Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans

Guihu Zhao1,2,1, Kuokuo Li3,1, Bin Li1,2, Zheng Wang1, Zhenghuan Fang3, Xiaomeng Wang3, Yi Zhang1, Tengfei Luo3, Qiao Zhou1, Lin Wang3, Yali Xie1, Yijing Wang3, Qian Chen1, Lu Xia3, Yu Tang1, Beisha Tang1,2, Kun Xia3 and Jinchen Li1,2,3,*

1National Clinical Research Centre for Geriatric Disorders, Department of Geriatrics, Xiangya Hospital, Central South University, Changsha, Hunan, China, 2Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China and 3Centre for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China

ABSTRACT

De novo mutations (DNMs) significantly contribute to sporadic diseases, particularly in neuropsychiatric disorders. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) provide effective methods for detecting DNMs and prioritizing candidate genes. However, it remains a challenge for scientists, clinicians, and biologists to conveniently access and analyse data regarding DNMs and candidate genes from scattered publications. To fill the unmet need, we integrated 580 799 DNMs, including 30 060 coding DNMs detected by WES/WGS from 23 951 individuals across 24 phenotypes and prioritized a list of candidate genes with different degrees of statistical evidence, including 346 genes with false discovery rates <0.05. We then developed a database called Gene4Denovo (http://www.genemed.tech/gene4denovo/), which allowed these genetic data to be conveniently catalogued, searched, browsed, and analysed. In addition, Gene4Denovo integrated data from >60 genomic sources to provide comprehensive variant-level and gene-level annotation and information regarding the DNMs and candidate genes. Furthermore, Gene4Denovo provides end-users with limited bioinformatics skills to analyse their own genetic data, perform comprehensive annotation, and prioritize candidate genes using custom parameters. In conclusion, Gene4Denovo conveniently allows for the accelerated interpretation of DNM pathogenicity and the clinical implication of DNMs in humans.

INTRODUCTION

De novo mutations (DNMs) are defined as variants observed in individuals that are not seen in either parent and these types of variants have been reported to play prominent roles in several genetic diseases (1,2). Trios-based whole-exome sequencing (WES) and whole-genome sequencing (WGS) are the most useful methods to detect DNMs and have been successful applied in prioritizing candidate genes for autism spectrum disorder (ASD) (3), congenital heart disease (CHD) (4), undiagnosed developmental disorder (UDD) (5), epileptic encephalopathy (EE) (6), intellectual disability (ID) (7), schizophrenia (SCZ) (8) and others. Combining analyses demonstrates that for DNMs in coding-regions including single nucleotide variants (SNVs), insertions and deletions (indels) are associated with 13–60% of neurodevelopmental disorders (1). In addition, other research has shown that 42% of individuals with UDD carry pathogenic DNMs in coding regions and it is estimated that 0.22–0.47% of births involve UDD influenced by DNMs (5). Furthermore, 13% of de novo missense variants and 43% of de novo loss-of-function (LoF) variants have been diagnosed in 12% and 9% of ASD, respectively (3). In addition to neurodevelopmental disorders, DNMs also contribute to neurodegenerative disorders, such as early onset Alzheimer disease (EOAD) (9) and early onset Parkinson disease (EOPD) (10).

Because of the high clinical and genetic heterogeneities in single complex disorders, it is essential to integrate the data on DNMs that is distributed in different publications in order to more effectively prioritize novel candidate genes using a uniform strategy, such as that previously reported for autism and other neuropsychiatric disorders by us (11–13) and other groups (14–16). The denovo-db (17) aggregates a large number of DNMs identified from next-generation sequencing studies and facilitates the interpretation of DNMs.
in humans. However, the denovo-db only includes basic annotation information and does not provide a list of DNM-based candidate genes. Given that some diseases, such as different types of neurodevelopmental disorders, share significant aetiologies and phenotypes, some studies have integrated DNMs of different diseases in order to prioritize novel candidate genes (14-16). Consequently, the Developmental Brain Candidate Gene Database (14) and NPdenovo database (18) were developed to present the integrated genetic data. However, these two databases focus only on DNMs of limited types of diseases.

With the increasing application of WES and WGS, greater numbers of DNMs will be detected in individuals with different phenotypes, adding to the challenge for scientists and clinicians to determine the pathogenicity of DNMs. For the advancement of precision medicine, great efforts will be needed to assess disease-causing variants and to identify candidate genes more precisely. In a previous study we demonstrated that integrating more genetic and clinical data sources can be beneficial for better interpretation of human variants and the prioritization of candidate genes (19). In the present study, we catalogued all published DNMs detected by WES/WGS, performed comprehensive variant-level and gene-level annotations, and prioritized statistically significant candidate genes. We then developed a user-friendly integrated database called Gene4Denovo which allows DNMs, candidate genes, and annotation information in humans to be conveniently searched, browsed, and analysed.

MATERIALS AND METHODS

DNM collection

We collected DNMs from original published WES/WGS studies with sample sizes >10 (3–11,20–69) (Supplementary Table S1, Figure S1). DNMs from the denovo-db (17) and NPdenovo (18) databases were also collected. The information collected for each DNM included chromosome, start position, end position, reference sequence, alternate sequence, individual identifier, phenotype, sequence platform, publication information, and PubMed identifier. If an individual identifier was not available, ‘NA’ was used to fill this category. We also used LiftOver to translate different versions of human reference genomes (hg18 or hg38) to reference genome hg19. The complementary DNA (cDNA) positions of DNMs from some publications were translated into genomic DNA (gDNA) positions using VarCards online function that we previously developed (http://varcards.biols.ac.cn/). Given that some samples overlapped with different studies, the redundant samples were removed. If a study had integrated samples of other published studies, the DNMs and sample size recorded in Gene4Denovo were the non-redundant samples and the integrated studies were not included. If samples of an original studies had not been integrated by any other studies, the sample size recorded in Gene4Denovo was the same as the original samples. DNMs of individuals with the same phenotype from different publications were merged. In addition to DNMs from unaffected controls and patients with different disorders, we also collected DNMs from one study with mixed phenotypes (17,53) and like denovo-db (17) filled the phenotypic information of the individuals in this study using ‘Mix’.

DNM annotation and candidate genes prioritization

We performed a comprehensive analysis of the collected DNMs (Supplementary Figure S1). ANNOVAR (70) was used to perform comprehensive annotation of the DNMs based on definitions of transcripts from RefSeq, UCSC known Gene, and Ensembl Gene. Based on the functional effects, DNMs were classified into the following different types: (i) LoF variants, including frameshift indels, splicing, stopgain, and stoploss variants, (ii) deleterious missense variants (Dmis), (iii) tolerant missense variants (Tmis), (iv) synonymous variants (Syn), (v) non-frameshift indels (NF) variants and (vi) noncoding variants. The pathogenesis of the missense variants were predicted using ReVe, which was recently developed by our group (71). Missense variants with a ReVe score higher than 0.7 were considered Dmis. The LoF and Dmis variants were referred to as putative functional (Pfun) variants.

