Database, 2018, 1–12
doi: 10.1093/database/bay119
Database update

Database update

Ensembl variation resources

Sarah E. Hunt*, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M. Armean, Stephen J. Trevanion, Paul Flicek and Fiona Cunningham*

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

*Corresponding author: Tel: +44(0)1223 494517; Fax: +44 (0)1223 494468; Email:seh@ebi.ac.uk

Correspondence may also be addressed to Fiona Cunningham. Tel:+ 44 (0)1223 494612; Fax:+44 (0)1223 494468; Email: fiona@ebi.ac.uk

Citation details: Hunt,S.E., McLaren,W., Gil,L. *et al.* Ensembl variation resources. *Database* (2018) Vol. 2018: article ID bay119; doi:10.1093/database/bay119

Received 31 July 2018; Revised 25 September 2018; Accepted 4 October 2018

Abstract

The major goal of sequencing humans and many other species is to understand the link between genomic variation, phenotype and disease. There are numerous valuable and well-established variation resources, but collating and making sense of non-homogeneous, often large-scale data sets from disparate sources remains a challenge. Without a systematic catalogue of these data and appropriate query and annotation tools, understanding the genome sequence of an individual and assessing their disease risk is impossible. In Ensembl, we substantially solve this problem: we develop methods to facilitate data integration and broad access; aggregate information in a consistent manner and make it available a variety of standard formats, both visually and programmatically; build analysis pipelines to compare variants to comprehensive genomic annotation sets; and make all tools and data publicly available.

Database URL: https://www.ensembl.org

Introduction

One of the most important challenges in contemporary biomedical research is to understand how changes in genome sequences lead to different phenotypes and influence disease. Collecting and interpreting genomic sequence variation is the key to addressing these and other questions. Indeed, recent results have provided insight into disease susceptibility (1), treatment response (2), development of desirable traits in animals (3) and crop plants (4) and understanding population genetics in many species (5). Human genetics study designs in particular are transitioning from targeting a handful of candidate genes to full exome or genome sequencing to find trait-associated loci in individuals, families or populations (6, 7). In all study designs, sites of genome variation alone are insufficient to draw conclusions. To interpret results, it is essential to have access to additional data such as allele frequencies in reference populations, reported disease associations and predicted functional impact on genes (8). To facilitate this important and growing area of research, Ensembl identifies,

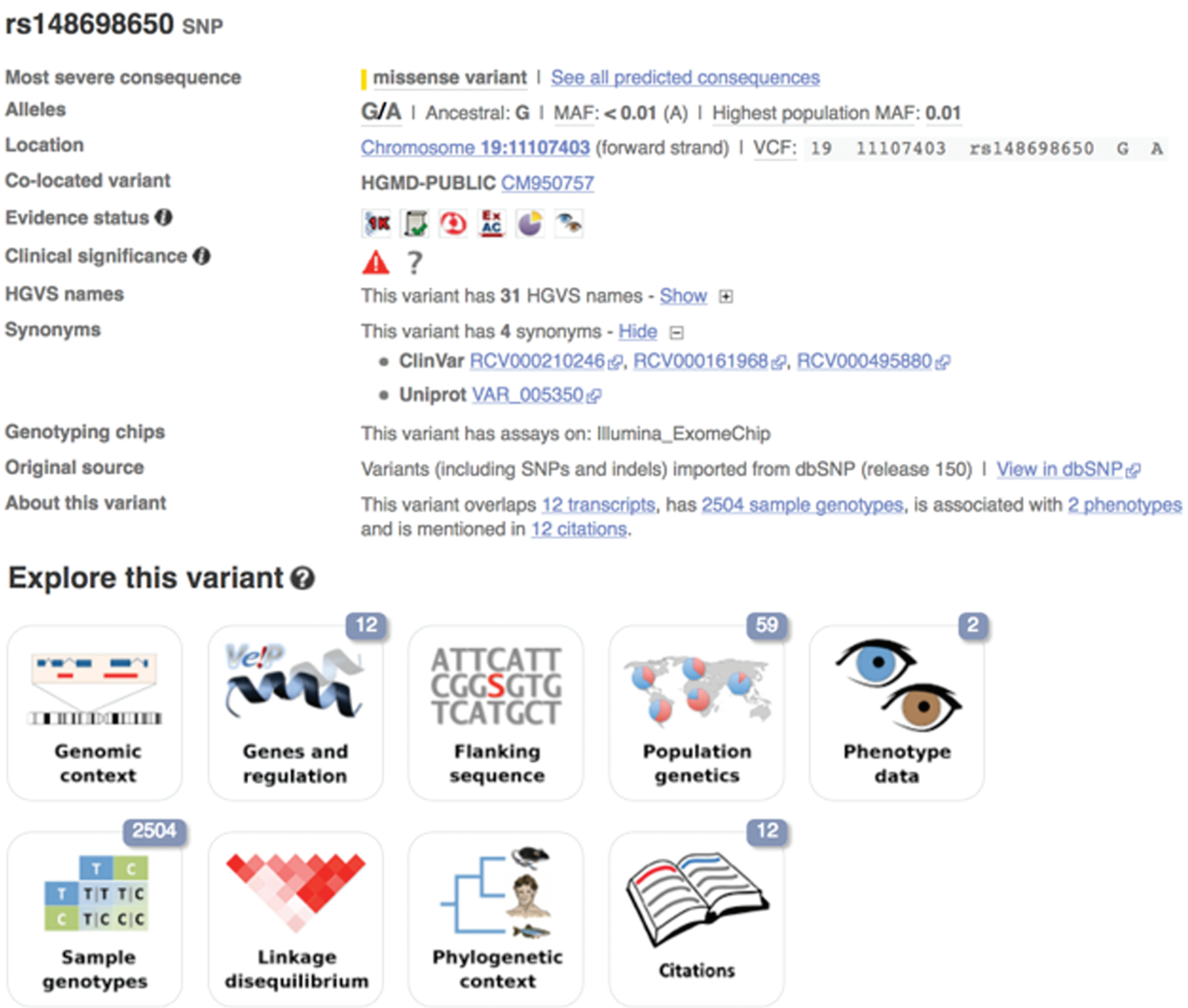


Figure 1. A variant summary view for rs148698650 that displays the global minor allele frequency from the 1000 Genomes Project, other variants at the same location and links to projects providing additional information about the variant. URL: https://www.ensembl.org/Homo_sapiens/Variation/Explore?v=rs148698650

integrates and organizes comparable data from different sources into a system that can be queried and browsed using uniform access and analysis methods.

