Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx445

HGVA: the Human Genome Variation Archive

Javier Lopez[1,*], Jacobo Coll[1], Matthias Haimel[2,3,4], Swaathi Kandasamy[2], Joaquin Tarraga[5], Pedro Furio-Tari[1], Wasim Bari[1], Marta Bleda[2,3,4], Antonio Rueda[1], Stefan Gräf[2,3,4], Augusto Rendon[1,2], Joaquin Dopazo[6,7,8,*] and Ignacio Medina[5,*]

[1]Genomics England, Charterhouse Square, London EC1M 6BQ, UK, [2]Department of Haematology, University of Cambridge, Cambridge CB2 0PT, UK, [3]Department of Medicine, University ofCambridge, Cambridge, CB2 0QQ, UK, [4]NIHR BioResource–Rare Diseases,Cambridge University Hospitals, CambridgeBiomedical Campus, Cambridge CB2 0QQ, UK, [5]HPC Service, UIS, University of Cambridge, Cambridge CB3 0FB, UK, [6]Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Hospital Virgen del Rocio, Sevilla 41013, Spain, [7]Functional Genomics Node (INB), FPS, Hospital Virgen del Rocio, Sevilla 41013, Spain and [8]Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), FPS, Hospital Virgen del Rocio, Sevilla 41013, Spain

Received February 24, 2017; Revised April 14, 2017; Editorial Decision April 26, 2017; Accepted May 19, 2017

ABSTRACT

High-profile genomic variation projects like the 1000 Genomes project or the Exome Aggregation Consortium, are generating a wealth of human genomic variation knowledge which can be used as an essential reference for identifying disease-causing genotypes. However, accessing these data, contrasting the various studies and integrating those data in downstream analyses remains cumbersome. The Human Genome Variation Archive (HGVA) tackles these challenges and facilitates access to genomic data for key reference projects in a clean, fast and integrated fashion. HGVA provides an efficient and intuitive web-interface for easy data mining, a comprehensive RESTful API and client libraries in Python, Java and JavaScript for fast programmatic access to its knowledge base. HGVA calculates population frequencies for these projects and enriches their data with variant annotation provided by CellBase, a rich and fast annotation solution. HGVA serves as a proof-of-concept of the genome analysis developments being carried out by the University of Cambridge together with UK's 100 000 genomes project and the National Institute for Health Research BioResource Rare-Diseases, in particular, deploying open-source for Computational Biology (OpenCB) software platform for storing and analyzing massive genomic datasets.

INTRODUCTION

Over the last decade, next generation sequencing techniques have become an indispensable asset for the identification of disease-causing genomic variants. Interestingly, a sheer wealth of mutations can be found in any human genome (∼5 million per human genome, ∼150 000 per human exome) many of which may provide a compelling story about how the variant may influence a trait (1,2). Transforming these data into scientifically or clinically useful information involves a heuristic filtering process during which tolerable, likely non-causal variants are excluded from the analysis. One of the most powerful metrics in the filtering procedure is the variant minor allele frequency within a given population. Variants that appear at high frequencies in other genomic studies are unlikely to be disease causing. Hence, large reference datasets such as the 1000 Genomes Project (1KGP), the NHLBI Exome Sequencing Project (ESP), the Exome Aggregation Consortium (ExAC) or the Genome of the Netherlands project (3–6), are generating a genomic variation knowledge base which sets an essential resource for identifying disease-causing genotypes. Due to the complexity of the data, it is challenging and laborious to integrate into analysis pipelines for disease gene prioritization. In addition, these large datasets are distributed across repositories, lack normalization and are delivered in different tastes of Variant Calling Format (VCF) files. Access to these datasets, standardizing the data across multiple projects and integration of those data in the analysis process is often extremely cumbersome, even for experienced bioinformaticians.