The transmitted and de novo association (TADA) model (72) was used to calculate the P-value and false discovery rate (FDR) for each gene with Pfun variants in each disorder (Supplementary Table S2). The TADA parameters for each disorder, including the background gene-level de novo mutation rate (GDNMR) of each gene, the fraction of risk genes among all human genes (π), the fold-enrichment (λ) and the relative risk (γ) were evaluated. The GDNMR was sourced from a previous study based on the trinucleotide model and several adjusted factors (73). The fraction of risk genes was evaluated by maximum likelihood estimation based on the number of Pfun DNMs and the number of genes with multiple Pfun DNMs, as described in previous studies (3,4,74). The fold-enrichment of LoF and Dmis were calculated by comparing the number of normalized LoF and Dmis variants in each case with the control. As previous studies, we normalized the number of LoF and Dmis using the number of de novo synonymous mutations in each case and the control. Finally, we calculated the relative risk of LoF and Dmis using the equation: \( π(γ − 1) = λ − 1 \). For some disorders with >500 samples, including ASD, CHD, UDD, ID, EE, SCZ and Tourette Disorder (TD), the parameters of the TADA model were re-evaluated (74). For other diseases with inadequate sample sizes, we used parameters estimated from all the integrated DNMs. Two strategies were used to prioritize candidate genes. In the first strategy, we performed TADA to calculate the FDR of each gene for each disorder. In the second strategy, we combined DNMs of each gene in all disorders and calculated the FDR. Genes with different FDR levels in either of the two prioritization strategies were classified using the following criteria, respectively: High confidence [0, 0.0001], Strong [0.0001, 0.001], Suggestive [0.001, 0.01], Positive [0.01, 0.05], Possible [0.05, 0.1] and Minor evidence [0.1, 0.2].

Variant-level data source

Initially, the allele frequencies of different populations were downloaded from various human genetic variation databases, including gnomAD (release 2.1.1), which contained variants of 125 748 exomes and 15 708 genomes.
from unrelated individuals sequenced as part of various disease-specific and population genetic studies including a total of 141 456 individuals (75); ExAC (release 1.0), which included 60 706 unrelated individuals sequenced as part of various disease-specific and population genetic studies (75,76);ESP6500 (release ESP6500SI-V2), which included 6503 exomes from European Americans and African Americans (77); 1000 Genomes Project (final phase of the project), which included genomic data for 2504 individuals from 26 different populations around the world (78); Kaviar genomic variant database (version 160 204-Public), which included integrated variants from 35 projects encompassing 13 200 genomes and 64 600 exomes (79); and HaploType Reference Consortium (HRC), which included 64 976 haplotypes from 20 studies of predominantly European ancestry (80). The predictive scores and pathogenicity consequences of missense variants were assessed based on 24 *in silico* methods, including ReVe (71), SIFT (81,82), PolyPhen2 HVAR (83), PolyPhen2 HDIV (83), LRT (84), MutationTaster (85), MutationAssessor (86), FATHMM (87), PROVEAN (88), MetaSVM (89), MetaLR (89), VEST 3.0 (90), M-CAP (91), CADD (92), GERP++ (93), DANN (94), FATHMM MKL (95), Eigen (96), GenoCanyon (97), fitCons (98), PhyloP (99), PhastCons (100), SiPhy (101) and REVEL (102). In addition, we extracted variant and related diseases or phenotype information from public disease- and phenotypes (104); COSMIC, a database of somatic reports on the relationships among human variations mutation information and related details, which also contains Cons (98), PhyloP (99), PhastCons (100), SiPhy (101) and REVEL (102). In addition, we extracted variant and related diseases or phenotype information from public disease-specific databases, including Clinical Interpretation of genetic variants (InterVar) (103); ClinVar, a database of somatic mutation information and related details, which also contains information relating to human cancers (105); ICGC, which catalogues genomic abnormalities in tumours (106); and the single nucleotide polymorphism database dbSNP v150 (107). Finally, we acquired the protein domain for each DNMs from InterPro (108) and the protein sequences across 21 species from the National Center for Biotechnology Information (NCBI) database HomoloGene (109).

Gene-level data source

A large amount of meaningful annotations for each gene was collected from public databases. Basic information and functional information of genes were sourced from the following: UniProt (release 201902), which is a collection of functional information on proteins (110); NCBI Gene, which includes gene-specific connections in the nexus of sequence, expression, function and homology data (111); NCBI BioSystems, (release 20170421), which categorizes the genes, proteins, and small molecules involved in the biological system (112); Gene ontology (GO; V1.4), which is a source of information on the functions of genes (113); and InBio Map (release 20160912), which includes information on protein–protein interactions (114). The genetic intolerance score of each gene were collected from residual variation intolerance score (RVIS), which is a gene-based score intended to help in the interpretation of human sequence data (115); the novel gene intolerance ranking system LoFTool (116); the heptanucleotide context intolerance score, which is an intolerance score quantifying the difference between the expected and observed numbers of functional variants at a gene (117); the gene damage index (GDI), which is the accumulated mutational damage for each human gene in the general population (118); Episcore, which is a computational method to predict haplinsufficiency using epigenomic features and is complementary to mutation intolerance metrics (119); and the probability of loss of function intolerance (pLI) score, which indicates the probability that a gene is intolerant to a loss of function mutation (75). In addition, disease-related or phenotype-related information of genes was extracted from Online Mendelian Inheritance in Man (OMIM) (120); ClinVar (104); Human Phenotype Ontology (HPO), which is a standardized vocabulary of phenotypic abnormalities encountered in human disease (121); mammalian phenotype from mouse genome informatics (MGI) (122); and InterPro, which is a resource that provides functional analysis of protein sequences (108). Furthermore, we collected gene expression data from BrainSpan, which contains data regarding gene expression in specific brain regions and covers several developmental stages (123); the Genotype-Tissue Expression project (GTEx), which involves the relationship among genetic variation and gene expression in multiple human tissues (124); and the protein subcellular map from the Human Protein Atlas, which is a map of protein expression across 32 human tissues (125). Finally, the drug–gene interactions data and gene druggability were sourced from the latest Drug-gene Interaction Database (DGIdb, v3.0), which assembled 56 039 drug-gene interaction claims (126).

**Database construction and interface**

Gene4Denovo ([http://www.genemed.tech/gene4denovo](http://www.genemed.tech/gene4denovo)) was developed using JavaScript, PHP, and Perl using a Linux platform on a Nginx web server. A front and back separation model was used. The front end was based on vue and used the UI Toolkit element, which supports all modern browsers across platforms, including Microsoft Edge, Safari, FireFox and Google Chrome. The back end was based on Laravel, a PHP web framework. The front and back separation model has a number of advantages, including simplicity of control, modularity and expandability. Gene4Denovo is compatible with all major browser environments and different operating systems, including Windows, Linux, and Mac. The data were stored in a MySQL database.

