D1130–D1137 Nucleic Acids Research, 2021, Vol. 49, Database issue *Published online 29 September 2020*
doi: 10.1093/nar/gkaa794

CSVS, a crowdsourcing database of the Spanish population genetic variability

María Peña-Chilet[1,2,3], Gema Roldán[1], Javier Perez-Florido[1,3,4], Francisco M. Ortuño[1,3,4], Rosario Carmona[1], Virginia Aquino[1], Daniel Lopez-Lopez[1,3], Carlos Loucera[1,3], Jose L. Fernandez-Rueda[1], Asunción Gallego[5], Francisco García-Garcia[6], Anna González-Neira[7], Guillermo Pita[7], Rocío Núñez-Torres[7], Javier Santoyo-López[8], Carmen Ayuso[9], Pablo Minguez[9,10], Almudena Avila-Fernandez[9], Marta Corton[9], Miguel Ángel Moreno-Pelayo[11], Matías Morin[11], Alvaro Gallego-Martinez[12,13], Jose A. Lopez-Escamez[12,13], Salud Borrego[14,15], Guillermo Antiñolo[14,15], Jorge Amigo[16], Josefa Salgado-Garrido[17], Sara Pasalodos-Sanchez[17], Beatriz Morte[18], The Spanish Exome Crowdsourcing Consortium, Ángel Carracedo[16,19], Ángel Alonso[17] and Joaquín Dopazo[1,2,3,4,*]

[1]Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Hospital Virgen del Rocío, Sevilla 41013, Spain, [2]Bioinformatics in Rare Diseases (BiER), Center for Biomedical Network Research on Rare Diseases (CIBERER), ISCIII, Sevilla 41013, Spain, [3]Computational Systems Medicine group, Institute of Biomedicine of Seville (IBIS) Hospital Virgen del Rocío, Sevilla 41013, Spain, [4]Functional Genomics Node, FPS/ELIXIR-ES, Hospital Virgen del Rocío, Sevilla 41013, Spain, [5]Sistemas Genomicos, Paterna, Valencia 46980, Spain, [6]Unidad de Bioinformática y Bioestadística, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain, [7]Human Genotyping Unit–Centro Nacional de Genotipado (CEGEN), Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain, [8]Edinburgh Genomics, The University of Edinburgh, Edinburgh EH9 3FL, UK, [9]Department of Genetics, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid 28040, Spain, [10]Center for Biomedical Network Research on Rare Diseases (CIBERER), ISCIII, Madrid 28040, Spain, [11]Servicio de Genética, Ramón y Cajal Institute of Health Research (IRYCIS) and Biomedical Network Research Centre on Rare Diseases (CIBERER), Madrid 28034, Spain, [12]Otology & Neurotology Group CTS 495, Department of Genomic Medicine, Centre for Genomics and Oncological Research (GENYO), Pfizer University of Granada, Granada 18016, Spain, [13]Department of Otolaryngology, Instituto de Investigación Biosanitaria, IBS. GRANADA, Hospital Universitario Virgen de las Nieves, Universidad de Granada, Granada 18016, Spain, [14]Department of Maternofetal Medicine, Genetics and Reproduction, Institute of Biomedicine of Seville (IBIS), University Hospital Virgen del Rocío/CSIC/University of Seville, Seville 41013, Spain, [15]Centre for Biomedical Network Research on Rare Diseases (CIBERER), Seville 41013, Spain, [16]Fundación Pública Galega de Medicina Xenómica, SERGAS, IDIS, Santiago de Compostela 15706, Spain, [17]Navarrabiomed-IdiSNA, Complejo Hospitalario de Navarra, Universidad Pública de Navarra (UPNA), IdiSNA (Navarra Institute for Health Research), Pamplona, Navarra 31008, Spain, [18]Undiagnosed Rare Diseases Programme (ENoD). Center for Biomedical Research on Rare Diseases (CIBERER), ISCIII, Madrid 28029, Spain and [19]Grupo de Medicina Xenómica, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), CIMUS, Universidade de Santiago de Compostela, Santiago de Compostela, España

Received August 13, 2020; Revised September 08, 2020; Editorial Decision September 10, 2020; Accepted September 10, 2020

ABSTRACT

The knowledge of the genetic variability of the local population is of utmost importance in personalized medicine and has been revealed as a critical factor for the discovery of new disease variants. Here, we present the Collaborative Spanish Variability Server (CSVS), which currently contains more than 2000 genomes and exomes of unrelated

[*]To whom correspondence should be addressed. Tel: +34 677910685; Email: joaquin.dopazo@juntadeandalucia.es

Spanish individuals. This database has been generated in a collaborative crowdsourcing effort collecting sequencing data produced by local genomic projects and for other purposes. Sequences have been grouped by ICD10 upper categories. A web interface allows querying the database removing one or more ICD10 categories. In this way, aggregated counts of allele frequencies of the pseudo-control Spanish population can be obtained for diseases belonging to the category removed. Interestingly, in addition to pseudo-control studies, some population studies can be made, as, for example, prevalence of pharmacogenomic variants, etc. In addition, this genomic data has been used to define the first Spanish Genome Reference Panel (SGRP1.0) for imputation. This is the first local repository of variability entirely produced by a crowdsourcing effort and constitutes an example for future initiatives to characterize local variability worldwide. CSVS is also part of the GA4GH Beacon network.