We present here the Human Genome Variation Archive (HGVA), an open access genetic variation resource. HGVA

*To whom correspondence should be addressed. Tel: +44 122 376 3554; Email: im411@cam.ac.uk
Correspondence may also be addressed to Javier Lopez. Tel: +44 122 376 3554; Email: javier.lopez@genomicsengland.co.uk
Correspondence may also be addressed to Joaquin Dopazo. Tel: +34 95 5313241; Email: joaquin.dopazo@juntadeandalucia.es

Table 1. List of studies and versions available at HGVA

Project name	Studies	Version/date
Reference GRCh37	1000 genomes project GRCh37	Phase 3 2016–05
	Exome Sequencing Project (http://evs.gs.washington.edu/EVS/)	2016-05
	Exome Aggregation Consortium	0.3.1 2016-05
	Genome of the Netherlands	Release 5 2016–05
	UK10K project	2016–05
	Spanish Medical Genome Project	2016–12
Reference GRCh38	1000 genomes project GRCh38	Phase 3 2016–10
Cancer GRCh37	QIMR Berghofer melanoma	2016–12
	Chronic myeloid leukemia–Russian Academy of Medical Sciences (http://www.med-gen.ru/en/)	2016–12
Platinum	Illumina platinum	2015–08

normalizes, annotates and merges together in a NoSQL database all variants from world-wide leading reference projects, which are the prime source for most genomic variation analysis pipelines. Of particular interest is the annotation of population allele frequencies, which can be especially difficult to calculate and parse, given the size of the data and the different analysis pipelines used. HGVA provides an efficient user-friendly web-interface, a comprehensive RESTful API and client libraries in Python, Java and JavaScript for fast programmatic access to its knowledge base.

MATERIALS AND METHODS

Variation projects

We have selected a number of high-profile projects which are the initial reference for most genomic variation analysis pipelines. These projects include germline and somatic (cancer) datasets mapped on to GRCh37 and GRCh38 assemblies. The full list of datasets currently loaded is shown in Table 1. Further details on all data sources can be found at the HGVA documentation site (http://docs.opencb.org/display/hgva/Datasets+and+Studies) and at the corresponding references (3,5,7–11).

Data processing

Data are processed and loaded into HGVA using the following pipeline (Figure 1). First, public VCF and pedigree (PED) files (where available) are downloaded from the different repositories. Second, data are transformed in a process which comprises several steps including variant parsing, normalization and conversion into our data models. VCF files are read using the Java library HTSJDK (https://samtools.github.io/htsjdk/), which provides a syntactic validation of the data. Further quality control actions to validate the data are included e.g. duplicate or overlapping variant detection. Malformed variants are excluded. Finally, the load stage will populate a NoSQL database(s) with the processed data and merge variants as appropriate. Loaded variants are then annotated and indexed to speed up queries.

Normalization is an essential operation of the transformation given that the different variant callers from different studies can generate different representations of the same variant calls. Normalization comprises the following tasks:

i) Chromosome nomenclature: remove prefixes. The list of known chromosome prefixes are: chrom, chrm, chr and ch.
ii) Reference/alternate trimming (12). This stage consists of removing the trailing (right trimming) and leading (left trimming) bases that are identical in both alleles.
iii) Multi-allelic split. For positions where multiple alleles exist. After splitting variants, the previously described normalization steps must be re-applied over each of the split alleles.

The result of the transform step is written into Avro (https://avro.apache.org/) file format to provide a high-performance and compact binary file. These avro files are subsequently loaded into a MongoDB database. Avro files from the same HGVA project (e.g. Reference GRCh37) are loaded and merged together in the same database. Loading variants from multiple files into a single database will effectively merge them. In most of the scenarios, with a proper normalization, merging variants is straightforward. Positions with multiple alternate alleles or overlapping variants, however, require additional logic.