**RESULTS AND WEB INTERFACE**

**DNMs and candidate genes**

The Gene4Denovo database fully integrated 580 799 DNMs from 23 951 individuals across 24 phenotypes from 59 publications, including 553 404 DNMs detected by WGS and 27 395 DNMs detected by WES (Table 1). Most of the DNMs and samples were collected from nine phenotypes that included 6511 patients with ASD (*n* = 280 782), 4293 patients with UDD (*n* = 8361), 2645 patients with CHD (*n* = 2990), 933 patients with EE (*n* = 1213), 1331 patients with ID (*n* = 1493), 1094 patients with SCZ (*n* = 1064), 812 patients with TD (*n* = 805), 3391 unaffected controls (*n* = 174 836) and 1548 individuals with Mix phenotypes (*n* = 107 834). Using comprehensive annotation, we
preferentially focused on 30,060 DNMs in coding regions and splicing sites (4,582 LoF, 6,651 Dmis, 11,781 Tmis, 6,550 Syn and 496 NF). The DNMs included 8,175 in ASD, 7,696 in UDD, 2,972 in CHD, 1,478 in ID, 1,165 in EE, 1,052 in SCZ, 781 in TD, 1,702 in Mix phenotype Mix 1, 1548 in Trios DNMs Coding DNMs

| Phenotype                                    | Abbreviation | Study | Trios | DNMs | Coding DNMs |
|----------------------------------------------|--------------|-------|-------|------|-------------|
| Autism spectrum disorder                     | ASD          | 11    | 6511  | 280,782 | 8175        |
| Undiagnosed developmental disorder          | UDD          | 1     | 4293  | 8361  | 7696        |
| Congenital heart disorder                    | CHD          | 1     | 2645  | 2990  | 2972        |
| Intellectual disability                      | ID           | 7     | 1331  | 1493  | 1478        |
| Epileptic encephalopathy                     | EE           | 7     | 933   | 1213  | 1165        |
| Schizophrenia                                | SCZ          | 7     | 1094  | 1064  | 1052        |
| Tourette disorder                            | TD           | 2     | 812   | 805   | 781         |
| Congenital diaphragmatic hernia              | CDH          | 1     | 362   | 470   | 470         |
| Craniosynostosis                             | CRAN         | 1     | 294   | 322   | 319         |
| Periventricular nodular heterotopia          | PNH          | 1     | 202   | 219   | 219         |
| Amyotrophic lateral sclerosis                | ALS          | 3     | 173   | 111   | 109         |
| Bipolar disorder                             | BP           | 1     | 79    | 71    | 68          |
| Early onset Parkinson disease                | EOPD         | 2     | 49    | 60    | 60          |
| Cerebral palsy                               | CP           | 1     | 98    | 61    | 59          |
| Neural tube defects                          | NTD          | 1     | 43    | 40    | 40          |
| Early-onset high myopia                      | EOHM         | 1     | 18    | 20    | 19          |
| Early onset Alzheimer disease                | EOAD         | 1     | 12    | 15    | 15          |
| Smith-Magenis syndrome                       | SMS          | 1     | 13    | 13    | 13          |
| Cantu syndrome                               | CS           | 1     | 14    | 6     | 6           |
| Sporadic infantile spasm syndrome            | SISS         | 1     | 10    | 5     | 5           |
| Acromelic frontonasal dysostosis             | AFD          | 1     | 4     | 4     | 4           |
| Anophthalmia/Microphthalmia                 | AM           | 1     | 25    | 4     | 4           |
| Control                                      | Control      | 9     | 3391  | 174,836 | 3629 |
| Mix phenotype                                | Mix          | 1     | 1548  | 107,834 | 1702 |
| Total                                        |              | 59    | 23,951 | 350,799 | 30,060 |

All DNMs reported in primary publications were integrated in Gene4Denovo database. ANNOVAR was performed to annotate these DNMs. Variants with functional effects of frameshift indels, stopgain, and stoploss, missense, synonymous, non-frameshift indels and splicing site (4582 LoF, 6651 Dmis, 6751 Tmis, 6550 Syn and 496 NF). The DNMs included 8,175 in ASD, 7,696 in UDD, 2,972 in CHD, 1,478 in ID, 1,165 in EE, 1,052 in SCZ, 781 in TD, 1,702 in Mix phenotype Mix 1, 1548 in Trios DNMs Coding DNMs

The quick search automatically recognizes a variety of key terms, such as gene symbol, genomic region, cytoband, transcript accession, the nucleic acid change in a certain transcript, the genomic coordinate of a variant, as well as the DNM identifier. Moreover, several examples of input query formats are available by clicking the ‘example’ link with the corresponding examples occurring in the input box. The advanced search (http://www.genemed.tech/)
Table 2. Summary of prioritized candidate genes in Gene4Denovo database

| Disease (trios) | FDR ≤ 0.0001 | 0.0001 < FDR ≤ 0.001 | 0.001 < FDR ≤ 0.01 | 0.01 < FDR ≤ 0.05 | 0.05 < FDR ≤ 0.1 | 0.1 < FDR ≤ 0.2 |
|----------------|--------------|-----------------------|-------------------|------------------|-----------------|-----------------|
| ASD (6511)     | 13           | 10                    | 29                | 26               | 53              |
| UDD (4293)     | 85           | 21                    | 43                | 50               | 40              | 69              |
| CHD (2645)     | 3            | 3                     | 4                 | 12               | 13              | 25              |
| ID (1331)      | 26           | 13                    | 18                | 16               | 14              | 34              |
| EE (933)       | 14           | 3                     | 14                | 8                | 29              |
| SCZ (1094)     | 0            | 0                     | 0                 | 1                | 9               |
| TD (812)       | 0            | 0                     | 0                 | 2                | 3               | 6               |
| CDH (362)      | 0            | 1                     | 0                 | 0                | 1               |
| CRAN (291)     | 0            | 1                     | 0                 | 0                |
| PNH (202)      | 1            | 0                     | 0                 | 0                | 0               |
| ALS (173)      | 0            | 0                     | 0                 | 0                | 0               |
| BP (79)        | 0            | 0                     | 0                 | 0                | 0               |
| EOPD (49)      | 0            | 0                     | 0                 | 0                |
| CP (98)        | 0            | 0                     | 0                 | 1                |
| NTD (43)       | 0            | 0                     | 0                 | 1                | 0               |
| SMS (13)       | 1            | 0                     | 0                 | 0                |
| CS (14)        | 1            | 0                     | 0                 | 0                |
| AFD (4)        | 0            | 1                     | 0                 | 0                |
| CD (19 012)    | 117          | 27                    | 46                | 60               | 47              | 88              |
| Total          | 132          | 36                    | 62                | 116              | 99              | 230             |

