Published online 3 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1039–D1048***
doi: 10.1093/nar/gkx1039

VarCards: an integrated genetic and clinical database for coding variants in the human genome

Jinchen Li[1,2,3,†], Leisheng Shi[1,†], Kun Zhang[1], Yi Zhang[1], Shanshan Hu[1], Tingting Zhao[1], Huajing Teng[4], Xianfeng Li[3,4], Yi Jiang[3], Liying Ji[1,*] and Zhongsheng Sun[1,4,*]

[1]Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang 325025, China, [2]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan 410078, China, [3]Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan 410078, China and [4]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China

Received August 15, 2017; Revised October 16, 2017; Editorial Decision October 17, 2017; Accepted October 18, 2017

ABSTRACT

A growing number of genomic tools and databases were developed to facilitate the interpretation of genomic variants, particularly in coding regions. However, these tools are separately available in different online websites or databases, making it challenging for general clinicians, geneticists and biologists to obtain the first-hand information regarding some particular variants and genes of interest. Starting with coding regions and splice sties, we artificially generated all possible single nucleotide variants (n = 110 154 363) and cataloged all reported insertion and deletions (n = 1 223 370). We then annotated these variants with respect to functional consequences from more than 60 genomic data sources to develop a database, named VarCards (http://varcards.biols.ac.cn/), by which users can conveniently search, browse and annotate the variant- and gene-level implications of given variants, including the following information: (i) functional effects; (ii) functional consequences through different *in silico* algorithms; (iii) allele frequencies in different populations; (iv) disease- and phenotype-related knowledge; (v) general meaningful gene-level information; and (vi) drug–gene interactions. As a case study, we successfully employed VarCards in interpretation of *de novo* mutations in autism spectrum disorders. In conclusion, VarCards provides an intuitive interface of necessary information for researchers to prioritize candidate variations and genes.

INTRODUCTION

In recent decades, next-generation sequencing (NGS) has resulted in a revolution in the rapid detection of large amounts of sequence variants in the human genome (1). Among NGS technologies, whole-exome sequencing is probably the most commonly used for prioritizing candidate mutations and genes underlying Mendelian, complex and undiagnosed genetic diseases as well as human cancers. However, only a small subset of functionally relevant variants, particularly in coding regions, are potentially associated with a given disease (2). To better interpret human variants for identifying disease-causing variations, a growing number of databases and tools have been successively developed (3). In addition, several organizations, such as The American College of Medical Genetics and Genomics, presented their guidelines for evaluating the causality between genetic variants and human diseases based on known genetic and clinical data sources, and functional studies (4–6).

Datasets from the 1000 Genomes Project (7), National Institutes of Health Heart, Lung and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (8), Exome Aggregation Consortium (ExAC) (9,10) and the Genome Aggregation Database (gnomAD) (9) provided large-scale reference genetic variations for multiple populations, which are critical for filtering out common variants that are less likely be disease-causing, allowing to identify rare variants. Additionally, a variety of *in silico* algorithms and tools, such as SIFT (11), PolyPhen2 (12) and MutationTaster (13), were developed to predict whether missense variants are damaging to the protein function or structure. To facilitate the process of querying missense predictions, the dbNSFP database, which has been widely used in the research community, was developed by integrating different algorithms and is constantly being updated (14–16). To date, dbNSFP v3.0 has integrated more than 20 algorithms (16). Wang and colleagues developed a functional annotation pipeline

*To whom correspondence should be addressed. Tel: +86 010 64864959; Fax: +86 10 84504120; Email: sunzsbiols@126.com
Correspondence may also be addressed to Liying Ji. Email: jiliying15@163.com
†These authors contributed equally to the paper as first authors.

named ANNOVAR for genetic variant annotation (17). To further facilitate web-based personal genome annotation, they also developed a web server called wANNOVAR (18). Through the command-line tool and web server, users can effectively analyze genomic variants (19). Furthermore, several other variant- and gene-level databases, such as InterVar (20), ClinVar (21), InterPro (22), denovo-db (23), COSMIC (24), OMIM (25), Ensembl (26), GenBank (27), The UCSC Genome Browser (28), UniProt (29), Gene Ontology (30) and DGIdb (31), were successfully developed to assist in the interpretation of genetic variants and prioritization of disease candidate genes.

Despite the great progress of these databases and tools in genomics and medical genetics, these resources are presented separately on various online websites, which cannot simultaneously perform both functional prediction and its clinical implication. It is also both tedious and time consuming to obtain first-hand core information regarding some variants and genes of interest. Therefore, there is a necessity to develop a convenient database through which users can retrieve general genetic and clinical knowledge for given variants in one integrated online database. To address this need, we developed VarCards, which provides an intuitive graphical user interface for querying genetic and clinical data regarding coding variants in the human genome.

MATERIALS AND METHODS

Variant-level data source

Based on definitions of transcripts from RefSeq (32), CCDS (33), UCSC known Gene (34) and Ensembl Gene (35) with the reference human genome (CRCh37/hg19), we retrieved the coding regions and splicing sites (2-base pairs of the splicing junctions), and artificially generated any possible single nucleotide variants (SNVs) of these regions. For example, for a given genomic position, if the nucleotide was cytosine (C) in the reference human genome, we artificially generated three SNVs: cytosine to thymidine (C>T), cytosine to guanine (C>G) and cytosine to adenine (C>A). In addition, we cataloged all reported insertions and deletions (INDELs) sourced from the general population variant database gnomAD (9), clinical variations database (ClinVar) (21), International Cancer Genome Consortium (ICGC) (36), Catalogue of Somatic Mutations In Cancer (COSMIC) (24) and *de novo* mutations database (denovo-db) (23).