Once the genomic variants are loaded and indexed into the database we annotate variants and calculate allele frequencies when genotypes are available. Variant annotation is carried out by using the variant annotator provided by the CellBase project (13) and includes these main annotation types: core variant data, consequence types, clinical annotations, deleteriousness scores and conservation scores. Further details on variant annotation sources are documented in the CellBase documentation (http://docs.opencb.org/display/cellbase/Data+sources+and+species). Allele frequency parsers were implemented for frequency formats provided by ExAC, ESP, GoNL, UK10K, MGP and the 1000 Genomes Project GRCh38 data as provided by ENSEMBL. Allele counts and genotype counts are calculated for all populations and super-populations defined by The 1000 Genomes Project GRCh37. Additional population frequencies can be easily incorporated in the future by simply creating new sample cohorts in the system and running the frequency calculation procedure.

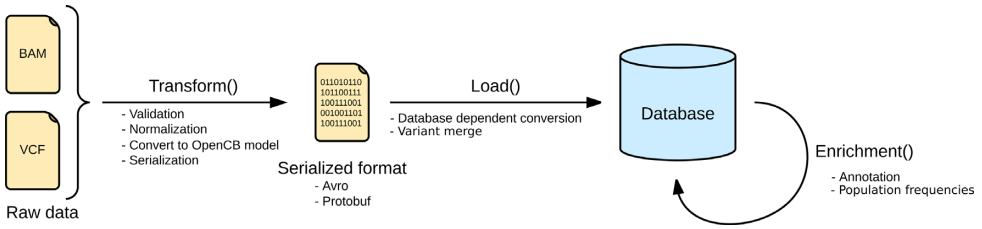

Figure 1. Pipeline for processing and loading variants into the Human Genome Variation Archive (HGVA).

RESULTS

Architecture

HGVA follows the standard client-server architecture. The server uses the OpenCGA (https://github.com/opencb/opencga) project from open-source for Computational Biology (OpenCB) to index VCF files using different NoSQL technologies and to provide a high-performance variant query engine through RESTful web services to the different clients developed: command-line (CLI) to query and download data, client libraries for programmatic access and a web-based data mining tool based on OpenCB Interactive Variant Analysis (IVA) (https://github.com/opencb/iva) for visual and interactive analysis.

The HGVA web-based application makes intensive use of the HTML5 standard and recent web technologies such as web components, currently available in all modern web browsers. All of the higher level interfaces (CLI and client libraries) make intensive use of HGVA's RESTful API (Figure 2). The Java implementation of this API was designed to ensure optimum performance, scalability and maintainability of the whole system. Likewise, the reference infrastructure that runs the system was carefully designed to provide a robust and efficient service: the REST API is accessible through an HAProxy load balancer server to ensure high availability. Currently, RESTful web services run in two Apache Tomcat servers with 8 cores and 32GB RAM each. Google Remote Procedure Call (gRPC) (http://www.grpc.io) servers are currently under testing and will be enabled soon to improve data streaming performance. The RESTful API allows to query either the variants or the metadata, such as: data files, sample data, populations, etc. Variant Storage module allows managing and querying genomic variant data. Currently MongoDB is used as the database; a replica-set with three MongoDB servers supports HGVA installation, each of them comprising 24 cores, 256GB RAM and 15TB storage space. A single Apache Solr (http://lucene.apache.org/solr/) server is used for full text search and faceted queries.

Interfaces

HGVA provides an intuitive web-based data mining tool (http://hgva.opencb.org) based on OpenCB IVA project (http://docs.opencb.org/display/iva) and a comprehensive RESTful web service API (http://bioinfo.hpc.cam.ac.uk/hgva/webservices) based on OpenCGA project (http://docs.opencb.org/display/opencga). To query genomic variants, some client libraries in Python, Java and JavaScript have been developed as well as a CLI interface.

The web interface home page provides a text entry box which accepts genomic regions, rs ids, variants or gene ids as search terms. Entering a search will automatically change the view to the main Variant Browser (Figure 3). This interface offers fast and interactive browsing with numerous filtering options that include genomic regions, pathogenicity, conservation scores, Human Phenotype Ontology (HPO) or Gene Ontology terms (14–16). For example, variant lists can be uploaded to annotate them against reference project data and could be combined with numerous filters. In addition, a checkbox list enables two different study search modes: (i) searched variants that are present in any of the selected datasets, (ii) searched variants must be present in at least one of the selected studies. Finally, an active filter bar is provided at the top which facilitates enabling/disabling search filters.