ASD, autism spectrum disorder; UDD, undiagnosed developmental disorder; CHD, congenital heart disorder; ID, intellectual disability; EE, epileptic encephalopathy; SCZ, schizophrenia; TD, tourette disorder; CDH, congenital diaphragmatic hernia; CRAN, craniosynostosis; PNH, periventricular nodular heterotopia; ALS, amyotrophic lateral sclerosis; BP, bipolar disorder; EOPD, early onset parkinson disease; CP, cerebral palsy; NTD, neural tube defects; SMS, smith-magenis syndrome; CS, cantu syndrome; AFD, acromelic frontonasal dysostosis. CD, combined all samples with different disorders. Number of genes with FDR < 0.2 in each disorder and cross disorders analysis were showed in this table. We ranked all candidate genes into six tiers based on the strength of false discovery rate (FDR). The total number of candidate genes were counted after removing redundancy.

Figure 1. Snapshot of variant-level implications in Gene4Denovo. Two approaches are available to access variant-level implications, the ‘Quick search’ and ‘Advanced search’. The results of a quick search for the KCNQ2 gene are shown as an example, including the functional effects at the transcript and protein levels, homology, predicted damaging severity of missense variants, allele frequencies in different populations, and information in disease-related databases.

gene4denovo/search) supports batch searches and allows users to specify annotated datasets. The advanced search provides options that include primary information, predictive algorithms for nonsynonymous variation, allele frequency in different populations, and disease-related and phenotype-related information. The advanced search also has six types of input forms that are similar to the quick search (gene symbol, genomic region, cytoband, transcript accessions, the nucleic acid change in a certain gene or transcript, and the genomic coordinate of a variant). To improve the user experience, the advanced search query form and the corresponding result sets are displayed on the same page. Of note, the search results can be freely exported as Excel files to download.
Variant-level implications in Gene4Denovo

Both quick and advanced searches provide access to detailed DNM annotation data. Search results are returned as a page that contains two tables. The first table is a summary of DNMs for each gene in each disorder while the second table displays all the detail information regarding variant-level annotations (Figure 1). The summary table synoptically presents the number of LoF, Dmis, Pfun and Tmis, synonymous, the non-frameshift and non-coding variants, the P-value, and the FDR for each gene in each disorder. The variants table presents detailed information for each DNM, including the following aspects: (i) the functional effect and reference information for each DNM; (ii) the predicted damaging scores and functional consequences of missense variants based on 24 in silico algorithms; (iii) the allele frequencies of different populations based on seven data sources; (iv) the disease-related information from seven popular related data sources and (v) the protein sequences across 21 species from HomoloGene, including the graphic presentation of multiple sequence alignment between species. The variants table can be filtered by functional effects, adding flexibility to the output. In addition, users can specify any of the mentioned data sources to limit the contents presented to those of specific interest.

Gene-level implications in Gene4Denovo

On the page of variants-level implications, users can click on the corresponding gene symbols in the summary table or variants table to access detailed information regarding the given genes. All genes containing DNMs were curated in Gene4Denovo, which currently includes the following six specified panels: (i) basic information, (ii) gene function, (iii) phenotype and disease, (iv) gene expression, (v) variants in different populations and (vi) drug–gene interaction (Figure 2, (Supplementary Table S4). The ‘basic information’ displays the integrated basic information for the gene, including the official gene name, synonyms, genomic coordinate, gene type and functional annotations, the genic intolerance score based on six studies (75,115–119), and a summary of the cellular function of the protein encoded by the gene sourced according to UniProt (110). The ‘gene function’ consists of five sub Panels, including (i) the molecular function retrieved from UniProtKB; (ii) gene ontology terms retrieved from Gene Ontology Consortium (113); (iii) domain information retrieved from InterPro (108); (iv) protein–protein interactions retrieved from InBio Map (114) and (v) biological pathway information retrieved from BioSystems (112). The ‘phenotype and disease’ panel consists of four sub Panels, including (i) phenotype data retrieved from OMIM (120); (ii) clinical variation data retrieved from ClinVar (104); (iii) mammalian phenotype data retrieved from MGI (122) and (iv) human phenotype ontology retrieved from HPO (121). The ‘gene expression’ panel consists of four sub Panels, including (i) spatio-temporal expression profiles retrieved from BrainSpan (123); (ii) cell diversity and expression in the human cortex based on single-cell RNA-seq from the Allen Brain Atlas; (iii) gene expression data in 31 primary tissues and 53 secondary tissues retrieved from GTEx (124) and (iv) subcellular location retrieved from The Human Protein Atlas (125). The ‘variants in different populations’ panel provides the number of variants with different functional effects at different threshold in different populations. The ‘drug–gene interaction’ panel provides data for drug–gene interactions and gene drugability, which is retrieved from DGIdb v3.0 (126).

Customized analysis section in Gene4Denovo

Gene4Denovo provides an interface to allow users to freely analyse their own genetic data (http://genemed.tech/gene4denovo/analysis). As shown in Figure 3, the analysis process includes four simple steps: (i) inputting an email address, (ii) choosing the Trio or Non-trio option of users’ genetic data, (iii) uploading genetic data files (VCF4 format) and (iv) inputting the basic information for each sample. If users select the Trio option, the users must select the paternal sample ID, maternal sample ID, children’s sample ID and the gender of the children. Gene4Denovo will automatically identify the DNMs, homozygous variants, compound heterozygous variants, and X-linked variants using default parameters. If the Non-trio option is chosen, the users must select the genotypes of each sample, including heterozygous, homozygous, wild type, and so on. Gene4Denovo will identify the user-defined co-segregated rare damaging variants using default parameters. It is noteworthy that users are able to specify cut-off values of quality control, the data sources of annotation and the parameters used for identifying rare damaging variants. In the quality control panel, users are able to specify several parameters used to detect high-confidence genetic variants, including the minimum QUAL, sequencing depth, allele depth, and genotype quality. There are four annotation sub Panels: (i) ‘Basic information annotation’ to specify three basic data sources of annotation (such as cytoband database, gene-level based databases, and Gene4Denovo), which refer to the identifier, putative functional DNMs, P-value and FDR of candidate genes in each disorder; (ii) ‘Pathogenicity prediction of missense variants’ to specify the methods and cut-off values for predicting deleterious missense variants; (iii) ‘Allele frequency in variant population’ to specify the cut-off values of allele frequency for detecting rare variants according to different population databases and (iv) ‘Clinical related database’ to specify clinical related database, such as InterVar, ClinVar, COSMIC, ICGC and NCI-60. After completing the analysis, Gene4Denovo sends an email to the designated email address that includes a link for downloading the analysis results.

