A., Haussler,D. and Pollard,K.S. (2011) Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol.*, **3**, 614–626.
33. Severin,J., Beal,K., Vilella,A.J., Fitzgerald,S., Schuster,M., Gordon,L., Ureta-Vidal,A., Flicek,P. and Herrero,J. (2010) eHive: An Artificial Intelligence workflow system for genomic analysis. *BMC Bioinformatics*, **11**, 240.
34. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W., Sandelin,A. *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
35. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
36. Forbes,S.A., Tang,G., Bindal,N., Bamford,S., Dawson,E., Cole,C., Kok,C.Y., Jia,M., Ewing,R. *et al.* (2010) COSMIC (the catalogue of Somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.*, **38**, D652–D657.
37. Cooper,D.N., Stenson,P.D. and Chuzhanova,N.A. (2006) The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr. Protoc. Bioinformatics*, Chapter 1, Unit 1.13.
38. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
39. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
40. 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.

41. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
42. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
43. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
44. Parker,A., Bragin,E., Brent,S., Pritchard,B., Smith,J.A. and Trevanion,S. (2010) Using caching and optimization techniques to improve performance of the Ensembl website. *BMC Bioinformatics*, **11**, 239.
45. Toronto International Data Release Workshop Authors. (2009) Prepublication data sharing. *Nature*, **461**, 168–170.