Results are presented in an interactive table which provides an initial view of the most significant variant information: genomic coordinates, rs id, most severe consequence type, population frequencies, CADD, GERP scores (17) and clinical annotation. Colored values and heat maps in this table aim to facilitate the identification of variants of interest. Selecting any of the variants allows the user to instantly visualize a more comprehensive variant information provided in the tabs at the bottom of the web page. This detailed information allows the user to contextualize the variant in an interactive genome browser and run the beacon network in order to determine if other resource projects that have reported the same variant.

Results can be downloaded in TSV or JSON file formats. Nevertheless, massive downloads of data are discouraged; as previously commented an efficient RESTful API and libraries in different languages are provided for programmatic access. Therefore, no more than 2000 filtered variants can be downloaded at a time. Moreover, a 'Share' button allows generation and export of an URL that saves current results so that they can be easily shared.

Clicking on the gene names in the main variant grid will open a gene view. In this view, gene annotation summary is detailed. Below, another variant grid lists only those variants related to this particular gene. Quick filters by 'Missense' and 'LoF' consequence types can be activated using the corresponding buttons. Finally, a 'Protein' tab displays a lollipop diagram which visualizes the list of variants affecting the corresponding protein, with different lollipop sizes and colors that facilitate the interpretation of variant effects in proteins.

HGVA has been extensively tested in different web browsers including Chrome 49+, Firefox 45+, Microsoft Edge 14+ and Safari 10+.

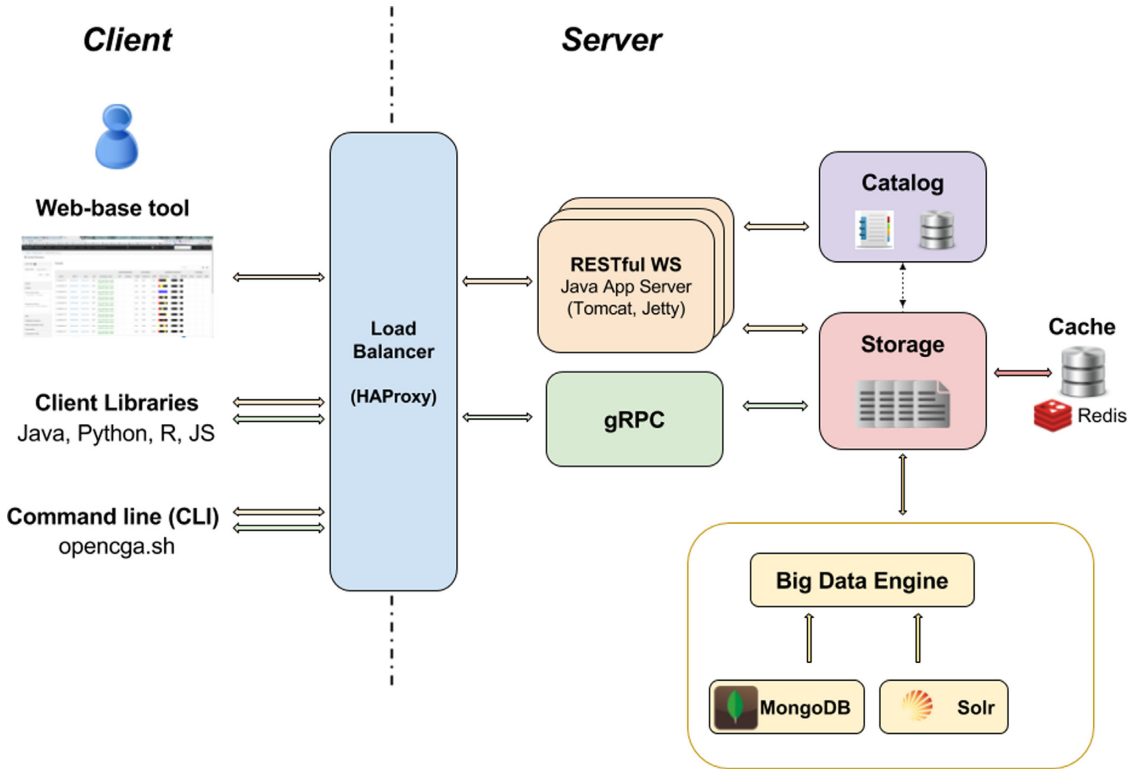

Figure 2. HGVA infrastructure/architecture.