CSVS can be accessed at: http://csvs.babelomics.org/.

INTRODUCTION

Sequencing technologies have experienced an unprecedented development during the last decade (1) that resulted in different international collaborative projects (2–4) which contributed to an extraordinary increase in the knowledge of the mutational spectrum of diseases. This generation of knowledge has been especially significant in diseases with high morbidity and mortality, caused by highly penetrant (typically protein-coding) variants (5,6). In fact, more than 4500 monogenic diseases can nowadays be directly diagnosed by personalized genomics (7), a possibility that might soon be extended to the whole spectrum of rare diseases with a genetic background (8). Among the strategies used to discover new disease variants, especially in monogenic disorders, frequency-based filtering has demonstrated to be a very useful tool (9). The rationale is as follows: variants that are relatively common in a control population (common variation) are likely benign (10), while rare variants (especially if they have functional consequences) found in multiple affected cases but absent in the control population are likely to cause disease (11–13). These filters search for genes or variants present in all (or most) affected individuals but in none (or very few) of the unaffected control individuals. Therefore, it seems clear that the availability of healthy controls is a decisive factor for the progress of discovery of new disease determinants.

From an historical perspective, the 1000 Genomes Project produced the first comprehensive catalogue of common human genetic variation (14). However, it is known that low frequency (with minor allele frequencies, MAF, under 5%) and rare (MAF under 0.5%) variants, typically population-specific (15), are poorly represented in such catalogue (14). Actually, recent studies have described a remarkable local component (16–18) and a high stratification level (19,20) in many rare variants with uncertain functional consequences. As a consequence of this, the risk of many diseases differs in distinct human populations according to their genetic backgrounds (21,22). In fact, the knowledge of the genetic variability of the local population has been revealed as a critical factor for the discovery of new disease variants (23). All these observations highlight the need for population-specific catalogues of genetic variation (24). However, only a few initiatives to study genetic variation at the population level have been carried out to date, which include a whole-genome sequence (WGS) study of 100 Malays (25), the *Genome of the Netherlands*, with low-resolution (∼13×) WGS data of 250 trio-families from across the entire country (15), the French-Canadians study of 109 exomes (26), the Medical Genome Project that produced a catalog of the healthy Spanish population with almost 270 exomes (23), the 3000 Finnish genomes (27) and the Icelandic population study of medium resolution (∼20×) WGS of 2636 individuals (28) or the high resolution (>30×) WGS of 1070 healthy Japanese individuals (29) and the recent genetic analysis of the Iranian population (30).

In spite of its recognized usefulness, large-scale sequencing projects of cohorts of local 'healthy' populations require expensive consortium-based projects to obtain a representative sample of the population targeted. Unfortunately, funding bodies that are prone to support research on diseases, tend to be, however, reluctant to fund projects that involve systematic sequencing of healthy individuals. In this scenario, a crowdsourcing strategy can provide a feasible alternative to traditional working schemas by organizing consortia that collect data from different groups that ultimately are collectively benefited of the sample size cooperatively obtained. Crowdsourcing is becoming a very popular strategy in biomedicine (31) and can be defined as 'the process of getting services, information, labor or ideas by outsourcing through an open call, especially through the Internet' (32). Recently some examples of crowdsourced research have demonstrated an increased accuracy in predicting breast cancer survival (33), response to drugs (34) or to toxic compounds (35) from both, clinical and genomic data, and show how 'crowdsourced data science challenges can achieve in months what would take years through conventional research approaches' (36).

MATERIALS AND METHODS

Subjects

The database contains detailed allelic frequencies corresponding to The MGP population, sequenced in the context of the Medical Genome Project (http://www.clinbioinfosspa.es/content/medical-genome-project), which includes 267 healthy, unrelated samples of Spanish origin (EGA, accession: EGAS00001000938), other healthy controls, patients of different diseases, accompanied in some cases of unrelated phenotypically healthy carriers. The sequences were contributed by different consortiums and projects, including groups from the Spanish Network for Research in Rare Diseases, CIBERER, results from the EnoD, (Undiagnosed Rare Diseases programme; https://www.ciberer.es/en/transversal-programmes/scientific-projects/undiagnosed-rare-diseases-programme-enod), the Project Genome 1000 Navarra (NAGEN 1000;

(https://www.nagen1000navarra.es/en), The RareGenomics (https://www.rare-genomics.com/) from Madrid, and other research groups and initiatives across Spain (37,38), which currently sum up a total of 2027 genomic and exomic sequences of unrelated Spanish individuals.

Testing sample locality

Ensuring the Spanish locality of the samples uploaded in the CSVS is key for the project. Here, we specifically developed a methodology to double-check the origin of each sample. Sequences belonging to different populations in the 1000 genomes project (14) were used to train a Machine Learning based decision model to discriminate Spanish samples from the rest of populations. Firstly, SNPs corresponding to the genomic regions shared by all the samples having a MAF > 0.01 were selected. Then, individual ancestry in 1000 genomes was estimated for 26 subpopulations using ADMIXTURE (39). Therefore, each individual is described by a vector of 26 features that correspond to the probabilities of belonging to any of the 26 subpopulations of 1000 genomes. Then, a machine learning binary classificatory was built using a well-known variant of the gradient boosting machine: extreme gradient boosting (*XGBoost*) (40) (see Supplementary Methods for details).