Here, we describe how the variation and phenotype data sources incorporated into Ensembl are harmonized, visualized and programmatically accessed. We focus on data sources relevant for genome interpretation, including for understanding human phenotypes and disease. Major changes since our last report [9, that described the variation resources present in Ensembl release 56 (September 2009)] include the incorporation of more extensive phenotype and disease annotation in multiple species, the inclusion of variation citation data, improved allele normalisation and equivalence algorithms and infrastructure changes to support the presentation and visualisation of many millions of genotypes.

Data access

The genome browser

For analysis of small areas of the genome, such as variation in a single gene or transcript, visual displays remain the key to explore, analyse and communicate scientific findings. We provide access to the data we incorporate through several different interactive displays via the Ensembl genome browser (www.ensembl.org).

From the Ensembl home page, searching for genomic regions, genes, phenotypes, diseases or variant identifiers provides access to different data views. For example, variant-specific web pages (see Figure 1) are a convenient way to visualize data aggregated from a wide range of different variant-focused projects combined with the results of our analyses. The information is presented

Figure 2. The 'Region in detail' view showing 5 of the available 66 tracks of variant data. URL: https://www.ensembl.org/Homo_sapiens/Location/View?r=10:79069035-79316528

both as a summary and as a set of detailed graphical views and tables. These are described in more detail below.

The Ensembl genome browser also provides displays of genomic regions in which variants are presented alongside transcripts, regulatory features and conservation information. Variants are coloured by the most severe predicted impact they have on gene function. This 'Region in detail' view (see Figure 2) can be configured to display tracks of specific variants, such as those from ClinVar, those with phenotype annotations or those assayed on selected genotyping arrays, such as the Illumina ImmunoChip (10) and the Affymetrix Chicken600K array (11).

Our dedicated gene pages include tables describing all variants in the gene region with many filtering options including by consequence or variant type. These are configurable to improve readability and analysis capabilities, especially in genes with a large quantity of variant data. We also provide configurable phenotype pages that show variants and genes associated with a phenotype, trait or disease. On both pages, data are imported from multiple sources and presented in a single table for easy comparison.

Application programming interfaces

To facilitate direct data access for flexible analysis and to support the web displays described above, we have developed and extended our methods for rapid bulk data retrieval. Custom queries can be written using our mature and comprehensive Perl application programming interface (API), for which extensive documentation and detailed tutorials are available on the Ensembl website. We have also implemented a Representational State Transfer (REST) API (12) for language agnostic data access. Together, these enable convenient integration of our tools and data into multiple other websites and analysis pipelines.

Variant Effect Predictor

The data assembled in the Ensembl databases can be used to annotate variants and predict their functional impact using the Ensembl Variant Effect Predictor (VEP) tool (13). VEP provides a simple, yet powerful, interface using the Ensembl API, which enables the user to annotate an entire human genome (around 4 million variants) in less than an hour and an exome (around 200 000 variants) in less than 5 minutes. It can be accessed via a web interface, a stand-alone script and a REST API. The web interface is integrated with other Ensembl views allowing navigation directly from VEP results to detailed information on any transcripts or to previously known variants that match an input variant. The stand-alone script is highly configurable and can be extended to incorporate additional data from Ensembl (using the Perl API) or other sources. For more detailed information, see (13).

Heterogeneous data integration

Over the past 8 years, we have increased the number of supported species, the range of data types we hold and volume of data we store, analyse and serve. For example, we have extended our schema to include variant data citation data and greatly improved our management of phenotype and disease annotations to incorporate additional data sources.

Short variants

The Ensembl variation databases currently contain publicly available data for 20 vertebrate species from sources including dbSNP (14), the European Variation Archive (EVA) (https://www.ebi.ac.uk/eva/) and ClinVar (15). To allow all information for a locus to be considered together, we collate