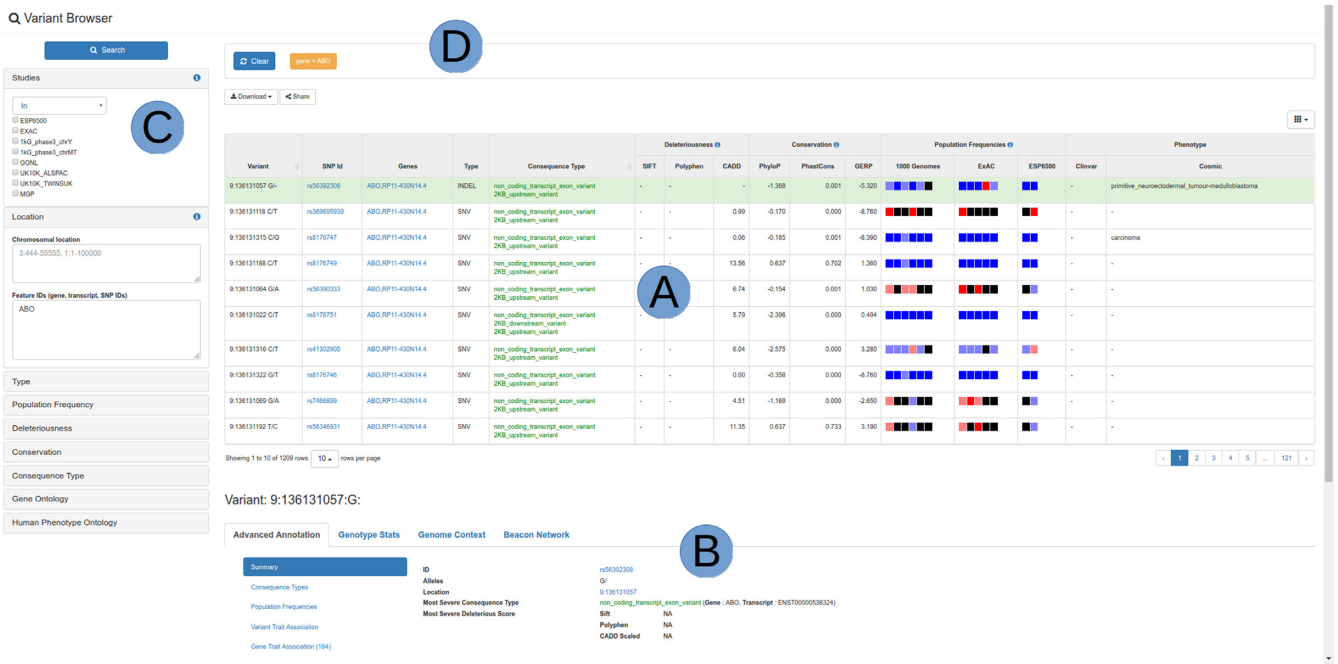

Figure 3. (**A**) Variant grid. Shows a summary of variant annotation data. (**B**) Detailed annotation for the corresponding variant selected in the variant grid. (**C**) Filters menu enabling filtering options from genomic region or gene names to CADD scores, HPO or Gene Ontology terms. (**D**) Filter bar display to facilitate quick editing/removal of selected filters.