Testing sample kinship and outlier sample detection

A test to determine undesired samples based on their percentage of novel variants introduced in the database, either by excess (potential noisy sample) or by defect (close relative or individual already in the database), has also been used to populate the CSVS database. A leave-one-out cross-validation (LOOCV) strategy was to build a distribution of percentages of variants contributed by any single sample to the pool of variants present in the rest of the database. Samples were considered potential outliers if overpass 1.5 times the interquartile range from first and third quartile in the distribution obtained (see Supplementary Methods for details).

Construction of the reference imputation panel

Two alternative reference panels were created for comparison purposes that include the CSVS WGS variant panel composed of 228 samples plus: (i) the entire 1000G reference panel (CSVS+1000G) and (ii) exclusively the Spanish population (IBS subpopulation) contained in the 1000G panel (CSVS+IBS), using the *Minimac3* imputation tool (41). The four longest chromosomes (chromosome 1–4) were used to estimate the correlation between real and imputed genotypes (r^2 parameter) and assess the imputation accuracy (see Supplementary Methods for details).

RESULTS

The CSVS database

Figure 1A shows how data contributed by different genomic projects undergo different quality control steps, including an artifact and kinship detection tests and locality test, described above. Then the original VCFs are aggregated as counts of variants, binned by ICD10 (https://www.icd10data.com/) disease categories, and inserted in the CSVS database.

The CSVS interface

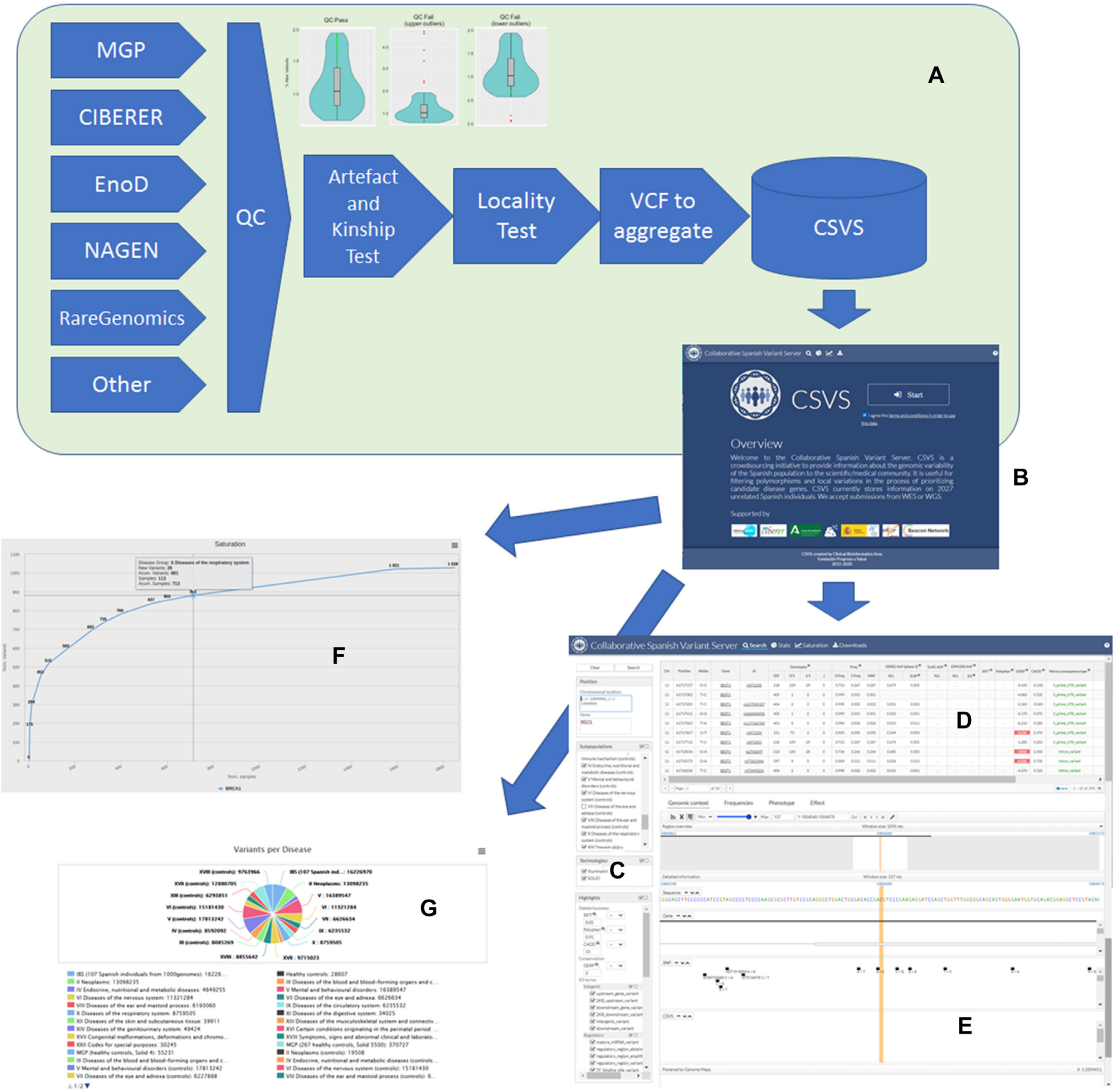
The initial screen (Figure 1B) requires the acceptance of the 'Terms and conditions for the use of the CSVS database' (http://csvs.babelomics.org/downloads/CSVSTermsAndConditions_use.pdf) before starting any operation. Once accepted, different options can be used.