Other sections in Gene4Denovo

Gene4Denovo also contains additional useful sections. These include (i) the browse section, which can be used for accessing gene-level summary implication efficiently; (ii) the upload section, which provides a user-friendly web-based process for uploading and archiving users’ DNMs; (iii) the download section, which allows users to freely access all released datasets in Gene4Denovo without login requirements and to download the complete de novo data files via http; (iv) the data source, which shows brief information regarding the integrated databases and (v) the tutorial section, which provides a further description of Gene4Denovo and details on how to get started.
Figure 2. Snapshot of gene-level implications in Gene4Denovo. The typical gene-level implications of the KCNQ2 gene are illustrated as an example, including basic information, gene functions, associated phenotypes and diseases, gene expression, variants in different populations, and drug–gene interactions.
Figure 3. Snapshot of analysis panel in Gene4Denovo. There are four steps in the analysis process: inputting an email address, choosing the Trio or Non-trio option, uploading the data files, and inputting the trio or genotype information. To increase flexibility, users are able to specify annotation datasets, such as functional effects, allele frequencies, and predicted damaging scores from any of the 24 in silico algorithms.
DISCUSSION

A DNM-based strategy for genome and exome analyses provides unprecedented opportunities to promote our knowledge regarding genetic pathogenic mechanisms in humans for complex disorders having high clinical and genetic heterogeneity (127–130). However, a major challenge is the scattered distribution of DNM data and annotated genetic and clinical data sources (14,17,18). To make the DNM and annotated data more accessible, we collected DNM data from various published WES/WGS studies, performed uniform comprehensive annotation, and prioritized a list of candidate genes. In addition, we developed a user-friendly, interactive, open-access web-based interface to browse, search, analyse, and download the integrated data. More than 60 popular genomic data sources were integrated into Gene4Denovo in order to provide users with comprehensive information regarding variants and genes.

Gene4Denovo accentuates the importance of integrating DNMs from different studies in a uniform manner. First, we found that integrating DNMs from multiple publications for a single disorder improved the power of prioritizing candidate genes due to increasing the sample sizes. Additionally, users will be able to analyse more DNMs by integrating their own in-house data with our database and then prioritize new candidate genes. Second, we found that integrating DNMs from different disorders resulted in an additional 84 candidate genes being prioritized, including eight genes that reached an FDR < 0.01. For example, KMT2C (FDR = 7.02 × 10^-3) was prioritized as a strong candidate gene by combining DNMs from seven disorders (ASD, CHD, UDD, ID, SCZ, ALS and BP). Third, integrated DNMs of unaffected controls were able to be used as negative control in the identification of pathogenic variants or candidate genes. For example, KDM5B carried 9, 3, 3, 3, 3, 3, 1, 1 Pfun DNMs in patients with ASD (FDR = 2.30 × 10^-3), CHD (FDR = 7.17 × 10^-3), UDD (FDR = 0.093), TD (FDR = 0.36), and CDH (FDR = 0.61), respectively. However, we found that KDM5B also carried five Pfun DNMs in control cohorts, suggesting that DNMs of KDM5B may not be associated with these diseases. This was consistent with a recent study that found KDM5B is a recessive gene associated with neurodevelopmental disorders (131). We fully prioritized 675 candidate genes. Of the candidate genes, 60.59% (409/675) carried Pfun DNMs in two or more disorders. This may be useful in identifying biomarkers that can be used in a translational setting for genetic counselling and clinical assessment. Some of the candidate genes have been well validated by functional experiments or clinical phenotypes, such as CHD8 (132), SCN2A (133,134), CACNA1E (135) and POGZ (136,137), while others need further functional validation.

All individuals carry ~70 DNMs in their genome, but only a small number of DNMs contribute to human diseases, making it a challenge to interpret the pathogenicity of DNMs and genes (56). To address this need, we integrated >60 genomic sources into Gene4Denovo in order to provide comprehensive analyses of DNMs and candidate genes. First, it has reported that approximately one third of DNMs of neurodevelopmental disorders are present in the general population and this type of DNM might do not contribute to risk of developing a disorder (138). Therefore, it was important for the population-based background allele frequency to be integrated into Gene4Denovo so to allow for a better understanding of pathogenic variants. Second, several computational methods have been used to predict deleterious variants in humans, but these methods provide inconsistent results (71). Therefore, Gene4Denovo offers prediction scores from 24 in silico algorithms and allows users to select one or a combination of multiple suitable methods. Third, Gene4Denovo integrated variant-level information from popular genetic database, such as ClinVar (104), OMIM (120) and HPO (139), which may help users to comprehensively evaluate the pathogenicity of genes and genetic variants. Fourth, Gene4Denovo integrated meaningful gene-level information, such as gene function and gene expression patterns, in order to provide users comprehensive information regarding a given gene from a one-stop database.

Despite of the advancement of other available databases related to DNMs, Gene4Denovo exhibits significant differences that represent major advances. The mirD-NMR database (140) focuses on gene-level background DNM rates predicted by four different methods instead of analysing DNMs themselves. The EpiDenovo database (141) provides the associations between embryonic epigenomes and DNMs in developmental disorders. Compared to the denovo-db (17), Gene4Denovo not only integrated more DNMs, but also provided more comprehensive annotation information collected from >60 genomic data sources. This extensive integration should further facilitate the interpretation of DNMs. In addition, Gene4Denovo prioritizes a list of candidate genes with different degrees of statistical evidence. This is important for biologists in selecting genes for functional validation and for geneticists and clinicians for genetic counselling. Furthermore, in order to facilitate research communities to take advantage of the integrated DNMs, candidate genes, and other genomic data sources, Gene4Denovo provides a user-friendly interface for detecting DNMs, homozygous variants, X-linked variants, and co-segregated variants for performing customized comprehensive annotations, and for prioritizing pathogenic variants and risk-genes.

There are some limitations to the present study that require further effort in order to be resolved. First, the candidate genes were prioritized using the TADA model, which influenced by several factors, such as GDNMR, the tools used for predicting deleterious missenses, and the enrichment of LoF and Dmis. We encourage users to download the integrated DNMs from Gene4Denovo database and prioritized candidate genes using different parameters and new models. For example, Nguyen, et al. developed a new method called extTADA and prioritized 288 candidate genes in neurodevelopmental disorders. Additional experiment validation and more detailed clinical phenotypes of patients are still needed. Second, despite noncoding variants being catalogued in the Gene4Denovo database, the current study prioritized candidate genes based only on DNMs in coding regions. DNMs in noncoding regions, such as those in promoters (45), are also likely to contribute to the risk of developing disorders. In the next version of Gene4Denovo, we plan to integrate both coding...
DNMs and noncoding DNMs for prioritizing candidate genes. Third, in order to provide uniform genetic data, Gene4Denovo focused on DNMs detected by WES/WGS, which are most wildly used in medical genetics. This means that some DNMs from targeted sequencing studies and case reports were not included. Since it is still challenging to accurately detect \textit{de novo} copy-number variations (CNVs) from NGS data, especially WES data, the current version of Gene4Denovo did not integrate \textit{de novo} CNVs. However, we plan to add \textit{de novo} CNVs in the next version of Gene4Denovo. In addition, we plan to continuously collect DNMs from the latest published WES/WGS studies and to update the Gene4Denovo database every six months. We also promote and highly appreciate users uploading their own DNA data and archives into Gene4Denovo by using the uploading interface.