In addition to the above described web interface, client libraries are also provided in Python, JavaScript and Java, which enable programmatic access to the whole REST functionality and facilitate script integration. Client libraries are the best way for users to query the REST API. Nevertheless, under certain circumstances users may require direct access the RESTful API. In order to provide a solution for those occasions the REST API is accessible and extensively specified at: http://bioinfo.hpc.cam.ac.uk/hgva/webservices. The REST API provides a number of methods and filters to query study metadata and variation data. A feature of particular interest to bioinformaticians is the extensive list of filtering parameters implemented to allow great flexibility on variant queries.

A detailed documentation at http://docs.opencb.org/display/hgva provides abundant information on the specification of the different interfaces as well as tutorials and an extensive list of usage examples. Of special interest for HGVA users will be the variant query functionality. A comprehensive summary of the functionality implemented in the different interfaces is provided at http://docs.opencb.org/display/hgva/Querying+Variants+with+the+Command+Line. It is also worth mentioning here that the CLI is a wrapper of the REST API and implements all its functionality.

Related tools

Some large-scale genomic projects, such as the Exome Sequencing Project, the ExAC or the Genome of the Netherlands, provide web interfaces to access to the aggregated data, however some data mining features such as complex queries are missing and the type of queries and the accompanying information are quite limited. Other resources, such as the European Variation Archive (EVA, http://www.ebi.ac.uk/eva) or dbSNP (18), are more oriented to archiving data and, similarly, the functionality provided through the web interfaces is limited. Moreover, these repositories rely on downloads of full datasets from their FTP servers and do not offer variant browsing functionality for high throughput data analysis, e.g. searches by lists of variants are not enabled. Furthermore, they do not provide libraries for programmatic access and offer quite limited functionality for remote programmatic access.

The ENSEMBL Variation resource adds valuable phenotype data and annotations from multiple sources to their collection. Nevertheless, the ENSEMBL Variation can provide variants lacking some normalization, which can lead to issues in the analysis. Web access to ENSEMBL variation data mining is provided through the BioMart interface and usage of the Perl API is not straightforward.

In contrast, HGVA is more oriented to offer rich informative data on human genomic variability, either from the user-friendly web interface or by programmatic access through advanced web services API. HGVA has selected and will keep integrating, high-profile studies which are the most widely used in genomic variation analysis pipelines.

DISCUSSION

HGVA is a resource to facilitate clean, fast, integrative access to key, high-quality reference project data brought together in one database. HGVA calculates population frequencies after normalization for these projects and enriches their data with CellBase annotation, one of the most comprehensive and fastest annotation solutions available to date. HGVA provides an efficient intuitive web-based user interface, an all-embracing RESTful API and client libraries in Python, Java and JavaScript for fast programmatic access to its knowledge base. Client libraries are easy to use, follow the logic of the REST API and are extensively documented at (http://docs.opencb.org/display/hgva/RESTful+Web+Services+and+Clients). In addition, a R client library is currently under development and will be made available in due course.

By providing efficient remote access HGVA prevents massive downloads of data required to use genomic allele frequencies in prioritization pipelines. Data size and complexity will increase as technologies and general reference resources improve. Movements of large datasets should be avoided since it is an expensive operation in terms of time and storage space. Local installations of databases are difficult to maintain and tracking of source database versions used in the analysis affects long term reproducibility of analysis results. Users of HGVA will always automatically access the most recent versions of reference data without the need of periodical downloads and re-installations involving hundreds of Gigabytes. The current release comprises over 650GB of compressed data. Moreover, novel communication protocols such as the gRPC, rarely used in Bioinformatics, offer powerful streaming features with high performance, scalability and flexibility. HGVA implements gRPC access for variant queries.

Although other tools provide web-based access to specific genomic repositories, none covers the functionality nor the wide variety of added-value annotations included in HGVA. This big data ready infrastructure solution is a timely proposal that addresses current computational challenges and is a useful framework for any human genome variation data analyst.