We downloaded the allele frequencies of different populations from various human genetic variation databases, including gnomAD (variants of 15 496 genomes and 123 136 exomes from seven populations worldwide) (9), ExAC (60 706 exomes from seven populations) (9,10), ESP (6503 exomes from European Americans and African Americans) (8), 1000 Genomes Project (genomic data for 2504 individuals from five populations) (7), Kaviar genomic variant database (integrated variants from 35 projects encompassing 13 200 genomes and 64 600 exomes) (37), Haplotype Reference Consortium (HRC, 64 976 haplotypes from individuals with predominantly European ancestry) (38) and CG69 (69 individuals with complete genomes) (39). In addition, we extracted variant and related diseases or phenotypes information from InterVar (20), ClinVar (21), denovo-db (23), COSMIC (24), ICGC (36) and GWAS Catalog (40). Furthermore, we obtained predictive scores and pathogenicity consequences of missense variants from 23 *in silico* algorithms or tools, including SIFT (41,42), PolyPhen2_HDIV (12), PolyPhen2_HVAR (12), LRT (43), MutationTaster (44), MutationAssessor (45), FATHMM (46), PROVEAN (47), MetaSVM (48), MetaLR (48), VEST3 (49), M-CAP (50), CADD (51), GERP++ (52), DANN (53), fathmm-MKL (54), Eigen (55), GenoCanyon (56), fitCons (57), PhyloP (58), PhastCons (59), SiPhy (60) and REVEL (61). In particular, the predicted damaging scores and functional consequences of the 23 algorithms were sourced from dbNSFP v3.0 database (16). Finally, some genomic features, such as the protein domain from InterPro (22) and repeat segment from segmental duplication database (62), were also cataloged.

Gene-level data source

Gene-level basic information and functional information were sourced from UniProt (29), NCBI Gene (63) and BioSystems (64). The Gene ontology (GO) terms from the Gene Ontology Consortium, protein domains from InterPro (22) and protein–protein interactions from InBio Map (65) were also integrated. We collected the genic intolerance score of each gene from three studies: (i) the residual variation intolerance score (RVIS) from Petrovski *et al*. (66), (ii) loss-of-function (LoF) intolerance (gene intolerance score based on loss-of-function variants in 60 706 individuals) from Fadista *et al*. (67) and (iii) the heptanucleotide context intolerance score from Aggarwala *et al*. (68). In addition, data for genes associated with different diseases or phenotypes were curated from Online Mendelian Inheritance in Man (OMIM) (25), ClinVar (21), Human Phenotype Ontology (HPO) (69) and MGI (mammalian phenotype from mouse genome informatics) (70). Furthermore, we collected gene expression data for various tissues from the genotype-tissue Expression Project (GTEx) (71) and the protein subcellular map from the Human Protein Atlas (72). To present an overall view of gene expression levels, the means and standard deviations across 31 primary tissues and 54 secondary tissues for each gene were calculated. Protein sequences across 21 species were sourced from HomoloGene at NCBI. Finally, the data for drug–gene interactions and gene druggability were sourced from the Drug-gene Interaction Database (DGIdb) (31) to assist with the precision medicine.

Combination and annotation

Similar to our previous studies (73–77), we performed the command line tool, ANNOVAR (17) to annotate all SNVs and INDELs with respect to variant-level data sources, including the following information: (i) functional effects of variants; (ii) functional prediction of missense mutations by 23 predictive algorithms; (iii) allele frequencies in different populations; (iv) reported variants in different disease- and phenotype-related databases; and (v) some other genome features, such as CytoBand. The gene-level data sources were integrated into the database by using our in-house script. LoF variants, including stop-gain, stop-loss, splic-

ing sites SNVs, and frameshift indels, and deleterious missense SNVs with an allele frequency of <0.0001 based on gnomAD (9) were regarded as potential extreme variants. Deleterious missense SNVs were predicted using a combination of 23 computational methods.

Database construction

A user-friendly web interface, VarCards (http://159.226.67.237/sun/varcards/ or http://varcards.biols.ac.cn/), was developed by combining jQuery with a PHP-based web framework CodeIgniter, supported by versatile browsing and searching functionalities, as our previous databases and web servers (73,76–78). Annotation information was stored in either MySQL database or flat files. Academic users can access genetic data or extended analysis results freely through the web interface with no requirement for the use of a username or password.

De novo mutation (DNM) annotations in a case study

In the current case study, DNMs from 2508 autism spectrum disorder (ASD) cases and 1911 unaffected siblings were sourced from the Simons Simplex collection (SSC) (79,80). VarCards was used to annotate all DNMs, and only coding and splicing site DNMs were retained for further analysis. Deleterious missense mutations were predicted by the combination of REVEL (61) and VEST3 (49). ASD candidate genes were sourced from ClinVar (21), OMIM (25) and SFARI Gene (81). The online tool DAVID (82) was employed to perform functional enrichment analysis.

RESULTS AND WEB INTERFACE

To assess the clinical significance of given variants, such as DNMs from sporadic families, homozygous variants from consanguineous families, and cosegregated heterozygous variants from multigeneration families, various genomic, genetic and clinical evidences must be systematically evaluated. VarCards provides integrated web interfaces to conveniently search, browse and annotate the variant- and gene-level implications of any given coding variants (Figure 1 and Table 1). For variant-level implications, users can obtained first-hand information, including: (i) whether this variant has been reported to be associated with diseases; (ii) allele frequencies in different populations; (iii) functional effects of transcript and protein levels; and (iv) deleteriousness predicted by various algorithms. In addition, VarCards provides general gene-level implications, such as basic information, genic intolerance, gene function, gene-related diseases, gene expression and target drug to assist users with prioritizing candidate genes.

Variant-level implications

Overall, 110 154 363 SNVs and 1 223 370 INDELs in coding regions or splicing sites are included in VarCards. Both general and advanced query interfaces are provided to access the detailed annotation data of these variants. Common search terms, such as genomic position and regions, gene symbol, and nucleic acid changes in a certain gene or transcript, are supported to allow users to quickly analyze variants of interest. Search results return as a page contained a table, which display all variant-level implications (Figure 2), including (i) functional effects at the protein and transcript levels in all four gene annotation systems; (ii) the predicted damaging scores and functional consequences of missense variants of 23 *in silico* algorithms; (iii) allele frequencies of different populations in gnomAD (9), ExAC (10), ESP (8), 1000G (7), Kaviar (37), HRC (38) and CG69 (39); and (iv) disease- and phenotype-related knowledge in dbSNP (83), ClinVar (21), denovo-db (23), InterVar (20), COSMIC (24), ICGC (36), GWAS Catalog (40) and InterPro (22). The search results can be flexibly filtered by several properties, such as functional effects, damaging scores and allele frequencies. To meet the needs of different users, VarCards also allows users to perform advanced searches by pasting a list or uploading a file containing a mass of search terms with specific formats, including VCF4, ANNOVAR, genomic coordinates and genomic regions (Figure 2). We encourage users to specify data sources of interest for advanced searches. Notably, users can freely export query results as Excel or CSV files or copy them to the clipboard.