In conclusion, Gene4Denovo offers a large number of freely available DNMs with uniform curation and annotations across 24 phenotypes. Gene4Denovo also provides a list of prioritized candidate gene and comprehensive genetic and clinical information for each DNM and gene. We hope the Gene4Denovo database will provide a great convenience for geneticists, biologists, and clinicians and accelerate the interpretation of DNM pathogenicity and its clinical implication.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank the members of the Centre for Medical Genetics, Central South University for their valuable discussion regarding this work.

FUNDING
National Natural Science Foundation of China [81801133 to J.C.L.]; Young Elite Scientist Sponsorship Program by CAST [2018QNRC001 to J.C.L.]; Innovation-Driven Project of Central South University [2018003304004 to J.C.L.]; Natural Science Foundation for Young Scientists of Hunan Province, China [2019JJ50974 to G.H.Z.]. Funding for open access charge: National Natural Science Foundation of China [81801133 to J.C.L.]; Innovation-Driven Project of Central South University [2018003304004 to J.C.L.]; Natural Science Foundation for Young Scientists of Hunan Province, China [2019JJ50974 to G.H.Z.].

Conflict of interest statement. None declared.

REFERENCES
1. Wilfert, A.B., Sulovari, A., Turner, T.N., Cox, B.P. and Eichler, E.E. (2017) Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. \textit{Genome Med.}, 9, 101.
2. Ronemus, M., Jossifov, I., Levy, D. and Wigler, M. (2014) The role of \textit{de novo} mutations in the genetics of autism spectrum disorders. \textit{Nat. Rev. Genet.}, 15, 133–141.
3. Jossifov, I., O’Roak, B.I., Sanders, S.I., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E. et al. (2014) The contribution of \textit{de novo} coding mutations to autism spectrum disorder. \textit{Nature}, 515, 216–221.
4. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang-W, Sierant, M.C. et al. (2017) Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. \textit{Nat. Genet.}, 49, 1593–1601.
5. Deciphering Developmental Disorders, S. (2017) Prevalence and architecture of de novo mutations in developmental disorders. \textit{Nature}, 542, 433–438.
6. EpiK-C and Epilepsy Phenome/Genome, P. Epilepsy Phenome/Genome, P., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T. et al. (2013) De novo mutations in epileptic encephalopathies. \textit{Nature}, 501, 217–221.
7. Lelieveld, S.H., Reijnards, M.R., Pfundt, R., Ntema, H.G., Kamsteeg, E.J., de Vries, P., de Vries, B.B., Willemsen, M.H., Kleeftstra, T., Lohner, K. et al. (2016) Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. \textit{Nat. Neurosci.}, 19, 1194–1196.
8. Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M. et al. (2014) \textit{De novo} mutations in schizophrenia implicate synaptic networks. \textit{Nature}, 506, 179–184.
9. Rovelet-Lecrux, A., Charbonnier, C., Wallon, D., Nicolas, G., Seaman, M.N., Pottie, C., Breusegem, S.Y., Mathur, P.P., Jenardhanan, P., Le Guennec, K. et al. (2015) De novo deleterious genetic variations target a biological network centered on Abeta peptide in early-onset Alzheimer disease. \textit{Mol. Psychiatry}, 20, 1046–1056.
10. Guo, J.F., Zhang, L., Li, K., Mei, J.P., Xue, J., Chen, J., Tang, X., Shen, L., Jiang, H., Chen, C. et al. (2018) Coding mutations in \textit{NUS1} contribute to Parkinson’s disease. \textit{PNAS}, 115, 11567–11572.
11. 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. \textit{Mol. Psychiatry}, 22, 1282–1290.
12. Li, J., Hu, S., Zhang, K., Shi, L., Zhang, Y., Zhao, T., Wang, L., He, X., Xia, K., Liu, C. et al. (2018) A comparative study of the genetic components of three subcategories of autism spectrum disorder. \textit{Mol. Psychiatry}, doi:10.1038/s41380-018-0081-x.
13. 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 \textit{de novo} mutation burdens in autism spectrum disorder. \textit{Am. J. Med. Genet. B Neuropsychiatr. Genet.}, 174, 568–577.
14. Gonzalez-Mantilla, A.J., Moreno-De-Luca, A., Ledbetter, D.H. and Martin, C.L. (2016) A cross-disorder method to identify novel candidate genes for developmental brain disorders. \textit{JAMA Psychiatry}, 73, 275–283.
15. Nguyen, H.T., Bryois, J., Kim, A., Dobyns, A., Huckins, L.M., Munoz-Manchado, A.B., Ruderfer, D.M., Genovese, G., Fromer, M., Xu, X. et al. (2017) Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. \textit{Genome Med.}, 9, 114.
16. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bukken, T.E., Luke, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A. et al. (2019) Neurodevelopmental disease genes implicated by \textit{de novo} mutation and copy number variation morbidity. \textit{Nat. Genet.}, 51, 106–116.
17. Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., Ha, F.S., Doebley, A., Bernier, R.A., Nickerson, D.A. and Eichler, E.E. (2017) \textit{De novo}-db: a compendium of human \textit{de novo} mutations. \textit{Nucleic Acids Res.}, 45, D804–D811.
18. Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., Li, X., Shao, Q., Ran, X., Li, Z. et al. (2016) Genes with \textit{de novo} mutations are shared by four neuropsychiatric disorders discovered from \textit{NP}denuvo database. \textit{Mol. Psychiatry}, 21, 290–297.
19. Li, J., Shi, L., Zhang, K., Yang, X., Hu, S., Zhao, T., Teng, H., Li, X., Jiang, Y., Ji, L. et al. (2018) VarCards: an integrated genetic and clinical database for coding variants in the human genome. \textit{Nucleic Acids Res.}, 46, D1039–D1048.
21. Kloosterman, W.P., Francioli, L.C., Hormozdari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameije, E.W., Moed, M.H., Koval, V., Renkens, I. et al. (2015) Characteristics of de novo structural changes in the human genome. Genome Res., 25, 792–801.

22. Heinezen, E.L., O’Neill, A.C., Zhu, X., Allen, A.S., Bahlo, M., Chelly, J., Chen, M.H., Dobyns, W.B., Freytag, S., Guerrini, R. et al. (2018) De novo and inherited private variants in MAP2B in periventricular nodular heterotopia. PLoS Genet., 14, e1007281.