This resource is the result of a long-term effort supported by Genomics England, the National Institute for Health Research (NIHR), the University of Cambridge and the Spanish Network for Research in Rare Diseases (CIBERER). It is deemed as proof-of-concept for the developments of OpenCGA and IVA carried out by the NIHR BioResource–Rare Diseases consortium and Genomics England to make more than 100 000 whole genomes accessible for analysis and interpretation. HGVA is an open-source and collaborative project with both frontend and backend being freely downloadable at https://github.com/opencb/iva and https://github.com/opencb/opencga. HGVA is available at http://hgva.opencb.org

AVAILABILITY

HGVA is powered by OpenCGA, IVA and CellBase projects. All these projects are open source collaborative initiatives available in the GitHub repository (http://github.com/opencb/opencga, http://github.com/opencb/iva and http://github.com/opencb/cellbase)

ACKNOWLEDGEMENTS

We are indebted to Kathy Stirrups for reviewing the English of the manuscript.

FUNDING

Spanish Ministry of Economy and Competitiveness [BIO2014-57291-R, in part]; 'Instituto de Salud Carlos III' (ISCIII) 'Plataforma de Recursos Biomoleculares y Bioinformáticos' [PT13/0001/0007]; European Regional Development Funds (ERDF) [EU H2020-INFRADEV-1-2015-1 ELIXIR-EXCELERATE (ref. 676559)]; National Institute for Health Research (NIHR) England [RG65966]. Funding for open access charge: MINECO (SPAIN); National Institute for Health Research (NIHR).
Conflict of interest statement. None declared.

REFERENCES

1. Goldstein,D.B., Allen,A., Keebler,J., Margulies,E.H., Petrou,S., Petrovski,S. and Sunyaev,S. (2013) Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.*, **14**, 460–470.
2. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
3. 1000 Genomes Project Consortium, Auton,A, Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
4. Boomsma,D.I., Wijmenga,C., Slagboom,E.P., Swertz,M.A., Karssen,L.C., Abdellaoui,A., Ye,K., Guryev,V., Vermaat,M., van Dijk,F. *et al.* (2014) The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.*, **22**, 221–227.
5. Auer,P.L., Reiner,A.P., Wang,G., Kang,H.M., Abecasis,G.R., Altshuler,D., Bamshad,M.J., Nickerson,D.A., Tracy,R.P., Rich,S.S. *et al.* (2016) Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. *Am. J. Hum. Genet.*, **99**, 791–801.
6. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
7. Xu,C., Tachmazidou,I., Walter,K., Ciampi,A., Zeggini,E., Greenwood,C.M.T. and UK10K Consortium (2014) Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.*, **38**, 281–290.
8. Dopazo,J., Amadoz,A., Bleda,M., Garcia-Alonso,L., Alemán,A., García-García,F., Rodriguez,J.A., Daub,J.T., Muntané,G., Rueda,A. *et al.* (2016) 267 spanish exomes reveal population-specific differences in disease-related genetic variation. *Mol. Biol. Evol.*, **33**, 1205–1218.
9. Arafeh,R., Qutob,N., Emmanuel,R., Keren-Paz,A., Madore,J., Elkahloun,A., Wilmott,J.S., Gartner,J.J., Di Pizio,A., Winograd-Katz,S. *et al.* (2015) Recurrent inactivating RASA2 mutations in melanoma. *Nat. Genet.*, **47**, 1408–1410.
10. Eberle,M.A., Fritzilas,E., Krusche,P., Källberg,M., Moore,B.L., Bekritsky,M.A., Iqbal,Z., Chuang,H.-Y., Humphray,S.J., Halpern,A.L. *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27**, 157–164.
11. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
12. Tan,A., Abecasis,G.R. and Kang,H.M. (2015) Unified representation of genetic variants. *Bioinformatics*, **31**, 2202–2204.
13. Bleda,M., Tarraga,J., de Maria,A., Salavert,F., Garcia-Alonso,L., Celma,M., Martin,A., Dopazo,J. and Medina,I. (2012) CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic Acids Res.*, **40**, W609–W614.
14. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
15. Köhler,S., Vasilevsky,N.A., Engelstad,M., Foster,E., McMurry,J., Aymé,S., Baynam,G., Bello,S.M., Boerkoel,C.F., Boycott,K.M. *et al.* (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
17. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
18. Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.