The search option. This is the main option and allows querying the CSVS database. In the left panel (Figure 1C) queries can be done by gene symbol or by chromosomal regions. Also, one or several disease categories can be excluded, and variants can be highlighted using different types of scores (e.g. SIFT (42), Polyphen (43), CADD (44), Gerp (45)) as well as Sequence Ontology terms for the variation consequences.

The results of the query (Figure 1D) include a list of the positions for which variation has been found in the Spanish population along with complementary data as: chromosome, position, reference allele and alternative allele, allelic frequencies in the Spanish population, allelic frequencies in the 1000 genomes populations and in the EVS populations, impact and conservation indexes (SIFT, Polyphen, CADD, Gerp), the wort of the consequence types assigned to the mutation and the phenotypes, corresponding to known clinical information for the variants, extracted from ClinVar (46), COSMIC (47) and are annotated interactively on each query using the CellBase (48) webservices. Also a visualization of the variant in the genomic context is provided, based on the Genome Maps browser (49). Additionally, some extra detailed information can be found on the population frequencies observed for the variant, the phenotype or the effect.

Contact request. An interesting option is the *Contact request* button, offered for any variant in the query results panel, which is a local equivalent of a Matchmaker exchange service (50), extensively used to contact the original contributor of a specific sequence.

Saturation plots. Saturation plots (Figure 1F) provide an interesting perspective on the general conservation of the gene studied and, consequently on the possibilities of discovering new variants into it. Genes highly constrained to change will saturate soon and a relatively low number of individuals will capture most of the tolerated mutation the gene can handle, while unconstrained genes will present a still growing slope, meaning that there are still many variants that can potentially be discovered. Discovering a new variant in a saturated gene (constrained to change) can be more relevant than the same finding in a non-saturated gene (unconstrained). Saturation has a clear functional component, that can easily be revealed by enrichment analysis of the genes ranked by saturation. Thus, when genes are ranked by their relative saturation, enrichment analysis using **enrichR** (51) shows how highly saturated genes (constrained) are enriched in functional terms related to meiosis, cell signaling, proliferation and homeostasis, while the



Welcome to the Collaborative Spanish Variant Server. CSVS is a crowdsourcing initiative to provide information about the genomic variability of the Spanish population to the scientific/medical community. It is useful for filtering polymorphisms and local variations in the process of prioritizing candidate disease genes. CSVS currently stores information on 2027 unrelated Spanish individuals. We accept submissions from WES or WGS.

Figure 1. (**A**) data is contributed by different genomic projects and pass through different quality control steps including an artefact and kinship test (that detects upper outliers, with an unexpected high ratio of private variants, most likely errors, and lower outliers, that are duplicates or close kinship individuals) and locality test before being inserted in the database. (**B**) Initial CSVS page. (**C**) Query panel in the Search option. (**D**) List of variants found in the Spanish population within the selected region along with complementary information on impact, conservation, other's population frequencies and phenotype. (**E**) genomic browser that displays the selected variant in its genomic context. (**F**) Saturation plot. (**G**) Updated contents of the database.

less saturated (unconstrained) are more related to sensory perception, immune response and similar functionalities (see Supplementary Results and Supplementary Figure S1). Figure 2 depicts how genes with high and low saturation are distributed along the chromosomes. Interestingly, sex chromosomes seem to be enriched in low saturated genes.

Downloads and statistics. Partial or total downloads of the aggregated data are possible upon the reception of the corresponding data download agreement duly signed.

The **Stats** option provides an updated view of the content of the CSVS database.

The Spanish Genome Reference Panel (SGRP1.0)

Supplementary Figure S2 shows the accuracy of the two reference panels derived for imputation in the Spanish population. Both reference panels including the CSVS WGS reference outperformed the 1000 genomes reference. The imputation accuracy increases when variants in rare sites were in-