23. Vetrini, F., Meeke, S., Rosenhof, J.A., Suri, M., Lewis, A.M., Nugent, K.M., Roeder, E., Littlejohn, R.O., Holder, S., Zhu, W. et al. (2019) De novo and inherited TFCT20 pathogenic variants are associated with intellectual disability, dysmorphic features, hypotonia, and neurological impairments with similarities to Smith-Magenis syndrome. Genome Med., 11, 12.

24. Xu, B., Ioniota-Laza, I., Roos, J.L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J.A. and Karayiorgou, M. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat. Genet., 44, 1365–1369.

25. Timberlake, A.T., Furey, C.G., Choi, J., Nelson-Williams, C. and Yale, E.T. (2017) Structural changes in the human genome. ABCC9, 1921–1929.

26. Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.W., Moed, M.H., Oegema, R., Nava, C. et al. (2019) Increased exonic de novo mutation rate in individuals with autism spectrum disorder. Science, 362, eaat6576.

27. Steinberg, K.M., Yu, B., Koboldt, D.C., Mardis, E.R. and Pamphlett, R. (2015) Exome sequencing of case-unaffected-parents trios reveals recessive and de novo genetic variants in sporadic ALS. Sci. Rep., 5, 9124.

28. Vissers, L.E., de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C. et al. (2012) Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med., 367, 1921–1929.

36. Helbig, K.L., Farwell Hagman, K.D., Shinde, J.D.N., Moske, C., Powis, Z., Li, S., Tang, S. and Helbig, I. (2016) Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. Genet. Med., 18, 898–905.

37. Kataoka, M., Matoba, N., Sawada, T., Kazuno, A.A., Ishiwata, M., Fujii, K., Matsuoka, K., Takata, A. and Kato, T. (2016) Exome sequencing for bipolar disorder points to roles of de novo loss-of-function and protein-altering mutations. Mol. Psychiatry, 21, 885–893.

38. Smith, J.D., Hing, A.V., Clarke, C.M., Johnson, N.M., Perez, F.A., Park, S.S., Horst, J.A., Mechem, B., Maves, L., Nickerson, D.A. et al. (2014) Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontaloss dysostosis. Am. J. Hum. Genet., 95, 235–240.

39. Slavotinek, A.M., Garcia, S.T., Chandrathilakille, G., Bardakjian, T., Ullah, E., Wu, D., Umeda, K., Lao, R., Tang, P.L. and Wan, E. et al. (2015) Exome sequencing in 32 patients with anophthalmia/microphthalmia and developmental eye defects. Clin. Genet., 88, 468–473.

40. Guipponi, M., Santoni, F.A., Setola, V., Gehrig, C., Rothamel, M., Cuenca, M., Guillou, D., Dikèos, D., Georgiopoulos, G., Papadimitriou, G. et al. (2014) Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. PLoS One, 9, e12745.

41. Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Schwartz, G.B., Collins, R.L. et al. (2018) Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science, 362, 860–863.

42. Bowling, K.M., Thompson, M.L., Amaral, M.D., Finnila, C.R., Caglayan, H., Craiu, D., De Jonghe, P., Guerrini, R., Helbig, K.L., Kitaygorodsky, A., Hernan, R., Aspelund, G. et al. (2017) De novo mutations in synaptic transmission genes including ELN and CAV3 are enriched in sporadic ALS. Genet. Med., 19, 1048–1053.

43. Dionne-Laporte, A., Spiegelman, D., Laporte, A.D., Nacuvas, C., Pelletier, N., Morin, F. et al. (2017) Genome-wide de novo CNV score identifies a new common risk factor for schizophrenia. Mol. Psychiatry, 22, 1734–1743.

44. Chesi, A., Stahl, B.T., Jovicic, A., Couthouis, J., Fasolino, M., Raphael, A.R., Yamazaki, T., Elia, I., Polak, M., Kelly, C. et al. (2013) Exome sequencing to identify de novo mutations in sporadic ALS. Nat. Neurosci., 16, 851–855.

45. An, J.Y., Lin, K., Zhu, L., Merlin, D., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L. et al. (2018) De novo mutations in schizophrenia implicate chromatin remodeling and neural pathways in non-syndromic midline craniosynostosis. Mol. Psychiatry, 23, 1270–1281.

46. Truak, H.Y., Liao, Y., Chumchob, M., Deevad, B., Hefner, L., Ow, S., Kim, Y., Yoon, S., Park, S.S., Helbig, K.L., Kato, T. et al. (2017) Identification of candidate genes involved in the etiology of sporadic Tourette syndrome by exome sequencing. Am. J. Med. Genet. Part B, 174, 712–723.

47. Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Doucette, A., Dionne-Laporte, A., Spiegelman, D., Henrich, E., Diao, O., et al. (2011) Increased exome de novo mutation rate in individuals with schizophrenia. Nat. Genet., 43, 860–863.

48. Takata, A., Miyake, N., Tsurusaki, Y., Fukui, R., Miyatake, S., Komishu, E., Kishima, I., Okada, T., Morikawa, M., Uno, Y. et al. (2018) Integrative analyses of De Novo mutations provide deeper biological insights into autism spectrum disorder. Cell Reports, 22, 734–747.

49. Chen, R., Davis, L.K., Guter, S., Wei, Q., Jacob, S., Potter, M.H., Cox, N.J., Cook, E.H., Sutcliffe, J.S. and Li, B. (2017) Leveraging blood serotonin as an endophenotype to identify de novo and rare variants associated in autism. Mol. Autism, 8, 14.
52. Lemay, P., Guyot, M.C., Tremblay, E., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., De Marco, P., Merelle, E., Massicotte, C. et al. (2015) Loss-of-function de novo mutations play an important role in severe human neural tube defects. *J. Med. Genet.*, 52, 493–497.

53. Jonsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Jonsson, E., Hardarson, M.T., Hjorleifson, K.E., Hjartarson, E., Egertsson, H.P., Gudjonsson, S.A. et al. (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549, 519–522.

54. Goldmann, J.M., Wong, W.S., Pinelli, M., Farrah, T., Bodian, D., Stiticr, A.B., Glusman, G., Vissers, L.E., Hoischen, A., Roach, J.C. et al. (2016) Postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.*, 19, 1217–1224.

55. Kong, A., Diekstra, F.P., Verde, J., Andersen, P.M., Dekker, A.M., Tiloca, C., Porteous, D., Smith, B., Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. and Roach, J.C. (2013) Large-scale analysis of de novo mutations in autism spectrum disorder. *Hum. Mutat.*, 38, 1534–1541.

56. Gulsuner, S., Walsh, W.A., Lee, M.K., Sulem, P., Risch, N., Zink, F., Jonsson, E., Thiruvahindrapuram, B., Patel, R.V., Wu, K.C., Xiao, X., Kloss, B.A. and Roach, J.C. (2012) Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.*, 15, 471–475.