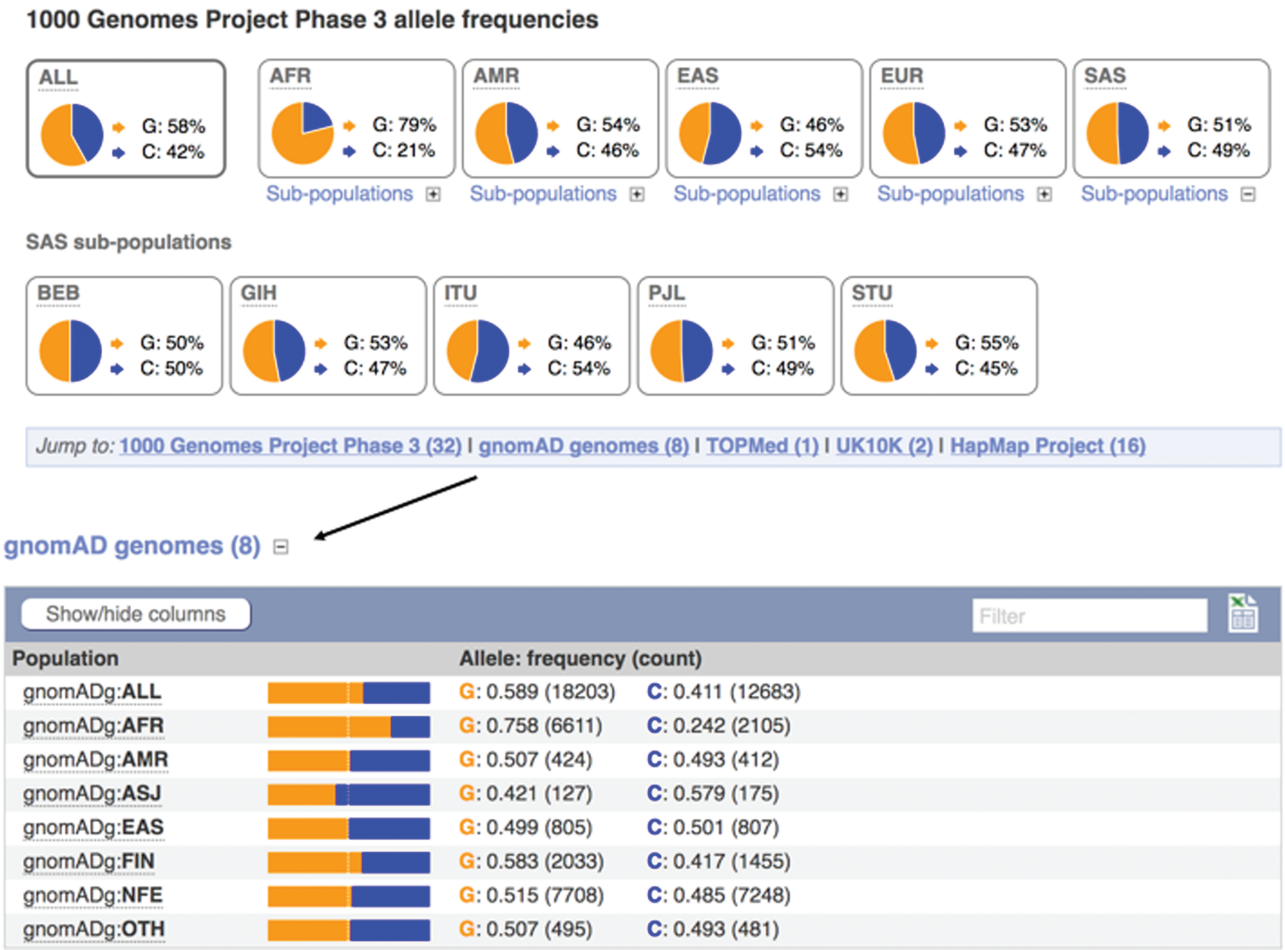


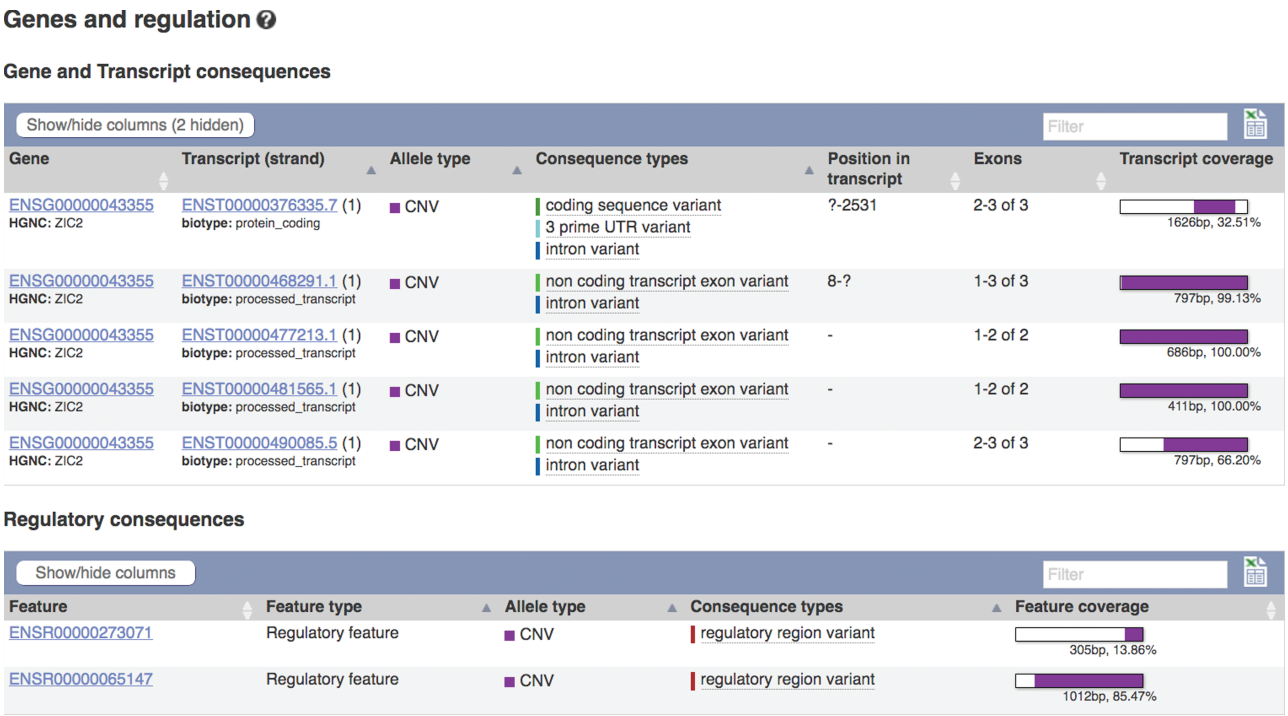
Figure 3. Displays of the frequency spectrum of variant rs1333049 in the 1000 Genomes Project and Genome Aggregation Database panels. URL: https://www.ensembl.org/Homo_sapiens/Variation/Population?v=rs1333049

data where possible and index the merged records with identifiers from multiple resources. This enables searching directly with accession numbers from multiple databases including dbSNP, UniProt (16), PharmGKB (17) and ClinVar.

Since our last report, the number of short variants held in Ensembl databases has risen from 56 million to over 1166 million as of Ensembl release 93 (July 2018). The distribution of data across species has also changed due to major variant discovery efforts in human and livestock species. Genotype and allele frequency data are available for an increasing number of species, from sources such as the Mouse Genomes Project (18) and the sheep and goat focused NextGen project (19).

The vast majority of our variant data are imported from (*The National Center for Biotechnology Information*) NCBI's dbSNP database. As an archive, dbSNP accepts all submitted variants, so entries vary in level of supporting evidence. To identify the most robust data from these multiple submissions, we have extended our quality control process to have two stages. As previously described, variants are considered as 'suspect' if they show certain characteristics, such as a mismatch between the reported variant alleles and the reference genome sequence at the specified location. We now also summarize the evidence supporting each variant such as inclusion in a large-scale reference project. Details of our data and processing procedure are listed at: https://www.ensembl.org/info/genome/variation/index.html.

When seeking to identify somatic mutations or variants involved in rare disease, it is standard practise to filter the variants found in a clinical sample for those already known to be common in the population (8). We therefore provide frequency data from a number of reference sets including the 1000 Genomes Project (20), which sequenced, at low coverage, the full genomes of 2504 individuals from 26 populations and the Genome Aggregation Database (gnomAD) (21), which collected and analysed the exomes of 123 136 individuals, and the full genomes of 15 496 individuals from 7 populations. Our variant 'Population Genet-



Genes and regulation

Gene and Transcript consequences

Show/hide columns (2 hidden) Filter

Gene	Transcript (strand)	Allele type	Consequence types	Position in transcript	Exons	Transcript coverage
ENSG00000043355 HGNC: ZIC2	ENST00000376335.7 (1) biotype: protein_coding	CNV	coding sequence variant 3 prime UTR variant intron variant	?-2531	2-3 of 3	1626bp, 32.51%
ENSG00000043355 HGNC: ZIC2	ENST00000468291.1 (1) biotype: processed_transcript	CNV	non coding transcript exon variant intron variant	8-?	1-3 of 3	797bp, 99.13%
ENSG00000043355 HGNC: ZIC2	ENST00000477213.1 (1) biotype: processed_transcript	CNV	non coding transcript exon variant intron variant	-	1-2 of 2	686bp, 100.00%
ENSG00000043355 HGNC: ZIC2	ENST00000481565.1 (1) biotype: processed_transcript	CNV	non coding transcript exon variant intron variant	-	1-2 of 2	411bp, 100.00%
ENSG00000043355 HGNC: ZIC2	ENST00000490085.5 (1) biotype: processed_transcript	CNV	non coding transcript exon variant intron variant	-	2-3 of 3	797bp, 66.20%