Gene-level implications

For genes containing given variants, VarCards provides seven specified panels to present gene-level implications (Figure 3). The 'Basic information' panel provides the following information: (i) primary information extracted from NCBI Gene (63), such as official gene name, synonyms and chromosomal location; (ii) a brief description of the cellular function of the protein encoded by the gene sourced from UniProt (29); and (iii) the genic intolerance score from three studies (66–68). The 'Gene function' panel provides information related to the protein entry name, length, subunit structure and domains, protein–protein interactions, GO terms and biological pathways. The 'Phenotype and disease' panel presents the reported disease-associated variants or genes from OMIM (25), ClinVar (21), denovo-db (23), MGI (70) and HPO (69). For the 'Gene expression' panel, the expression levels across 31 primary tissues and 54 secondary tissues are illustrated using a bar chart. For the 'Homology' panel, multiple alignments of protein amino acid sequences across 21 species are presented to assist the user in evaluating evolutionarily conserved sites. In addition, quick links to the interested gene at ENSEMBL (26) and TreeFam (84) are listed below the panel. Via the 'Variants in different populations' panel, users can inspect the number of variants with different functional effects and allele frequencies to preliminary estimate the general mutation rate in different populations. For the 'Drug-gene interaction' panel, the drug-gene interactions and gene druggability data sourced from DGIdb 2.0 (31) are presented in a real-time manner. Notably, only core information of gene-level implications is shown in VarCards. Links to external resources with detailed information are provided and can be easily accessed for academic users.

Browsing and customized annotations

Users can access variant- and gene-level implications via the browse function in the VarCards database. Moreover,

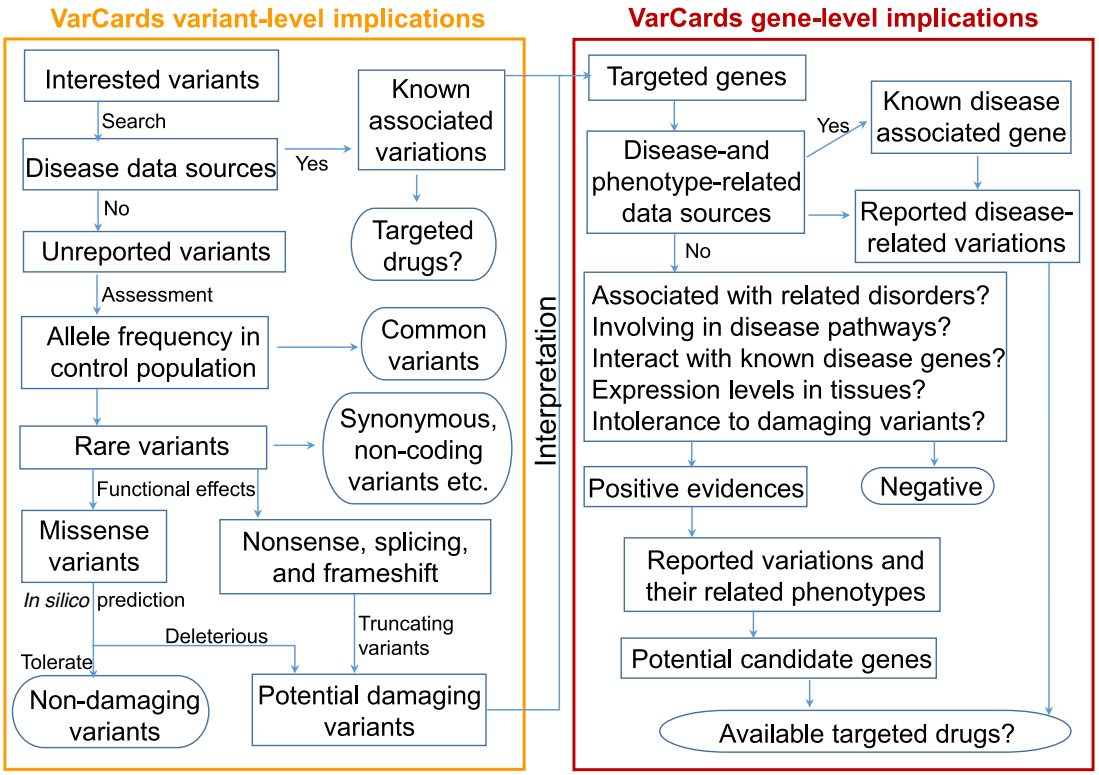

Figure 1. A general workflow of VarCards. A mass of genomic, genetic and clinical data sources should be systematically evaluated for prioritizing candidate variants and genes underlying genetic diseases. Various variant-level and gene-level implications have been integrated in VarCards.

Table 1. Summary of integrated data sources in VarCards

Category	Data source
Part one: variation-level implication	
Allele frequency	dbSNP, gnomAD, ExAC, 1000 Genomes, ESP, Kaviar, HRC, CG69
Missense prediction	SIFT, PolyPhen2_HDIV, PolyPhen2_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, MetaSVM, MetaLR, VEST, M-CAP, CADD, GERP++, DANN, fathmm-MKL, Eigen, GenoCanyon, fitCons, PhyloP, PhastCons, SiPhy, REVEL
Disease-related	InterVar, ClinVar, denovo-db, COSMIC, ICGC, GWAS Catalog,
Other data	RefSeq, InterPro, Segmental duplication
Part two: gene-level implication	
Basic information	UniProt, HomoloGene, Ensembl, NCBI Gene
Genic intolerance	RVIS, LoFtool, heptanucleotide context intolerance score
Gene function	UniProt, Gene Ontology, InterPro, InBio Map, BioSystems
Disease-related	OMIM, MGI, ClinVar, HPO
Gene expression	UniProt, GTEx, The Human Protein Atlas
Target drug	DGIdb

dbSNP, single nucleotide polymorphism database; gnomAD, genome aggregation database; ESP, NHLBI GO Exome Sequencing Project; HRC, haplotype reference consortium; Kaviar, Kaviar Genomic Variant Database, CG69, allele frequency in 69 human subjects sequenced by Complete Genomics. OMIM, online mendelian inheritance in man; MGI, mouse genome informatics; COSMIC, catalogue of somatic mutations in cancer; ICGC, international cancer genome consortium; HPO, human phenotype ontology; GTEx, Genotype-Tissue Expression.