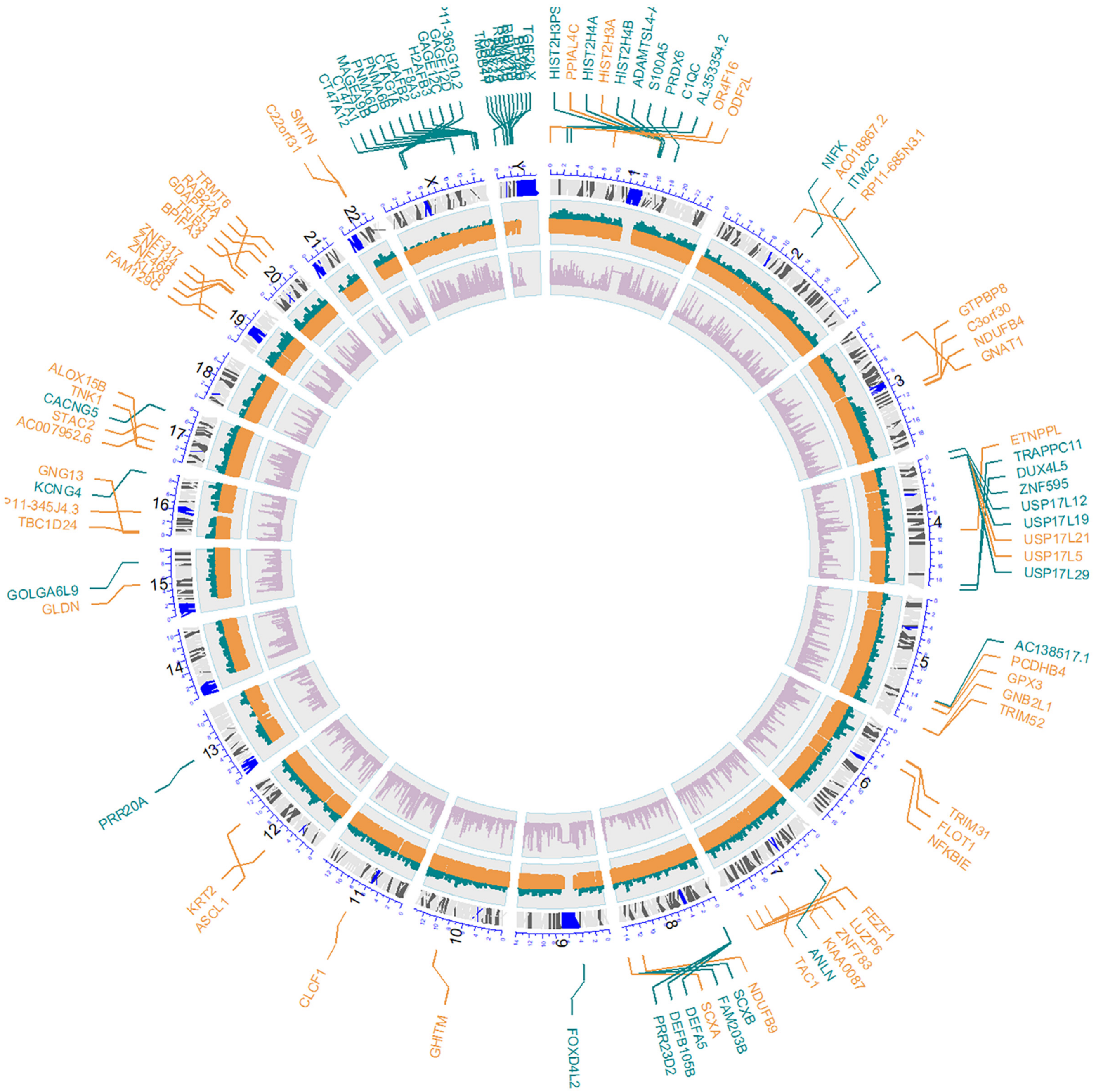


Figure 2. Circos plot showing the different genes with high saturation (orange) and low saturation (green) along the chromosomes, which were significantly enriched in functional terms in Supplementary Figure S1.

cluded (MAF > 0.005). The most realistic imputation panel includes CSVS and the IBS population of the 1000 genomes.

Variants of pharmacogenomic interest

Interindividual genetic variability in genes involved in drug-metabolizing enzymes and transporters have been linked to differences in the efficacy and toxicity of many medications: Moreover, genetic differences between human populations are becoming increasingly recognized as important factors accounting for interindividual variations in drug responsiveness (52,53). Approximately one-fifth of new drugs approved in the past years demonstrated differences in response across ethnic groups, leading to population-specific prescribing recommendations (54). In spite of the consensus about the existence of a relative homogeneity within European populations, population-specific differences in the Spanish population were recently reported (23). Using the individuals of the CSVS repository, we addressed how population-specific differences in those genes involved in drug Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) could affect in the rates and risks for drug inefficacy and/or adverse drug reactions in the Spanish population. We estimated the allele frequencies of a total of 142 pharmacogenetic variants described in the PharmGKB database (55) with pharmacogenetic clinical recommendations (PharmGKB variants level 1A and 1B) and a total of 40 of these were found to be polymorphic in the CSVS. When compared with the allele frequencies calculated from genetic data of 30 000 European non-Finnish individuals (gnomAD (56)), no relevant frequency differences between the general European population and the Spanish population were observed, being the most different rs2228001 (level 1B) in XPC gene, rs2108622 (level 1A) in CYP4F2 gene and rs3892097 (level 1A) In CYP2D6 gene (P-value $\leq 1 \times 10^{-10}$). Regarding the non-polymorphic variants, we observed that all of them are low-frequency variants (lower than 0.00065) and we do not expect to find a heterozygous individual due to the sample size in our repository (Supplementary Table S1).