Regulatory consequences

Show/hide columns Filter

Feature	Feature type	Allele type	Consequence types	Feature coverage
ENSR00000273071	Regulatory feature	CNV	regulatory region variant	305bp, 13.86%
ENSR00000065147	Regulatory feature	CNV	regulatory region variant	1012bp, 85.47%

Figure 4. Transcripts and regulatory features overlapping SV nsv916030. URL: https://www.ensembl.org/Homo_sapiens/StructuralVariation/Mappings?sv=nsv916030

ics' page displays frequency distributions across the typed populations (see Figure 3). To show how common an allele is in any assayed ethnic group, we report the highest minor allele frequency observed in any sample set, including 1000 Genomes Project's regional sub-populations. This facilitates improved filtering of common variants from potential disease alleles.

Structural variants

The European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) Database of Genomic Variants archive (DGVa) and NCBI's dbVar (22) are peer archives of structural variant (SV) information. We import SV data from these archives for nine species. To facilitate filtering, we annotate SVs with any transcripts or regulatory features they overlap (see Figure 4) and report variant type and consequence using standardized Sequence Ontology (SO) terms (23). We also calculate population frequencies for SVs discovered in the 1000 Genomes Project.

Citations

Publications contain valuable information including variant to disease associations, but it can be cumbersome to collate an extensive list of references. Variant identifiers are manually extracted from the literature by projects such as the The National Human Genome Research Institute - European Bioinformatics Institute genome wide association study catalog (NHGRI-EBI GWAS Catalog) (24); computationally mined by the University of California, Santa Cruz (25) and by Europe PubMed Central (26); and referenced in submissions to dbSNP and ClinVar. These alternative approaches yield different sets of data. Since our last report, we have implemented data collation and access methods for citations from all of these sources, and citations are now available in Ensembl for nine species. Our variant 'Citations' page lists publications that describe the variant, with links to abstracts and full text where available. We thereby avoid the need to collate lists of references by providing immediate access to a simple overview of the publications discussing a particular variant and easy navigation to detailed reports.

Phenotype and disease information

We aggregate multiple distinct sources of phenotype, disease and trait annotations into a common structure to mitigate many of the challenges of using these data. At present, there is no common format used to exchange such data and projects often use different conventions to describe the same condition. The type of information available for these annotations is also highly heterogeneous and dependent on data source. Annotations can be associated with variants, SVs, genes or quantitative trait loci (QTL), causing additional complexity when seeking all phenotype informa-

tion for a genomic locus. Our data integration methodology facilitates the analysis of heterogeneous annotations from different sources via a single interface.

For human, our primary data sources include ClinVar, OMIM (27) and the NHGRI-EBI GWAS Catalog. For other species, we import data from the Animal QTL database (28), Online Mendelian Inheritance in Animals (OMIA) (29) and a number of species-specific projects such as the Rat Genome Database (30), the Zebrafish Model Organism Database (31) and the International Mouse Phenotyping Consortium (32). We hold phenotype descriptions as used by the data providers and, to facilitate improved querying across conditions described differently in different studies, we map these to ontology terms in the experimental factors (33), human phenotype (34), clinical measurement (35) and mammalian phenotype (36) ontologies, where possible. In Ensembl release 93 (July 2018), 80% of the phenotype descriptions present in our human database were mapped to an ontology term. The mean number of descriptions mapping to an ontology accession is 2.2 but over 8% of accessions map to 5 or more descriptions. An extreme example is deafness (EFO:0001063), which is linked to 125 different descriptions, including 'Deafness, autosomal recessive, 53' and 'Deafness, autosomal dominant, 20'.

These data can be accessed via the Ensembl genome browser by searching for disease or phenotype descriptions or ontology terms. Our phenotype pages display a table of genes, variants and QTLs annotated as being associated with the phenotype, trait or disease (see Figure 5). By default, locus names, genomic locations and the source of the annotation are displayed with links out to the data source and any publications in PubMed. These links provide access to more detailed evidence supporting the assertions. We also display clinical significance assertions and review status from ClinVar, where available, using clear icons to distinguish the different statuses and reporting when conflicting reports have been submitted. When annotations are viewed clustered by ontology term, similar diseases or those sharing phenotypes are displayed. Our variant 'Phenotypes' page lists any phenotypes, traits or diseases that are reported to be associated with the variant. Our gene 'Phenotypes' page shows annotations grouped by gene and also displays phenotypes associated with orthologues in other species, as it is possible that function is shared. Our APIs also provide data access by variant, gene, region and phenotype ontology term.

Data sets with licensing or access restrictions

Some valuable data have license restrictions that prohibit redistribution or limit the detail that can be shown in Ensembl. In other cases, distribution is restricted by specific consent agreements with the research participants. In both cases, we seek to maximize data discoverability while complying with known restrictions. For example, we incorporate the public versions of the Human Gene Mutation Database (HGMD) (37) and the Catalogue of Somatic Mutations in Cancer (COSMIC) (38) data sets into Ensembl. We report these data in region-based and identifier-based searches and provide links to each project's website for additional information. As of Ensembl release 93 (July 2018), there are over 139 000 HGMD identifiers and over 4 million COSMIC identifiers with genomic locations. Variant locations from the DECIPHER (39) project and Leiden Open Variation Database (40), which are restricted by consent agreements, are not available via our APIs but can be viewed as tracks on the 'Region in detail' view with links to the project websites.

A full list of data sources and versions used is available at: https://www.ensembl.org/info/genome/variation/species/sources_documentation.html.

Variant annotation

To enable integration and discovery of the increasing wealth of variant data in many species we have implemented a number of high-throughput annotation pipelines since our last report. We have also developed tools and REST services to provide dynamic data analysis.

Predicted variant effect on gene function

Ensembl creates comprehensive gene annotation for over 100 vertebrate species (41) and regulatory element annotation for human and mouse (42). We analyse variants in Ensembl with respect to this annotation and predict the consequences of each allelic change on any overlapping feature to provide guidance as to its possible functional impact. We use SO terms to describe these consequences, enabling comparison with other annotation sources and querying by both specific and generic terms (for example, both missense and synonymous variants can be extracted using the generic term 'exonic'). To assess the potential deleteriousness of missense variants, we employ the Sorting Intolerant From Tolerant (SIFT) (43) and PolyPhen2 (44) packages, which use protein homology and structural information to predict the impact of a change in amino acid sequence on protein stability and function. SIFT results are available for our most highly accessed species—including human, chicken, cow, dog, goat, horse, mouse, pig, rat and zebrafish—while PolyPhen2 is specific to human. Many additional functional effect algorithms are available for human variants via VEP.