57. Lim, E.T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A.S., Zhang, X., D’Gama, A.M., Kim, S.N., Hild, R.S., Goldberg, A.P. et al. (2017) Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.*, 20, 1217–1224.

58. van Doormaal, P.T.C., Ticozzi, N., Weishaupt, J.H., Kenna, K., Diekstra, F.P., Verroti, E., Andersen, P.M., Dekker, A.M., Tiloca, C., Marroquin, N. et al. (2017) The role of de novo mutations in the development of amyotrophic lateral sclerosis. *Hum. Mutat.*, 38, 1534–1541.

59. Gulsuner, S., Walsh, W.A., Lee, M.K., Le, M.K., Casadei, S., Rippey, C., Shahin, H., Consortium on the Genetics of, S. and Group, P.S. (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell.*, 154, 518–529.

60. De Rubeis, S., He, X., Goldberg, A.P., Poulton, C.S., Samocha, K., Cicek, A.E., Koe, Y., Liu, L., Fromer, M., Walker, S. et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515, 209–215.

61. Kun-Rodrigues, C., Ganos, C., Guerreiro, R., Schneider, S.A., Spiegelman, D., Henrion, E., Diallo, O., De Marco, P., Merelle, E., Massicotte, C. et al. (2015) Functional, molecular, and phenotypic consequences of amino acid substitutions and indels. *Nat. Genet.*, 47, 593–939.

62. Goldmann, J.M., Wong, W.S., Pinelli, M., Farrah, T., Bodian, D., Stiticr, A.B., Glusman, G., Vissers, L.E., Hoischen, A., Roach, J.C. et al. (2016) Postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.*, 19, 1217–1224.

63. Ja, Y., Zhao, T., Zhang, Y., Hua, Z., Shi, L., Chen, Y., Wang, X. and Sun, Z. (2018) Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.*, 46, 7793–7804.

64. He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D. et al. (2013) Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.*, 9, e1003671.

65. Samocha, K.E., Robinson, E.B., Sanders, J.S., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Malick, S., Kirby, A. et al. (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, 46, 944–950.

66. Wilsey, A.J., Fernandez, T.V., Yu, D., King, R.A., Dietrich, A., Xing, J., Sanders, S.J., Mandell, J.D., Huang, A.Y., Richer, P. et al. (2017) De novo coding variants are strongly associated with tourette disorder. *Neuron*, 94, 486–499.

67. 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.

68. Karczewski, K.J., Weisburd, B., Solomonson, M., Ruderer, D.M., Kavanagh, D., Hamamcy, 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.

69. 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.

70. Genomes Project, C., 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.

71. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. and Roach, J.C. (2011) Kaviar: an accessible system for testing SNV novelty. *Biobinfochem.*, 27, 3216–3217.

72. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. et al. (2014) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, 46, 1279–1283.

73. Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, 31, 3812–3814.

74. Vasar, R., Adusumalli, S., Leng, S.N., Sikic, M. and Ng, P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, 11, 1–9.

75. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Grasinskaia, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nucleic Acids Res.*, 38, 659–659.

76. Shihabi, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N. 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.

77. 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, e46888.
psychiatric disorders: from discovery to application. \textit{Front Genet.}, \textbf{10}, 258.

131. Martin,H.C., Jones,W.D., McIntyre,R., Sanchez-Andrade,G., Sanderson,M., Stephenson,J.D., Jones,C.P., Handsaker,J., Gallone,G., Bruntraeger,M. \textit{et al.} (2018) Quantifying the contribution of recessive coding variation to developmental disorders. \textit{Science}, \textbf{362}, 1161–1164.

132. Bernier,R., Golzio,C., Xiong,B., Stessman,H.A., Coe,B.P., Penn,O., Witherspoon,K., Gerds.T., Baker,C., Vulto-van Silfhout,A.T. \textit{et al.} (2014) Disruptive CHD8 mutations define a subtype of autism early in development. \textit{Cell}, \textbf{158}, 263–276.

133. Ben-Shalom,R., Keeshen,C.M., Berrios,K.N., An,J.Y., Sanders,S.J. and Bender,K.J. (2017) Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. \textit{Biol. Psychiatry}, \textbf{82}, 224–232.

134. Wolff,M., Johannesen,K.M., Hedrich,U.B.S., Masnada,S., Rubboli,G., Gardella,E., Lesca,G., Ville,D., Milh,M., Villard,L. \textit{et al.} (2017) Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. \textit{Brain}, \textbf{140}, 1316–1336.

135. Helbig,K. L., Lauerer,R.J., Bahr,J.C., Souza,I.A., Myers,C.T., Uysal,B., Schwarz,N., Gandini,M.A., Huang,S., Keren,B. \textit{et al.} (2018) De novo pathogenic variants in CACNA1E cause developmental and epileptic encephalopathy with contractures, macrocephaly, and Dyskinesias. \textit{Am. J. Hum. Genet.}, \textbf{103}, 666–678.

136. Zhao,W., Tan,J., Zhu,T., Ou,J., Li,Y., Shen,L., Wu,H., Han,L., Liu,Y., Jia,X. \textit{et al.} (2019) Rare inherited missense variants of POGZ associate with autism risk and disrupt neuronal development. \textit{J. Genet. Genomics}, \textbf{46}, 247–257.

137. Zhao,W., Quan,Y., Wu,H., Han,L., Bai,T., Ma,L., Li,B., Xun,G., Ou,J., Zhao,J. \textit{et al.} (2019) POGZ de novo missense variants in neuropsychiatric disorders. \textit{Mol. Genet. Genomic Med.}, \textbf{7}, e900.

138. 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. \textit{et al.} (2017) Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. \textit{Nat. Genet.}, \textbf{49}, 504–510.

139. Kohler,S., Carmody,L., Vasilevsky,N., Jacobsen,I.O.B., Danis,D., Gourdin,J.P., Gargano,M., Harris,N.L., Matentzoglu,N., McMurry,J.A. \textit{et al.} (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. \textit{Nucleic Acids Res.}, \textbf{47}, D1018–D1027.

140. Jiang,Y., Li,Z., Liu,Z., Chen,D., Wu,W., Du,Y., Ji,L., Jin,Z.B., Li,W. and Wu,J. (2017) mirDNMR: a gene-centered database of background de novo mutation rates in human. \textit{Nucleic Acids Res.}, \textbf{45}, D796–D803.

141. Mao,F., Liu,Q., Zhao,X., Yang,H., Guo,S., Xiao,L., Li,X., Teng,H., Sun,Z. and Dou,Y. (2018) EpiDenovo: a platform for linking regulatory de novo mutations to developmental epigenetics and diseases. \textit{Nucleic Acids Res.}, \textbf{46}, D92–D99.