Apart from the genetic variants already recommended to be implemented in the clinical setting, it was found that genetic variability with functional impact was governed by few high-frequency variants for some genes, but the functionality of the majority of pharmacogenes is dominated by rare genetic variants (57). In addition, local variability in these ADMET genes could also be very relevant for explaining a substantial part of the unexplained interindividual differences in drug response and toxicities at the population-specific level, so that it is mandatory to have available population-specific catalogs of these pharmavariants (mainly rare) to explore their contribution to predictions of drug response. To examine this, we studied the variability of the Spanish population captured by our repository in a total of 421 well-known pharmacogenes involved in drug pharmacokinetics and/or drug response (Supplementary Table S2). High-impact variants within those pharmacogenes were defined according to the Variant Effect Predictor (58) as those having having the following consequence types: frameshift, splice acceptor, splice donor, start lost, stop gained, stop lost, transcript ablation and transcript amplification. Additionally, deleterious missense variants categorized as deleterious by CONDEL (59) or having a LoFtool score (60) lower than the first quartile corresponding to the most intolerant variants.

As before, the same comparison with the corresponding European non-Finnish variants rendered a total of 318 high impact variants and 235 likely deleterious missense single nucleotide variants in the pharmacogenes studied. Interestingly, 18 (5.6%) high impact variants and 18 (7.6%) missense variants identified were present in our Spanish population while no heterozygotes were observed in these positions across ~30 000 healthy individuals of the European non-Finnish population. Also, a non-negligible percentage of private variation was observed in these genes encoding proteins involved in drug metabolism, transport, and response, and this information can be used to pinpoint relevant private genetic variants to be included in the design of population-specific pharmacogenetic genotyping arrays to be utilized in the implementation of pharmacogenetic diagnostics in the clinical setting (Supplementary Table S3).

CSVS Beacon

Since 2017, CSVS makes its genomic information discoverable through the GA4GH Beacon network (https://beacon-network.org/). In order to improve the performance of the CSVS Beacon API we set up an SQLite database specific for this purpose. Although CSVS stores data in 1-base it can respond to queries in both 1-base or 0-base (Beacon requests data in 0-base). A form to directly make Bacon-style queries is also available (http://ucscbeacon.clinbioinfosspa.es/).

DISCUSSION

The genetic variability of the local population is recognized as one of the most relevant factors in the discovery of new disease variants, especially in mendelian diseases (6,8,23). However, genomic data of healthy individuals belonging to the local population of interest are often scarce when not unavailable. The CSVS provides an original solution to this problem. The CSVS is a continuously growing resource that collects genomic or exomic sequences of the Spanish local population, no matter whether these come from healthy or diseased individuals. The main objective is using the repository as a pseudo-control population for finding new disease-causing variants and genes, with the idea that 'disease A is a healthy control for disease B'. Despite gene pleiotropy cannot be completely ruled out, data are binned at higher disease ICD10 categories, where this gene property can be considered negligible. Actually, resources like Disgenet (61) can be used in case of doubt, and will be incorporated to automatically exclude the proper disease categories, in future CSVS versions. Since the collection of population-specific genomic data from individuals with different diseases are easier to collect than those from healthy donors, CSVS provides an example for the construction of population-specific pseudo-control repositories by means of crowdsourcing (31). Moreover, the CSVS Beacon and the *Contact request* option makes of CSVS a tool with high potential of discoverability. Thus, CSVS sets the ground and it is an example for future federated European infrastructures (62).

DATA AVAILABILITY

CSVS is an open resource available at http://csvs.babelomics.org/.

The CSVS code, as well as the code of the different tests used is available in the corresponding github repository: https://github.com/babelomics/CSVS.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The Spanish Exome Crowdsourcing Consortium is a *de facto* consortium currently composed by: Fátima Al-Shahrour, Rafael Artuch, Javier Benitez, Luis Antonio Castaño, Ignacio del Castillo, Aitor Delmiro, Carmina Espinos, Roser González, Daniel Grinberg, Encarnación

Guillén, Pablo Lapunzina, Esther Lopez, Ramón Martí, Montserrat Milá, José Mª Millán, Virginia Nunes, Francesc Palau, Belen Perez, Luis Pérez Jurado, Rosario Perona, Aurora Pujol, Feliciano Ramos, Antonia Ribes, Jordi Rosell, Eulalia Rovira, Jordi Surrallés, Isabel Tejada and Magdalena Ugarte.

FUNDING

Spanish Ministry of Economy and Competitiveness [SAF2017-88908-R, PT17/0009/0006 to J.D.; PI19/00321 and CIBERER ACCI-06/07/0036 to C.A., PI14-948, PI17-1659 and CIBERER ACCI-06/07/0036 to M.A.M.P.]; Regional Government of Madrid, RAREGenomics-CM [B2017/BMD-3721 to C.A. and B2017/BMD3721 to M.A.M.P.]; all co-funded with European Regional Development Funds (ERDF) as well as EU H2020-INFRADEV-1-2015-1 ELIXIR-EXCELERATE [676559]; University Chair UAM-IIS-FJD of Genomic Medicine and the Ramon Areces Foundation also supported this work. Funding for open access charge: Spanish Ministry of Economy and Competitiveness [SAF2017-88908-R].

Conflict of interest statement. None declared.