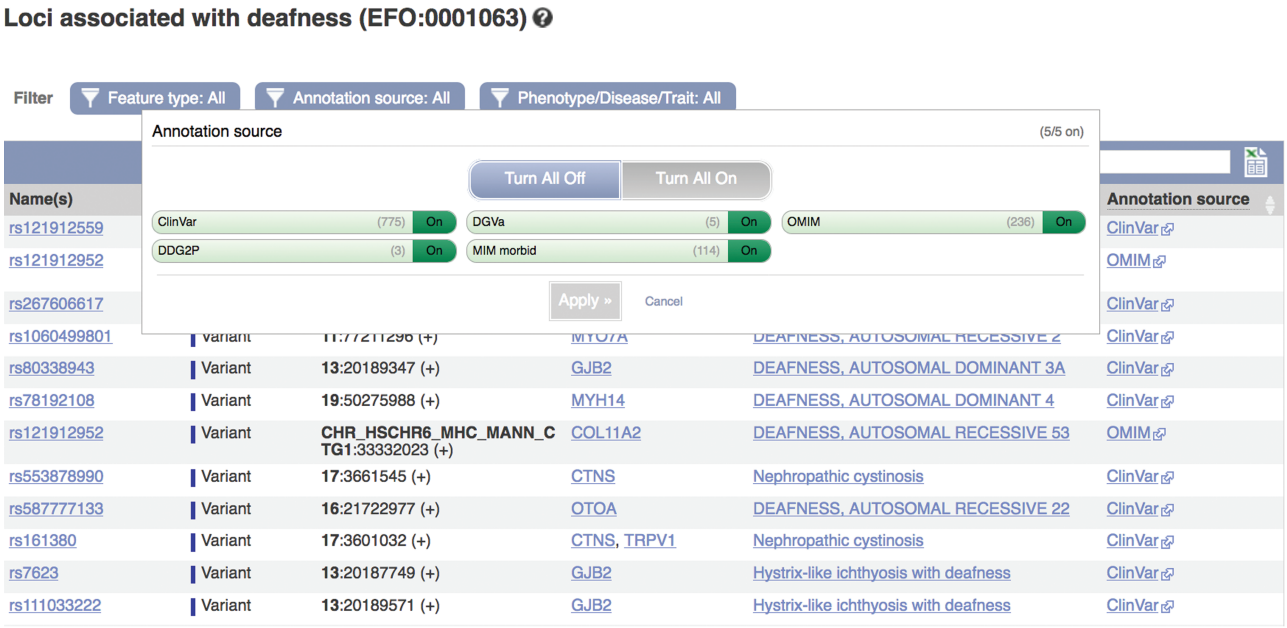


Figure 5. A phenotype table showing variants associated with deafness. URL: https://www.ensembl.org/Homo_sapiens/Phenotype/Locations?oa=EFO:0001063

These results can be viewed grouped by gene or variant. Our variant 'Genes and regulation' page displays the predicted functional consequences for each transcript and regulatory feature the variant overlaps; the position in the transcript, CDS and protein sequences with amino acid and codon change are listed where available. Our transcript 'Variant' page (see Figure 6) displays the predicted functional consequence of each variant in the region, with amino acid position and change, and SIFT and PolyPhen2 scores, where available. To help assess potential deleteriousness, the evidence supporting each variant and the minor allele frequency in the 1000 Genomes Project samples is reported. These results are presented in a table that can be interactively filtered on any of these attributes as well as variant type, location and data source.

Conservation

Allele conservation is one of the strongest predictors that genomic sequence modifications are not tolerated (45). To indicate allele conservation, we use Ensembl's comparative genomics resources (46) to display the sequence flanking a variant aligned with genomic sequence from relevant sets of other species on the variant 'Phylogenetic context' page. We also predict the ancestral allele at each variant location in primate species.

Allele equivalence

Repositories suffer from the lack of a consistent variant reporting standard, allowing equivalent insertions or deletions in repetitive sequence to be reported at different locations depending on whether the change is left or right justified (47). Problems in interpretation can arise when disease annotations are attached to a variant at one location and frequency information to an equivalent variant at a different location. We have implemented a method to systematically extract all variants in overlapping five megabase regions, normalize each allele individually to the most 3′ position at which it can be described and identify variants with equivalent alleles. We list variants with equivalent alleles on our variant page to allow scattered information about a genomic change to be considered together. Over 3 million human variants have equivalent alleles listed in our current release.

Linkage disequilibrium

Variants found to have associations with disease in large-scale studies are rarely causative, so an understanding of linkage disequilibrium (LD) in the region around the variant is essential. Our variant 'Linkage Disequilibrium' page displays LD results calculated in the sample populations typed in the 1000 Genomes Project, supporting tagSNP selection and the interpretation of association results (see Figure 7). In March 2016 we added a REST service, which can be used in analysis pipelines to access both D' and r^2 values by region or variant, and in December 2017, we released a web tool that provides LD results for a single variant, group of variants or region.