VarCards implements a function for customized annotation by which users can conveniently annotate their variants in VCF or ANNOVAR formats. For different annotation needs, users can flexibly specify their data source of interest and cutoff of extreme variants including functional effects, allele frequencies and predicted damaging scores from any of the 23 *in silico* algorithms. After the variant file is uploaded, an annotate job will run in the backend, and when the job is completed, an email containing a download link for retrieving the results will be sent to the user.

Case study

DNMs play essential roles in the etiology of ASD, as shown in our previous studies (73–75). We cataloged 3397 and 2285 DNMs of 2508 ASD cases and 1911 unaffected siblings, respectively, from SSC (79,80) (Figure 4A). After removing noncoding variants, 2723 exonic DNMs retained in ASD, including 1114 DNMs that were presented in gnomAD (9) and 1609 DNMs that were novel variants. Consistent with previous studies (85), we found that the former category

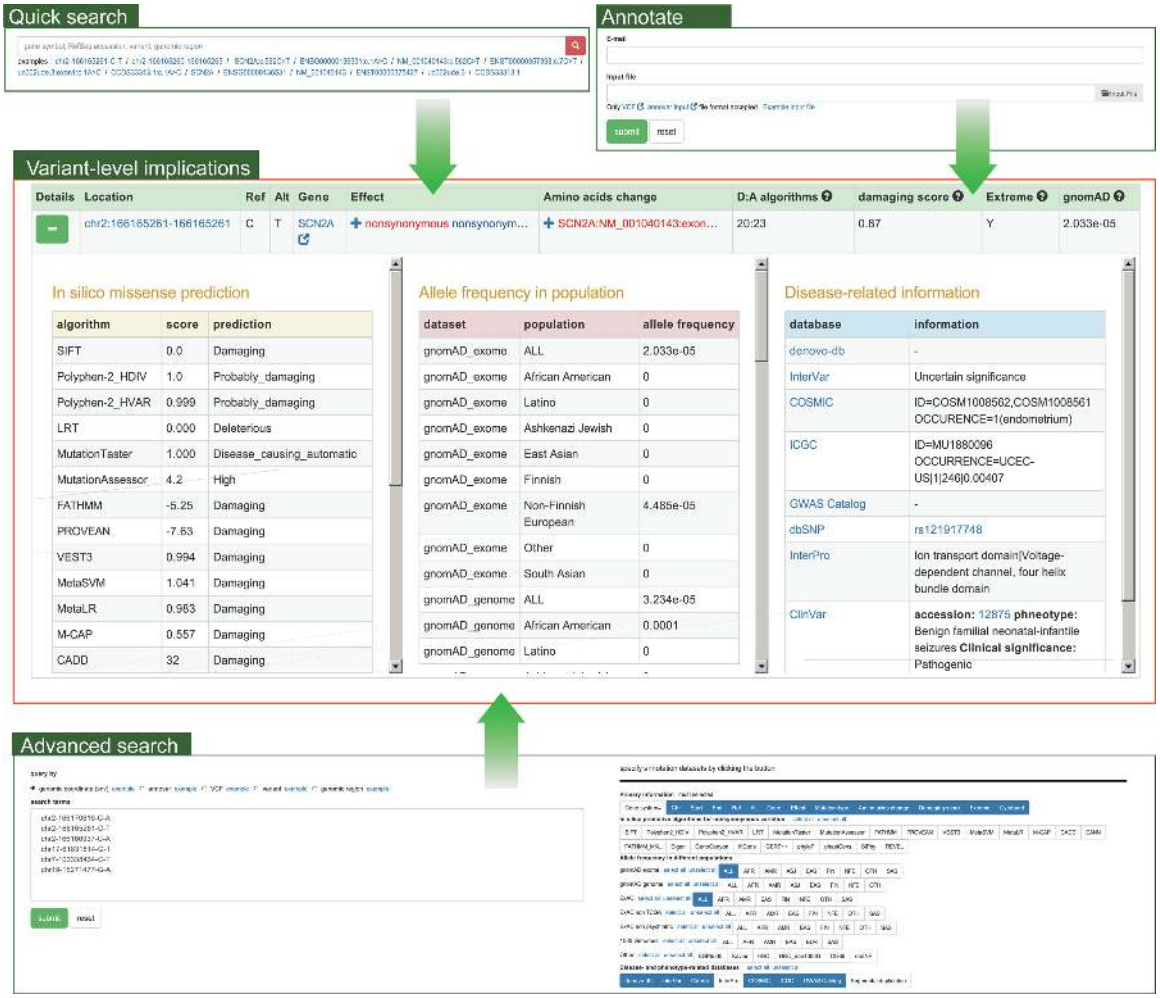

Figure 2. Snapshot of variant-level implications in VarCards. There are three approaches to access variant-level implications, including 'Quick search', 'Advanced search' and 'Annotate'. As an example, the results of a quick search for the variant 'SCN2A:c.562C>T', including functional effects at the transcript and protein levels, predicted the damaging severity of missense variants, allele frequencies in different populations and information in disease-related databases.

of DNMs did not show significant differences in patients with ASD when compared with siblings, whereas the mutation rate of LoF and predicted deleterious missense variants (i.e. putative functional DNMs) rather than tolerated missense, synonymous and nonframeshift INDELs (i.e. non-functional DNMs) of the novel DNM, was significantly higher than that in the control ($P < 0.05$, Figure 4B). We found that 600 (23.92%) patients with ASD harbored functional novel DNMs, 1015 (40.47%) patients harbored non-functional novel DNMs or other DNMs that presented in gnomAD, and 893 (35.61%) patients did not harbor any exonic DNMs (Figure 4C).