REFERENCES

1. Mardis,E.R. (2017) DNA sequencing technologies: 2006–2016. *Nat. Protoc.*, **12**, 213.
2. 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.
3. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
4. Fu,W., O'Connor,T.D., Jun,G., Kang,H.M., Abecasis,G., Leal,S.M., Gabriel,S., Rieder,M.J., Altshuler,D., Shendure,J. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
5. Boycott,K.M., Rath,A., Chong,J.X., Hartley,T., Alkuraya,F.S., Baynam,G., Brookes,A.J., Brudno,M., Carracedo,A., den Dunnen,J.T. *et al.* (2017) International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.*, **100**, 695–705.
6. Boycott,K.M., Vanstone,M.R., Bulman,D.E. and MacKenzie,A.E. (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.*, **14**, 681–691.
7. Wenger,A.M., Guturu,H., Bernstein,J.A. and Bejerano,G. (2017) Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet. Med.*, **19**, 209.
8. Boycott,K.M., Hartley,T., Biesecker,L.G., Gibbs,R.A., Innes,A.M., Riess,O., Belmont,J., Dunwoodie,S.L., Jojic,N., Lassmann,T. *et al.* (2019) A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell*, **177**, 32–37.
9. Rehm,H.L., Bale,S.J., Bayrak-Toydemir,P., Berg,J.S., Brown,K.K., Deignan,J.L., Friez,M.J., Funke,B.H., Hegde,M.R. and Lyon,E. (2013) ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.*, **15**, 733.
10. 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.
11. Ng,S.B., Turner,E.H., Robertson,P.D., Flygare,S.D., Bigham,A.W., Lee,C., Shaffer,T., Wong,M., Bhattacharjee,A. and Eichler,E.E. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
12. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.
13. Ng,S.B., Bigham,A.W., Buckingham,K.J., Hannibal,M.C., McMillin,M.J., Gildersleeve,H.I., Beck,A.E., Tabor,H.K., Cooper,G.M., Mefford,H.C. *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, **42**, 790–793.
14. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
15. The_Genome_of_the_Netherlands_Consortium. (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.
16. Nelson,M.R., Wegmann,D., Ehm,M.G., Kessner,D., St Jean,P., Verzilli,C., Shen,J., Tang,Z., Bacanu,S.A., Fraser,D. *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**, 100–104.
17. Kryukov,G.V., Pennacchio,L.A. and Sunyaev,S.R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.
18. Marth,G.T., Yu,F., Indap,A.R., Garimella,K., Gravel,S., Leong,W.F., Tyler-Smith,C., Bainbridge,M., Blackwell,T., Zheng-Bradley,X. *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.
19. Mathieson,I. and McVean,G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.
20. Moreno-Estrada,A., Gravel,S., Zakharia,F., McCauley,J.L., Byrnes,J.K., Gignoux,C.R., Ortiz-Tello,P.A., Martinez,R.J., Hedges,D.J., Morris,R.W. *et al.* (2013) Reconstructing the population genetic history of the Caribbean. *PLoS Genet.*, **9**, e1003925.
21. Corona,E., Chen,R., Sikora,M., Morgan,A.A., Patel,C.J., Ramesh,A., Bustamante,C.D. and Butte,A.J. (2013) Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genet.*, **9**, e1003447.
22. Fernandez,R.M., Bleda,M., Luzon-Toro,B., Garcia-Alonso,L., Arnold,S., Sribudiani,Y., Besmond,C., Lantieri,F., Doan,B., Ceccherini,I. *et al.* (2013) Pathways systematically associated to Hirschsprung's disease. *Orphanet. J. Rare. Dis.*, **8**, 187.
23. Dopazo,J., Amadoz,A., Bleda,M., Garcia-Alonso,L., Alemán,A., García-García,F., Rodriguez,J.A., Daub,J.T., Muntané,G. and Rueda,A. (2016) 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. *Mol. Biol. Evol.*, **33**, 1205–1218.
24. Bustamante,C.D., Burchard,E.G. and De la Vega,F.M. (2011) Genomics for the world. *Nature*, **475**, 163–165.
25. Wong,L.P., Ong,R.T., Poh,W.T., Liu,X., Chen,P., Li,R., Lam,K.K., Pillai,N.E., Sim,K.S., Xu,H. *et al.* (2013) Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.*, **92**, 52–66.
26. Casals,F., Hodgkinson,A., Hussin,J., Idaghdour,Y., Bruat,V., de Maillard,T., Grenier,J.C., Gbeha,E., Hamdan,F.F., Girard,S. *et al.* (2013) Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLos Genet.*, **9**, e1003815.
27. Lim,E.T., Wurtz,P., Havulinna,A.S., Palta,P., Tukiainen,T., Rehnstrom,K., Esko,T., Magi,R., Inouye,M., Lappalainen,T. *et al.* (2014) Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.*, **10**, e1004494.
28. Gudbjartsson,D.F., Helgason,H., Gudjonsson,S.A., Zink,F., Oddson,A., Gylfason,A., Besenbacher,S., Magnusson,G., Halldorsson,B.V., Hjartarson,E. *et al.* (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435–444.
29. Nagasaki,M., Yasuda,J., Katsuoka,F., Nariai,N., Kojima,K., Kawai,Y., Yamaguchi-Kabata,Y., Yokozawa,J., Danjoh,I., Saito,S. *et al.* (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.*, **6**, 8018.
30. Fattahi,Z., Beheshtian,M., Mohseni,M., Poustchi,H., Sellars,E., Nezhadi,S.H., Amini,A., Arzhangi,S., Jalalvand,K. and Jamali,P. (2019) Iranome: a catalog of genomic variations in the Iranian population. *Hum. Mutat.*, **40**, 1968–1984.