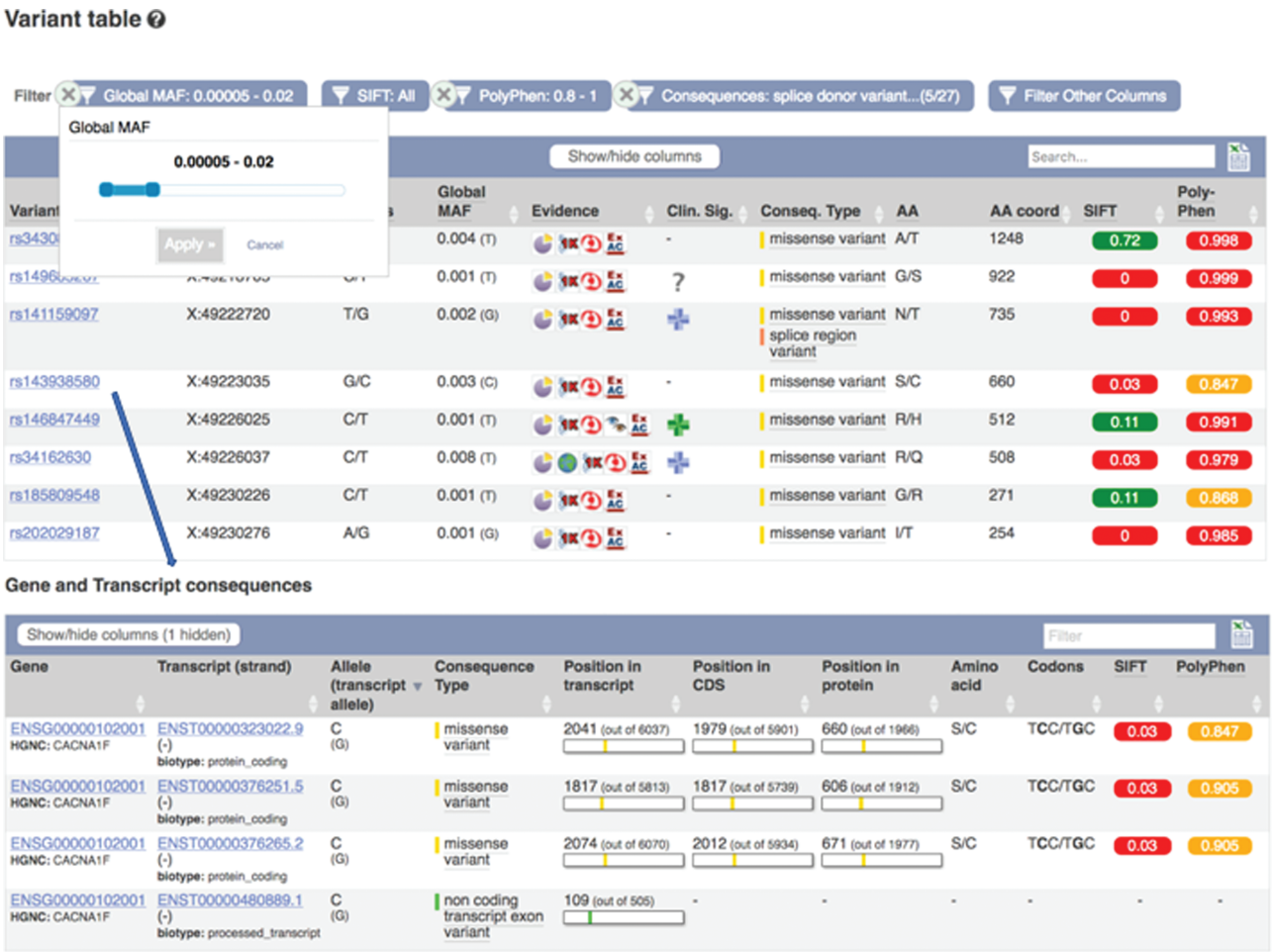


Figure 6. Part of a table showing predicted consequences for the variants overlapping transcript ENST00000323022.9. Table of transcripts overlapping variant rs143938580. URLs: https://www.ensembl.org/Homo_sapiens/Transcript/Variation_Transcript/Table?t=ENST00000323022, https://www.ensembl.org/Homo_sapiens/Variation/Mappings?v=rs143938580

Transcript haplotype frequencies

The reference human genome is a mix of sequences from several different individuals (48) and, as such may contain regions, including very rare alleles, which have not been observed together in a single individual. In studies investigating disease or drug response, it is useful to know the protein forms common in different populations, rather than considering only the translation represented by the reference sequence. (58).

In August 2016, we first provided protein haplotype sequence frequency data for human genes, calculated using the phased genotypes from the 1000 Genomes Phase 3. We consider the effect of all variant alleles acting together on protein function and include SIFT and PolyPhen2 predictions of deleteriousness. Each haplotype is assigned a description showing the allele change and the frequency in the continental populations is calculated. These data are available in both the genome browser (from the 'Haplotypes' tab on the human Transcript pages) and via dedicated REST endpoints. A small number of reference haplotypes were unobserved in the 1000 Genomes samples. The Genome Reference Consortium is now reviewing these.

Variant Recoder

Variants can be named in a variety of ways in literature and database resources, (for example, the identifiers 'ENSP00000420705.2:p.Ser737Asn', 'NM_007294.3:c.5522G>A', 'rs80357368' and 'RCV000236784' all refer to the same variant) causing difficulties in interpretation. To resolve this issue, we have implemented a REST service to return Human Genome Variation Society (HGVS) descriptions and other known names for an input identifier. The Variant Recoder takes accessions from databases such as dbSNP, UniProt and ClinVar as input as well as HGVS at genomic, transcript and protein level. It also decodes some common forms of incorrect HGVS descriptions, such as gene name with a

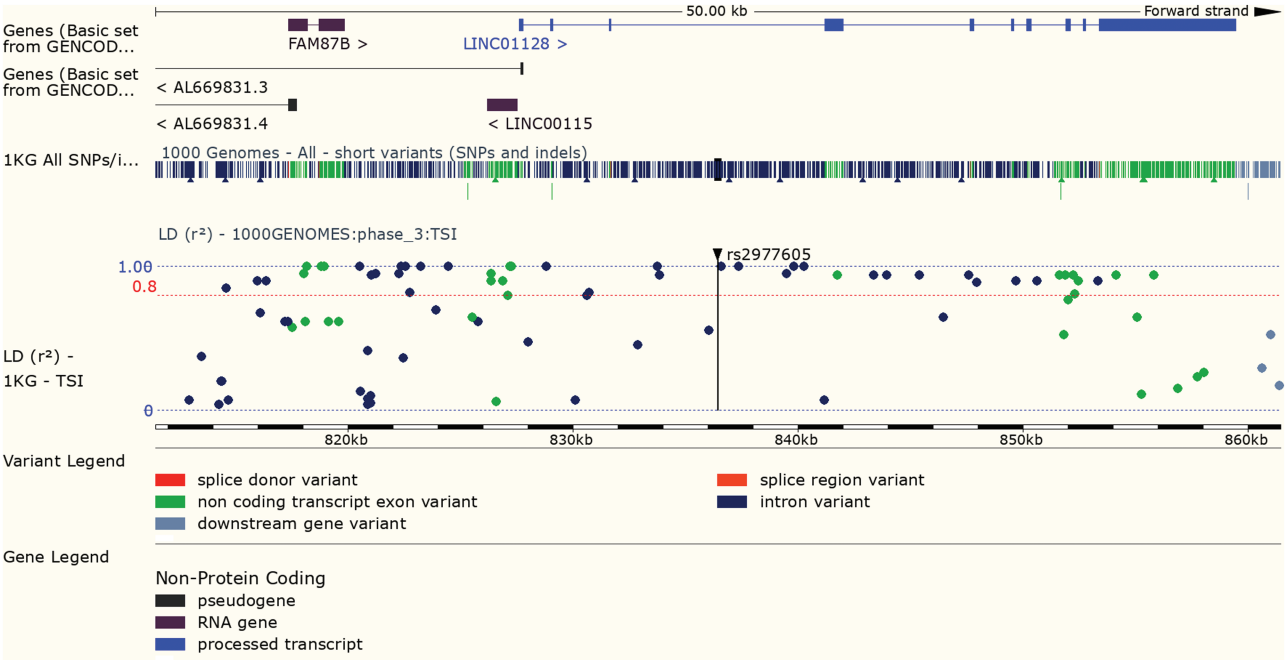


Figure 7. LD results for a variant can be viewed adjacent to gene structure in both Manhattan and Haploview style plots. URL: https://www.ensembl.org/Homo_sapiens/Variation/LDPlot? v=rs2977605;pop1=373537

protein level change, where possible. Warnings are issued when invalid HGVS is input, or results are potentially ambiguous.