A previous study estimated that 45% of *de novo* LoF mutations and 13% of *de novo* missense mutations accounted for 9 and 12% of ASD cases, respectively (80). For the 600 cases with functional DNMs, we then prioritized their candidate genes based on clinical, genetic and biological information from VarCards and SFARI Gene (81). As a result, we found that 126 (21%) cases harbored functional DNMs in strong ASD candidate genes; 114 (19%) cases harbored functional DNMs in suggested ASD candidate genes; 41 (6.83%) cases harbored functional DNMs in genes associated with other neurodevelopmental disorders; 20 (3.33%) cases harbored functional DNMs in genes involved in known ASD pathways; and 299 (49.83%) cases harbored functional DNMs in genes without sufficient evidence supporting their identity as ASD candidate genes (Figure 4C). In total, 301 of 2508 (12%) ASD cases were found to have possible ASD risk DNMs and genes, which was higher than that reported in a previous study (86) in which *de novo* LoF mutations in ASD candidate genes accounted for 5% of patients. Finally, we found that candidate genes in the 301 cases were enriched in multiple biological processes in GO and KEGG pathways, including *in utero* embryonic development (GO:0001701, adjusted $P = 6.5 \times 10^{-4}$), circadian entrainment (hsa04713, adjusted $P = 6.3 \times 10^{-4}$), dopaminergic synapse (hsa04728, adjusted $P = 7.9 \times 10^{-4}$), glutamatergic synapse (hsa04724, adjusted $P = 0.001$), covalent chromatin modification (GO:0016569, adjusted $P = 0.024$), and the canonical Wnt signaling pathway (GO:0060070, adjusted $P = 0.032$).

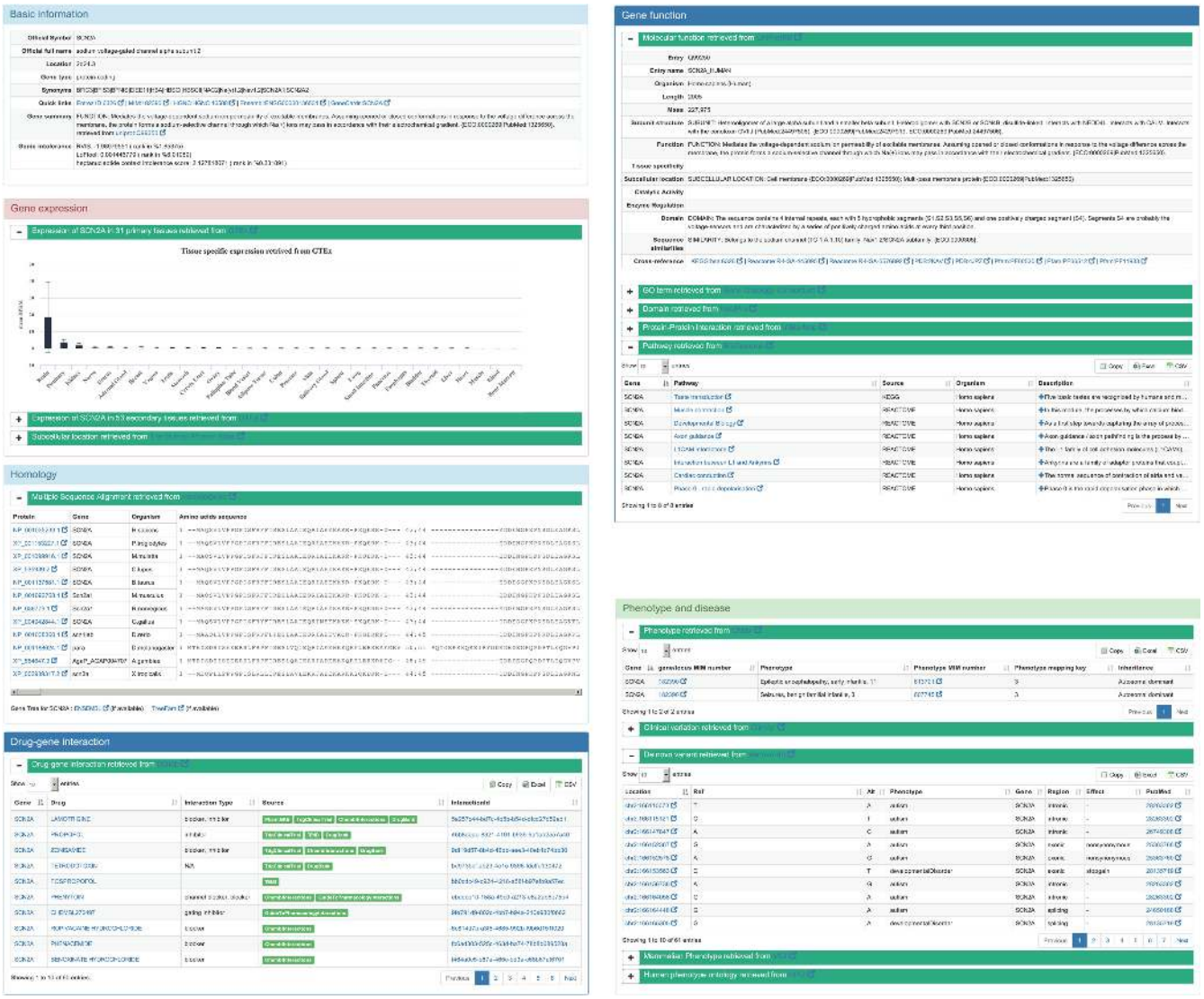

Figure 3. Snapshot of gene-level implications in VarCards. As example, the typical gene-level implications of the *SCN2A* gene are illustrated, including basic information, gene functions, associated phenotypes and diseases, gene expression, homology, variants in different population and drug–gene interactions.

DISCUSSION

Analysis of numerous variants detected by NGS technologies provide us unprecedented opportunities to prioritize clinically significant variations and genes underlying human genetic diseases (1). The major challenge is to interpret the close relationships between genotypes and phenotypes (87). Several scattering distributed genetic, genomic, and clinical data sources can assist in prioritizing disease-causing or disease-risk variations. In this study, we retrieved the most important core information from more than 60 genetic, genomic and clinical data sources and integrated them into the VarCards database, allowing clinicians, geneticists and biologists to conveniently analyze the first-hand general variant- and gene-level implications without having to search various websites or annotate variants by command line.