31. Khare,R., Good,B.M., Leaman,R., Su,A.I. and Lu,Z. (2015) Crowdsourcing in biomedicine: challenges and opportunities. *Brief. Bioinform.*, **17**, 23–32.
32. Estellés-Arolas,E. and González-Ladrón-de-Guevara,F. (2012) Towards an integrated crowdsourcing definition. *J Inf Sci*, **38**, 189–200.
33. Margolin,A.A., Bilal,E., Huang,E., Norman,T.C., Ottestad,L., Mecham,B.H., Sauerwine,B., Kellen,M.R., Mangravite,L.M., Furia,M.D. *et al.* (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.*, **5**, 181re1.
34. Plenge,R.M., Greenberg,J.D., Mangravite,L.M., Derry,J.M., Stahl,E.A., Coenen,M.J., Barton,A., Padyukov,L., Klareskog,L., Gregersen,P.K. *et al.* (2013) Crowdsourcing genetic prediction of clinical utility in the rheumatoid arthritis responder challenge. *Nat. Genet.*, **45**, 468–469.
35. Eduati,F., Mangravite,L.M., Wang,T., Tang,H., Bare,J.C., Huang,R., Norman,T., Kellen,M., Menden,M.P., Yang,J. *et al.* (2015) Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotech.*, **33**, 933–940.
36. Davis,S., Button-Simons,K., Bensellak,T., Ahsen,E.M., Checkley,L., Foster,G.J., Su,X., Moussa,A., Mapiye,D., Khoo,S.K. *et al.* (2019) Leveraging crowdsourcing to accelerate global health solutions. *Nat. Biotechnol.*, **37**, 848–850.
37. Gallego-Martinez,A. and Lopez-Escamez,J.A. (2019) Genetic architecture of Meniere's disease. *Hear. Res.*, 107872.
38. Gui,H., Schriemer,D., Cheng,W.W., Chauhan,R.K., Antiňolo,G., Berrios,C., Bleda,M., Brooks,A.S., Brouwer,R.W. and Burns,A.J. (2017) Whole exome sequencing coupled with unbiased functional analysis reveals new Hirschsprung disease genes. *Genome Biol.*, **18**, 48.
39. Alexander,D.H., Novembre,J. and Lange,K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
40. Chen,T. and Guestrin,C. (2016) In: *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.
41. Das,S., Forer,L., Schönherr,S., Sidore,C., Locke,A.E., Kwong,A., Vrieze,S.I., Chew,E.Y., Levy,S. and McGue,M. (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284.
42. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
43. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **76**, 7.20.21–27.20.41.
44. 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.
45. 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.
46. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2017) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
47. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2018) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, **47**, D941–D947.
48. 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.
49. Medina,I., Salavert,F., Sanchez,R., de Maria,A., Alonso,R., Escobar,P., Bleda,M. and Dopazo,J. (2013) Genome Maps, a new generation genome browser. *Nucleic Acids Res.*, **41**, W41–W46.
50. Philippakis,A.A., Azzariti,D.R., Beltran,S., Brookes,A.J., Brownstein,C.A., Brudno,M., Brunner,H.G., Buske,O.J., Carey,K. and Doll,C. (2015) The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.*, **36**, 915–921.
51. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
52. Kubo,K., Ohara,M., Tachikawa,M., Cavallari,L., Lee,M., Wen,M., Scordo,M., Nutescu,E., Perera,M. and Miyajima,A. (2017) Population differences in S-warfarin pharmacokinetics among African Americans, Asians and whites: their influence on pharmacogenetic dosing algorithms. *Pharmacogenomics J.*, **17**, 494–500.
53. Meyer,U.A. (2004) Pharmacogenetics–five decades of therapeutic lessons from genetic diversity. *Nat. Rev. Genet.*, **5**, 669–676.
54. Ramamoorthy,A., Pacanowski,M., Bull,J. and Zhang,L. (2015) Racial/ethnic differences in drug disposition and response: review of recently approved drugs. *Clin. Pharmacol. Ther.*, **97**, 263–273.
55. Barbarino,J.M., Whirl-Carrillo,M., Altman,R.B. and Klein,T.E. (2018) PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **10**, e1417.
56. Koch,L. (2020) Exploring human genomic diversity with gnomAD. *Nat. Rev. Genet.*, **21**, 448–448.
57. Ingelman-Sundberg,M., Mkrtchian,S., Zhou,Y. and Lauschke,V.M. (2018) Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum. Genomics*, **12**, 26.
58. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
59. González-Pérez,A. and López-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
60. Fadista,J., Oskolkov,N., Hansson,O. and Groop,L. (2017) LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, **33**, 471–474.
61. Piñero,J., Queralt-Rosinach,N., Bravo,À., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028.
62. Saunders,G., Baudis,M., Becker,R., Beltran,S., Béroud,C., Birney,E., Brooksbank,C., Brunak,S., Van den Bulcke,M. and Drysdale,R. (2019) Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.*, **20**, 693–701.