Methods

Ensembl databases are built using MySQL and data input and analysis pipelines are written in Perl, normally utilising the eHive (49) workflow management system. While our database schema is subject to change, our Perl API is stable with changes deployed and announced in a controlled way. Specifically, we aim to support deprecated functionality for at least a year to provide ample time for those using our API in their pipelines to make the necessary updates.

Reducing sequencing costs have facilitated a rapid increase in the quantity of variant data available for a number of species, motivating us to regularly revise and optimize our analysis and storage methods. Key compute-intensive API functions, such as checking whether a variant overlaps another genomic feature, have been rewritten in C and can be optionally used through the Perl-XS interface. This brings considerable performance improvements when analysing large numbers of variants. We have also modified our API to use tabix (50), an efficient file access tool, to extract genotype data from Variant Call Format (51) files, removing the need to load large datasets into MySQL. Variant locations are stored in databases, enabling look up by names such as dbSNP refSNP identifier or ClinVar accession, followed by rapid extraction of genotype and allele frequency data from files.

To ensure our tools and data are compatible with other systems, we champion standards for data formatting and have adopted and contributed to the development of many standards. We drove the collaboration to develop the SO and use SO terms to describe both the type of change a variant represents and its consequence on overlapping genomic features (24). Consequences are annotated on the immutable Locus Reference Genomic (52) transcripts as well as the current Ensembl gene set. All variants are annotated using the HGVS (53) nomenclature, which has become the preferred way to describe variants in the clinical community. HGVS descriptions using Ensembl, RefSeq and LRG transcripts are provided where possible.

Discussion

Genome browsers provide an integrated view of biological knowledge. This is essential to aid understanding basic questions about biological function, to provide data for evolutionary studies, and as a basis for genomics to have an impact on healthcare. Ensembl is one of a small number of projects providing variation data within a genome browser. The University of California, Santa Cruz Genome Browser provides a set of detailed and configurable views of genes,

variants and other features but does not support as many species as Ensembl and does not currently provide REST API access. dbSNP provides the most detailed information available on the discovery of individual variants, but provides more limited browser capabilities and programmatic access.

Future work

There are now a multitude of different algorithms available to predict the potential pathogenicity of human variants. We already provide 15 algorithms via VEP and later this year will add further predictions to our variant consequence views, alongside the existing SIFT and PolyPhen2 results. We will also extend the information we make available for interpreting variants outside coding regions. Making sense of large number of, often conflicting, results can be a challenge. We will employ simple colour-coding to provide a results overview and configurable tables to allow algorithm and score selection.

We will also enhance the protein annotations we provide and display variants in the context of protein structures. The number of phenotype and disease annotations in the public domain is increasing; between Ensembl release 56 and 93, the number of short variants with such annotations has risen from less than a thousand to more than 300 thousand. In this time, the proportion of variants with only a single annotation has dropped from 90% to 67% although a key problem is redundancy of information across different resources. We will continue to improve our existing methods to link related information, merge identical records relating to the same assertion and provide detailed provenance to enable filtering and extraction of preferred subsets of data.

Future development will also focus on integrating data at increasing scale. We anticipate large numbers of additional species will be sequenced for variant detection, while the number of sequenced individuals within many species continues to rise (54, http://gnomad.broadinstitute.org/, http://www.1000bullgenomes.com/). For human, this will improve the already extensive catalogue of rare variation across different global populations. Federation with other data distributors is the key to our plan for supporting this increase in the number of species studied and sample depth. Responsibility for the accessioning of variant data for all species but human is transferring from dbSNP to the EVA, providing opportunities for streamlining our processes and data storage. We already display genotype data direct from EVA for variants within the Ensembl system and will extend this functionality to provide Ensembl browser views of variant data held entirely within EVA.

We have played a role in the Global Alliance for Genomics and Health (55), a project whose aim is to accelerate progress in human health by developing harmonized approaches to enable effective and responsible sharing of genomic and clinical data. In particular, we are involved in defining standard models and data exchange formats for variant and variant annotation data. We will adopt relevant standards to be able to consume data from other key resources and to ensure we maximize the interoperability and usability of the data we provide in Ensembl.

Efficient data extraction methods and intuitive displays will be essential to derive full benefit from these increasing data types and volumes. We are currently engaged in a complete redesign of the Ensembl website and seek to implement simple views for the novice and detailed, configurable results selection and extraction tools to meet the specific needs of domain experts.

Conclusion

Ensembl creates unique tools and visualisation for variation data and makes them freely available to the global scientific community. The data are comprehensively updated 4 times per year; each new release incorporates the current public knowledge for approximately 20 species and makes data available through a set of mature and stable interfaces. All data and software developed within the project are freely available. Archive web sites are maintained for at least 5 years to support replication of analyses. Our infrastructure is species agnostic and is used by other projects, such as Ensembl Genomes (56) and GRAMENE (57). We facilitate advances in genomic science by the provision of robust tools and comprehensive data sets to expedite analysis.

Acknowledgements

We acknowledge the contributions of former team members including Graham Ritchie and Pontus Larsson. We also thank the rest of the Ensembl team, especially the members of Ensembl Outreach for supporting our users and providing feedback on new features. We thank the EMBL-EBI and Wellcome Sanger Institute systems teams for maintaining the Ensembl computer systems.

Funding

Ensembl receives majority funding from the Wellcome Trust (grant numbers WT095908, WT098051, WT108749/Z/15/Z). This project has also received funding from the Biotechnology and Biological Sciences Research Council (BB/I025506/1, BB/I025360/2, BB/M011615/1); Open Targets; the Wellcome

Trust (WT200990/Z/16/Z) and the European Molecular Biology Laboratory. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007823. This work has also received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 200754 (GEN2PHEN) and grant agreement n° 222664 (Quantomics). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics).

Conflict of interest. Paul Flicek is a member of the scientific advisory boards for Fabric Genomics, Inc., and Eagle Genomics, Ltd.