Despite of the advance of other available tools, VarCards shows significant differences. The dbNSFP (14–16) focus on functional effects of non-synonymous SNVs and their annotations. In addition, dbNSFP is a locally installed database and therefore doesn't provide any web interface to search, browse and annotate genetic variants, which is not easily accessible. The command-line tools, such as ANNOVAR (17) and WGSA (88), and the application program interface MyVariant.info (89), are developed to perform functional annotation of genetic variant. These tools can analyze mass of variants, but they does not provide any web interface and their results are not clearly visualized for the end-users. It is not convenient for researchers without sufficient bioinformatics skills. The web servers, such as wANNOVAR (18,19), VEP tool (90), Phenolyzer (91), wInterVar (20) and SeqMule (92) have been developed to analyze the genomic variants and predict functional consequences. However, results are reported in tab-delimited, CSV or VCF formats, which may not be intuitive enough for general clinicians, geneticists and biologists. In addition, these tools

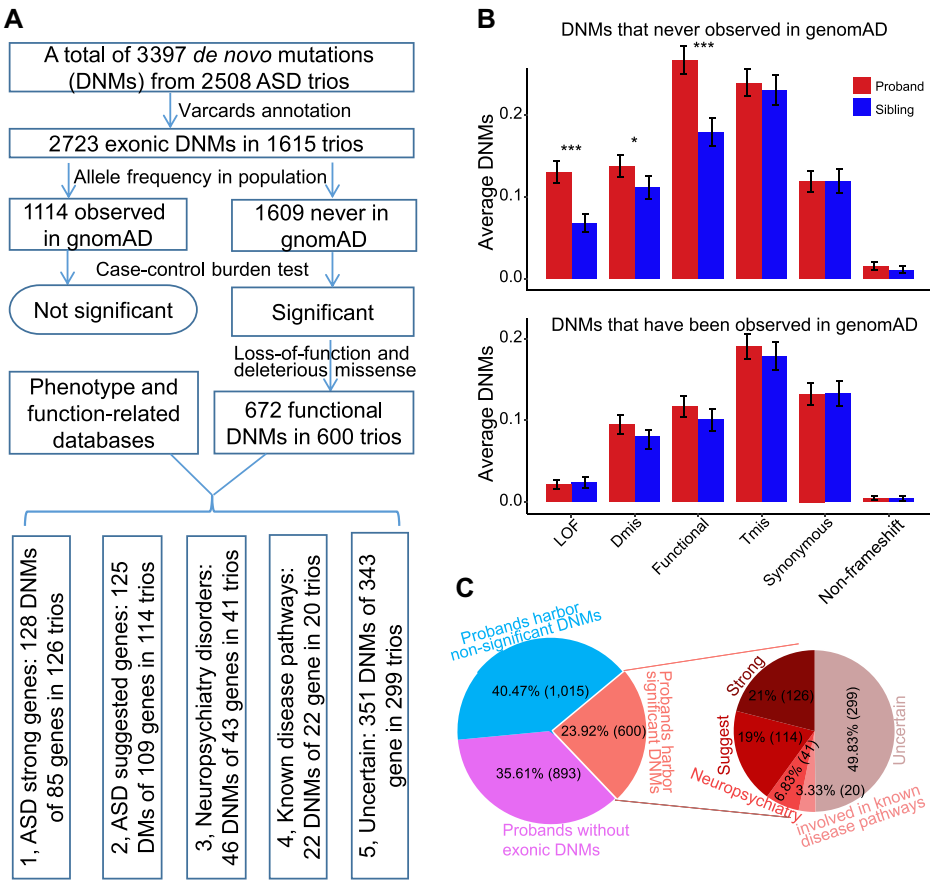

Figure 4. Case study of *de novo* mutations in ASD. (**A**) Workflow of data analysis. The LoF and predicted deleterious missense DNMs that had never been previously observed in the general population (based on gnomAD) and were found to be associated with ASD. These DNMs were identified in 600 ASD cases, accounting for 23.92% of ASD cohorts. We then classified these 600 ASD cases into five classes based on evidence of the associations of DNM-targeted genes with ASD, other neuropsychiatric disorders, and known disease pathways (see also in panel C). (**B**) Average number of DNMs classified by functional effects and their allele frequencies in gnomAD were compared between ASD and sibling. (**C**) Pie charts illustrating the percentages of ASD cases that harbored significant functional DNMs, non-functional DNMs, or non-coding DNMs. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

mainly focus on variant-level annotations and the gene-level information has not been annotated sufficiently in the web server. Moreover, when we query a small number of variants using web server, the results cannot be immediate shown because new submitted jobs usually need to be queued. Compared with these tools or web servers, VarCards not only provides similar annotation function, but also provides a more intuitive online interface for researchers without sufficient bioinformatics skills to accessibly obtain the first-hand genetic, genomic and clinical information of any coding variants within a short time.

To interpret whether a variant is significantly contributes to human disease, performing systematic and quantitative evaluations of positive and negative evidences regarding to its pathogenicity are urgently needed. There are several issues we would like to emphasize here. First, since the evidence of clinically significant variations from disease- and phenotype-related databases, such as ClinVar (21), COSMIC (24), OMIM (25) and HGMD (93), was mostly provided by individual studies or manually collected from scientific literature, different criteria and methodological biases certainly occurred in assessing the pathogenicity of genetic variants. Users should note the possibility of false-positive data in these data sources when interpreting known disease-contributing variations and genes (20,94). Second, VarCards provided prediction scores of 23 *in silico* algorithms for missense variants, and users should also note the potential limitations of specificity and sensitivity of these methods (16). Third, to reduce false-positive results in the identification of disease candidate genes, we encourage users to replicate their findings in more samples, perform functional experiment studies and carefully examine the clinical data of patients. Considering the complex processes of genetic testing, VarCards did not directly identify disease-causing variations, but provided various publically available data sources containing information on the given variants. It is expected that some users or groups will be able to flexibly prioritize candidate variations and genes based on their own criteria and genetic data according to the needs of their specific study.