References

1. Trynka,G., Hunt,K.A., Bockett,N.A. *et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193–1201.
2. Bourgeois,S., Jorgensen,A., Zhang,E.J. *et al.* (2016) Multifactorial analysis of response to warfarin in a UK prospective cohort. *Genome Med.*, **8**, 2.
3. von Holdt,B.M., Shuldiner,E., Koch,I.J. *et al.* (2017) Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Sci. Adv.*, **3**, e1700398 http://doi.org/10.1126/sciadv.1700398.
4. Scheben,A., Batley,J. and Edwards,D. (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.*, **15**, 149–161.
5. Metzger,J., Karwath,M., Tonda,R. *et al.* (2015) Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics*, **16**, 764.
6. The UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.
7. Vardarajan,B.N., Barral,S., Jaworski,J. *et al.* (2018) Whole genome sequencing of Caribbean Hispanic families with late-onset Alzheimer's disease. *Ann. Clin. Transl. Neurol.*, **5**, 406–417.
8. Richards,S., Aziz,N., Bale,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–423.
9. Chen,Y., Cunningham,F., Rios,D. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293.
10. Parkes,M., Cortes,A.; van Heel,D.A., *et al.* (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.*, **14**, 661–673.
11. Kranis,A., Gheyas,A.A., Boschiero,C. *et al.* (2013) Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics*, **14**, 59.
12. Yates,A., Beal,K., Keenan,S. *et al.* (2015) The Ensembl REST API: Ensembl data for any language. *Bioinformatics*, **31**, 143–145.
13. McLaren,W., Gil,L., Hunt,S.E. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
14. Sherry,S.T., Ward,M.H., Kholodov,M. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
15. Landrum,M.J., Lee,J.M., Benson,M. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
16. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
17. Whirl-Carrillo,M., McDonagh,E., Hebert,J. *et al.* (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, **92**, 414–417.
18. Adams,D.J., Doran,A.G., Lilue,J. *et al.* (2015) The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mamm. Genome*, **26**, 403–412.
19. Alberto,F.J., Boyer,F., Orozco-terWengel,P. *et al.* (2018) Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.*, **9**, 813. doi:10.1038/s41467-018-03206-y.
20. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
21. Lek,M., Karczewski,K.J., Minikel,E. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
22. Lappalainen,I., Lopez,J., Skipper,L. *et al.* (2013) dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
23. Cunningham,F., Moore,B., Ruiz-Schultz,N. *et al.* (2015) Improving the Sequence Ontology terminology for genomic variant annotation. *J. Biomed. Semantics*, **6**, 32.
24. MacArthur,J., Bowler,E., Cerezo,M. *et al.* (2017) The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
25. Tyner,C., Barber,G.P., Casper,J. *et al.* (2017) The UCSC Genome Browser Database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
26. The Europe PMC Consortium (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.*, **43**, D1042–D1048.
27. Amberger,J.S., Bocchini,C.A., Schiettecatte,F. *et al.* (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
28. Hu,Z.-L., Park,C.A. and Reecy,J.M. (2016) Developmental progress and current status of the animal QTLdb. *Nucleic Acids Res.*, **44**, D827–D833.
29. Lenffer,J., Nicholas,F.W., Castle,K. *et al.* (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599.
30. Shimoyama,M., De Pons,J., Hayman,G.T. *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
31. Howe,D.G., Bradford,Y.M., Conlin,T. *et al.* (2013) ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res.*, **41**, D854–D860.

32. Brown,S.D.M. and Moore,M.W. (2012) The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. *Mamm. Genome*, **23**, 632–640.
33. Malone,J., Holloway,E., Adamusiak,T. *et al.* (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
34. Köhler,S., Doelken,S.C., Mungall,C.J. *et al.* (2014) The Human Phenotype Ontology Project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.
35. Smith,J.R., Park,C.A., Nigam,R. *et al.* (2013) The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications. *J. Biomed. Semantics*, **4**, 26.
36. Smith,C.L., Goldsmith,C.-A.W. and Eppig,J.T. (2004) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
37. Stenson,P.D., Mort,M., Ball,E.V. *et al.* (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.
38. Forbes,S.A., Beare,D., Gunasekaran,P. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
39. Firth,H.V., Richards,S.M., Bevan,A.P. *et al.* (2009) DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am. J. Hum. Genet.*, **84**, 524–533.
40. Fokkema,I.F.A.C., Taschner,P.E.M., Schaafsma,G.C.P. *et al.* (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.*, **32**, 557–563.
41. Aken,B.L., Ayling,S., Barrell,D. *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, article ID baw093, https://doi.org/10.1093/database/baw093.
42. Zerbino,D.R., Johnson,N., Juetteman,T. *et al.* (2016) Ensembl regulation resources. *Database*, **2016**, article ID bav119, https://doi.org/10.1093/database/bav119.
43. 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. Protocols*, **4**, 1073–1081.
44. Adzhubei,I.A., Schmidt,S., Peshkin,L. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248–249.
45. Kircher,M., Witten,D.M., Jain,P. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
46. Herrero,J., Muffato,M., Beal,K. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**, article ID bav096, https://doi.org/10.1093/database/bav096.
47. Assmus,J., Kleffe,J., Schmitt,A.O. *et al.* (2013) Equivalent indels—ambiguous functional classes and redundancy in databases. *PLoS ONE*, **8**, e62803 https://doi.org/10.1371/journal.pone.0062803.
48. Schneider,V.A., Graves-Lindsay,T., Howe,K. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
49. Severin,J., Beal,K., Vilella,A.J. *et al.* (2010) eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics*, **11**, 240.
50. Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
51. Danecek,P., Auton,A., Abecasis,G. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
52. MacArthur,J.A.L., Morales,J., Tully,R.E. *et al.* (2014) Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.*, **42**, D873–D878.
53. den Dunnen,J.T., Dalgleish,R., Maglott,D.R., *et al.* (2016) HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.*, **37**, 564–569.
54. Alonso-Blanco,C., Andrade,J., Becker,C. *et al.* (2016) 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, **166**, 481–491.
55. The Global Alliance for Genomics and Health (2016) A federated ecosystem for sharing genomic, clinical data. *Science*, **352**, 1278–1280.
56. Kersey,P.J., Allen,J.E., Allot,A. *et al.* (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, **46**, D802–D808.
57. Tello-Ruiz,M.K., Naithani,S., Stein,J.C. *et al.* (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.*, **46**, D1181–D1189.
58. Spooner,W., McLaren,W., Slidel,T. *et al.* (2018) Haplosaurus Computes Protein Haplotypes for Use in Precision Drug Design. *Nature Communications*, **9**, 4128, https://doi.org/10.1038/s41467-018-06542-1.