VarCards will be updated continuously to provide the research community an up-to-date resource, not only update the data sources that we integrated, but also the integrated more new datasets that may be useful for medical genetics.

To improve the VarCards database in further updates, we encourage users to provide feedback with any suggestions or data sources. In the first phase, VarCards focused on variants in coding regions and splicing sites, accounting for 85% of disease-causing variations in Mendelian disorders (2). By reducing sequencing costs, new high-throughput technologies and analysis methods give us additional opportunities to investigate regulatory variants and functional elements in noncoding regions (95). However, the clinical interpretation of variants in noncoding regions still remains a major challenge (96). We plan to update the VarCards database in the next phase for rapid interpretation of noncoding variants. In summary, VarCards provides an intuitive interface of genetic, genomic, and clinical knowledge of coding variants, accelerating the prioritization of candidate variations and genes.

ACKNOWLEDGEMENTS

We thank the members of the Beijing Institute of Life Science, Chinese Academy of Sciences for their valuable discussion regarding this work.

FUNDING

National Key R&D Program of China [2016YFC0900400]; National Natural Science Foundation of China [31571301]. Funding for open access charge: National Key R&D Program of China [2016YFC0900400]; National Natural Science Foundation of China [31571301].
Conflict of interest statement. None declared.

REFERENCES

1. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
2. Rabbani,B., Tekin,M. and Mahdieh,N. (2013) The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.*, **59**, 5–15.
3. Biesecker,L.G. and Green,R.C. (2014) Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.*, **370**, 2418–2425.
4. 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) Supplementary Information for 'Guidelines for investigating causality of sequence variants in human disease'. *Nature*, **508**, 469–476.
5. Richards,S., Aziz,N., Bale,S., Bick,D., Das,S., Gastier-Foster,J., Grody,W.W., Hegde,M., Lyon,E., Spector,E. *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.
6. Matthijs,G., Souche,E., Alders,M., Corveleyn,A., Eck,S., Feenstra,I., Race,V., Sistermans,E., Sturm,M., Weiss,M. *et al.* (2016) Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.*, **24**, 2–5.
7. 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
8. Fu,W., O'Connor,T.D., Jun,G., Kang,H.M., Abecasis,G., Leal,S.M., Gabriel,S., Altshuler,D., Shendure,J., Nickerson,D.A. *et al.* (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
9. 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.
10. Karczewski,K.J., Weisburd,B., Thomas,B., Solomonson,M., Ruderfer,D.M., Kavanagh,D., Hamamsy,T., Lek,M., Samocha,K.E., Cummings,B.B. *et al.* (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.*, **45**, D840–D845.
11. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
12. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
13. Jian,X., Boerwinkle,E. and Liu,X. (2014) In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.*, **16**, 497–503.
14. Liu,X., Jian,X. and Boerwinkle,E. (2011) dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
15. Liu,X., Jian,X. and Boerwinkle,E. (2013) dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E2402.
16. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
17. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
18. Chang,X. and Wang,K. (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.*, **49**, 433–436.
19. Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.*, **10**, 1556–1566.
20. Li,Q., Wang,K., McPherson,J.D., Lyon,G.J., Wang,K., Quintáns,B., Ordóñez-Ugalde,A., Cacheiro,P., Carracedo,A., Sobrido,M.J. *et al.* (2017) InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.*, **100**, 267–280.
21. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
22. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.Y., Dosztanyi,Z., El-Gebali,S., Fraser,M. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
23. Turner,T.N., Yi,Q., Krumm,N., Huddleston,J., Hoekzema,K., Stessman,H.A.F., Doebley,A.L., Bernier,R.A., Nickerson,D.A. and Eichler,E.E. (2017) NAR Breakthrough Article denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.*, **45**, D804–D811.
24. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
25. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (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.
26. 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.
27. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
28. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
29. Wasmuth,E.V. and Lima,C.D. (2016) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, 1–12.
30. Thomas,P.D. (2017) Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium. *Nucleic Acids Res.*, **45**, D331–D338.

31. Wagner,A.H., Coffman,A.C., Ainscough,B.J., Spies,N.C., Skidmore,Z.L., Campbell,K.M., Krysiak,K., Pan,D., McMichael,J.F., Eldred,J.M. *et al.* (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
32. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
33. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M.J., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
34. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
35. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.
36. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabé,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
37. Glusman,G., Caballero,J., Mauldin,D.E., Hood,L. and Roach,J.C. (2011) Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*, **27**, 3216–3217.
38. McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A., Kang,H.M., Fuchsberger,C., Danecek,P., Sharp,K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
39. Carnevali,P., Baccash,J., Halpern,A.L., Nazarenko,I., Nilsen,G.B., Pant,K.P., Ebert,J.C., Brownley,A., Morenzoni,M., Karpinchyk,V. *et al.* (2012) Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.*, **19**, 279–292.
40. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
41. Ng,P.C. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
42. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
43. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
44. Schwarz,J.M., Rödelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
45. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
46. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L.A., Edwards,K.J., Day,I.N.M. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum. Mutat.*, **34**, 57–65.
47. Choi,Y., Sims,G.E., Murphy,S., Miller,J.R. and Chan,A.P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, **7**, e46688.
48. Dong,C., Wei,P., Jian,X., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
49. Carter,H., Douville,C., Stenson,P.D., Cooper,D.N. and Karchin,R. (2013) Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14**(Suppl. 3), S3.
50. Jagadeesh,K.A., Wenger,A.M., Berger,M.J., Guturu,H., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586.
51. Kircher,M. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
52. 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.
53. Quang,D., Chen,Y. and Xie,X. (2015) DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
54. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N.M., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
55. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
56. Lu,Q., Hu,Y., Sun,J., Cheng,Y., Cheung,K.-H. and Zhao,H. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, **5**, 10576.
57. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
58. Siepel,A., Pollard,K.S. and Haussler,D. (2006) New methods for detecting lineage-specific selection. *Lect. Notes Comput. Sci.*, **3909**, 190–205.
59. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.D.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
60. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*. **25**, i54–i62.
61. Ioannidis,N.M., Rothstein,J.H., Pejaver,V., Middha,S., McDonnell,S.K., Baheti,S., Musolf,A., Li,Q., Holzinger,E., Karyadi,D. *et al.* (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.*, **99**, 877–885.
62. She,X., Jiang,Z., Clark,R.A., Liu,G., Cheng,Z., Tuzun,E., Church,D.M., Sutton,G., Halpern,A.L. and Eichler,E.E. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.
63. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
64. Geer,L.Y., Marchler-Bauer,A., Geer,R.C., Han,L., He,J., He,S., Liu,C., Shi,W. and Bryant,S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.
65. Li,T., Wernersson,R., Hansen,R.B., Horn,H., Mercer,J., Slodkowicz,G., Workman,C.T., Rigina,O., Rapacki,K., Stærfeldt,H.H. *et al.* (2016) A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods*, **14**, 61–64.
66. Petrovski,S., Gussow,A.B., Wang,Q., Halvorsen,M., Han,Y., Weir,W.H., Allen,A.S. and Goldstein,D.B. (2015) The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.*, **11**, e1005492.
67. Fadista,J., Oskolkov,N., Hansson,O. and Groop,L. (2016) LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, **33**, 471–474.
68. Aggarwala,V. and Voight,B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.*, **48**, 349–355.
69. 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.
70. Eppig,J.T., Smith,C.L., Blake,J.A., Ringwald,M., Kadin,J.A., Richardson,J.E. and Bult,C.J. (2017) Mouse genome informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Methods Mol. Biol.*, **1488**, 47–73.

71. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
72. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
73. Li,J., Cai,T., Jiang,Y., Chen,H., He,X., Chen,C., Li,X., Shao,Q., Ran,X., Li,Z. *et al.* (2015) Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry*, **21**, 1–8.
74. Li,J., Wang,L., Guo,H., Shi,L., Zhang,K., Tang,M., Hu,S., Dong,S., Liu,Y., Wang,T. *et al.* (2017) Targeted sequencing and functional analysis reveal brain-size-related genes and their networks in autism spectrum disorders. *Mol. Psychiatry*, **22**, 1282–1290.
75. Li,J., Wang,L., Yu,P., Shi,L., Zhang,K., Sun,Z.S. and Xia,K. (2017) Vitamin D-related genes are subjected to significant de novo mutation burdens in autism spectrum disorder. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, **174**, 568–577.
76. Ran,X., Li,J., Shao,Q., Chen,H., Lin,Z., Sun,Z.S. and Wu,J. (2015) EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. *Nucleic Acids Res.*, **43**, D893–D899.
77. Li,J., Jiang,Y., Wang,T., Chen,H., Xie,Q., Shao,Q., Ran,X., Xia,K., Sun,Z.S. and Wu,J. (2015) mirTrios: an integrated pipeline for detection of de novo and rare inherited mutations from trios-based next-generation sequencing. *J. Med. Genet.*, **52**, 275–281.
78. Mao,F., Xiao,L., Li,X., Liang,J., Teng,H., Cai,W. and Sun,Z.S. (2016) RBP-var: a database of functional variants involved in regulation mediated by RNA-binding proteins. *Nucleic Acids Res.*, **44**, D154–D163.
79. Fischbach,G.D. and Lord,C. (2010) The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, **68**, 192–195.
80. Iossifov,I., O'roak,B.J., Sanders,S.J., Ronemus,M., Krumm,N., Levy,D., Stessman,H.A., Witherspoon,K., Vives,L., Patterson,K.E. *et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **13**, 216–221.
81. Abrahams,B.S., Arking,D.E., Campbell,D.B., Mefford,H.C., Morrow,E.M., Weiss,L.A., Menashe,I., Wadkins,T., Banerjee-Basu,S. and Packer,A. (2013) SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism*, **4**, 36.
82. Jiao,X., Sherman,B.T., Huang,D.W., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
83. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
84. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J.M., Guo,Y., Heacute;riché,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
85. Kosmicki,J.A., Samocha,K.E., Howrigan,D.P., Sanders,S.J., Slowikowski,K., Lek,M., Karczewski,K.J., Cutler,D.J., Devlin,B., Roeder,K. *et al.* (2017) Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.*, **49**, 504–510.
86. De Rubeis,S., He,X., Goldberg,A.P., Poultney,C.S., Samocha,K., Ercument Cicek,A., Kou,Y., Liu,L., Fromer,M., Walker,S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**, 209–215.
87. Lyon,G.J. and Wang,K. (2012) Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Med.*, **4**, 58.
88. Liu,X., White,S., Peng,B., Johnson,A.D., Brody,J.A., Li,A.H., Huang,Z., Carroll,A., Wei,P., Gibbs,R. *et al.* (2016) WGSA: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.*, **53**, 111–112.
89. Xin,J., Mark,A., Afrasiabi,C., Tsueng,G., Juchler,M., Gopal,N., Stupp,G.S., Putman,T.E., Ainscough,B.J., Griffith,O.L. *et al.* (2016) High-performance web services for querying gene and variant annotation. *Genome Biol.*, **17**, 91.
90. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
91. Yang,H., Robinson,P.N. and Wang,K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods*, **12**, 1–6.
92. Guo,Y., Ding,X., Shen,Y., Lyon,G.J. and Wang,K. (2015) SeqMule: automated pipeline for analysis of human exome/genome sequencing data. *Sci. Rep.*, **5**, 14283.
93. Stenson,P.D., Mort,M., Ball,E. V., Evans,K., Hayden,M., Heywood,S., Hussain,M., Phillips,A.D. and Cooper,D.N. (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.
94. Shearer,A.E., Eppsteiner,R.W., Booth,K.T., Ephraim,S.S., Gurrola,J., Simpson,A., Black-Ziegelbein,E.A., Joshi,S., Ravi,H., Giuffre,A.C. *et al.* (2014) Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. *Am. J. Hum. Genet.*, **95**, 445–453.
95. Kwasnieski,J.C., Fiore,C., Chaudhari,H.G. and Cohen,B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, **24**, 1595–1602.
96. Spielmann,M. and Mundlos,S. (2016) Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.*, **25**, R157–R165.