D796–D803 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 30 October 2016*
doi: 10.1093/nar/gkw1044

mirDNMR: a gene-centered database of background *de novo* mutation rates in human

Yi Jiang[1,†], Zhongshan Li[1,†], Zhenwei Liu[1], Denghui Chen[2], Wanying Wu[3], Yaoqiang Du[2], Liying Ji[1], Zi-Bing Jin[4], Wei Li[2,*] and Jinyu Wu[1,*]

[1]Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou 325000, China, [2]Zhejiang Provincial Key Laboratory of Medical Genetics, School of Laboratory Medicine and Life Sciences, Wenzhou Medical University, Wenzhou 325000, China, [3]Beijing Institutes of Life Science, Chinese Academy of Science, Beijing 100101, China and [4]The Eye Hospital of Wenzhou Medical University, The State Key Laboratory Cultivation Base and Key Laboratory of Vision Science, Ministry of Health, Wenzhou 325000, China

Received July 30, 2016; Revised September 29, 2016; Editorial Decision October 19, 2016; Accepted October 22, 2016

ABSTRACT

***De novo* germline mutations (DNMs) are the rarest genetic variants proven to cause a considerable number of sporadic genetic diseases, such as autism spectrum disorders, epileptic encephalopathy, schizophrenia, congenital heart disease, type 1 diabetes, and hearing loss. However, it is difficult to accurately assess the cause of DNMs and identify disease-causing genes from the considerable number of DNMs in probands. A common method to this problem is to identify genes that harbor significantly more DNMs than expected by chance, with accurate background DNM rate (DNMR) required. Therefore, in this study, we developed a novel database named mirDNMR for the collection of gene-centered background DNMRs obtained from different methods and population variation data. The database has the following functions: (i) browse and search the background DNMRs of each gene predicted by four different methods, including GC content (DNMR-GC), sequence context (DNMR-SC), multiple factors (DNMR-MF) and local DNA methylation level (DNMR-DM); (ii) search variant frequencies in publicly available databases, including ExAC, ESP6500, UK10K, 1000G and dbSNP and (iii) investigate the DNM burden to prioritize candidate genes based on the four background DNMRs using three statistical methods (TADA, Binomial and Poisson test). As a case study, we successfully employed our database in candidate gene prioritization for a sporadic complex disease: intellectual disability. In conclusion, mirDNMR (https://www.wzgenomics.cn/mirdnmr/) can be widely used to identify the genetic basis of sporadic genetic diseases.**

INTRODUCTION

De novo mutations (DNMs) arise spontaneously either in germline cells (*de novo* germline mutation) or shortly after fertilization (post-zygotic mutation). DNMs represent extremely rare genetic variants that contribute to sporadic genetic diseases, such as autism spectrum disorders (1,2), intellectual disability (3,4), epileptic encephalopathy (5–7), schizophrenia (8–10), mental retardation (4,11), amyotrophic lateral sclerosis (12,13), congenital heart disease (14), type 1 diabetes (15) and hearing loss (16,17). Recently, trio-based whole exome/genome sequencing (WES/WGS) was found to be the best way to identify DNMs in probands with the rise of next-generation sequencing. However, not all DNMs cause sporadic disease. For a given proband, an average of 74 *de novo* single nucleotide variants (SNVs) and three *de novo* insertions/deletions (INDELs) arise spontaneously across the genome (18), and few are considered pathogenic. Therefore, the challenge is to accurately identify pathogenic DNMs and disease genes among the numerous DNMs detected in probands.

DNMs in the same gene within large cohorts repeatedly detected by trio-based WES/WGS indicate disease risk (5,19,20). However, multiple DNMs can be found in one gene by chance, and this often happens in larger genes or within mutation hotspots in the case of large sequencing samples (18). Therefore, identification of disease genes solely based on recurrent DNMs can yield false positives and requires statistic methods for reliable inference. Sequencing of well-matched control samples has identified candidate genes with significantly recurrent DNMs in several large-scale studies (1,21). However, direct case-control comparisons to evaluate significantly recurrent DNM may

[*]To whom correspondence should be addressed. Tel: +86 577 88831309; Email: iamwujy@gmail.com
Correspondence may also be addressed to Wei Li. Email: liweiwzmc@163.com
[†]These authors contributed equally to the paper as first authors.

lack statistical power because of the scarcity of DNMs in each gene (18). Furthermore, sequencing of sufficient control samples is not feasible considering current WES/WGS costs. In principle, accurate estimates of background DNM rates (DNMRs) for each gene could be used to calculate significantly excess of DNMRs (22). In fact, gene-specific DNMRs have already been calculated by several studies using different models, and they have successfully identified disease-causal candidate genes via statistical analysis of DNMRs (2,23–25).

In this study, we constructed a database named mirDNMR for gene level background DNMRs predicted by four different methods based on GC content (DNMR-GC), sequence context (DNMR-SC), multiple factors (DNMR-MF), and local DNA methylation level (DNMR-DM). Variant frequencies in human genetic variation databases were also incorporated into mirDNMR, including ExAC, ESP6500, UK10K, 1000G and dbSNP. Overall, the mirDNMR is a database of gene-centered background DNMRs and population variations, with several functionalities for disease-causal gene prioritization included.

DATA COLLECTION AND METHODS

Data source

The background DNMRs calculated using four different methods (DNMR-GC(23), DNMR-SC(2), DNMR-MF(24) and DNMR-DM) were obtained from three published works and one of our unpublished work (Figure 1, Supplementary materials). DNMR-GC was obtained from work by Sanders *et al.* (23). Briefly, at first, DNMRs for each nucleotide were calculated based on actual sequencing data. As was reported that the average DNMR of GC bases are 1.76-fold greater than that of AT bases (26), the gene-based background DNMRs was calculated considering both gene sizes and GC contents (26). DNMR-SC was obtained from work by Samocha et al (2). DNMR-SC was predicted based on tri-nucleotide sequence context. Based on human–chimpanzee intergenic genome regions sequence comparisons, a mutation rate matrix was constructed to determine the mutation rates of each type of tri-nucleotide variation. The gene-level DNMRs were calculated by summing up DNMRs of all coding nucleotides for different mutation types, separately. DNMR-MF was predicted based on primate substitution rates, and then adjusted by sequence context, transcription strand and recombination rate to obtain final DNMR. The primate substitution rates were calculated based on human-chimpanzee comparisons on 1-MB genomic region scales. The contribution of multiple other factors to DNMR was estimated based on actual sequencing data of 250 trios. Gene level DNMR was determined by DNMR of the 1-MB genomic region overlapping with the gene mid-point for different variant types. DNMR-DM was obtained from our unpublished results, which was predicted based on local DNA methylation levels of human sperm. Because spontaneous deamination of 5-methylcytosine results in ∼14-fold higher C>T substitution rates than the genome-wide average, we assumed that DNA methylation increases base substitution variants of C>T and built a DNMR model accordingly for each gene. All of the four background DNMRs were incorporated into the mirDNMR database. Furthermore, we calculated background DNMRs of LoF, missense and synonymous variants for DNMR-SC, DNMR-MF and DNMR-DM, respectively. In addition, a unified DNMR (named as DNMR-average) was calculated by averaging the four background DNMRs (Supplementary Figure S1). The high correlation (Pearson coefficient > 0.9) among the four background DNMRs indicates the uniformity. Meanwhile, the four background DNMRs and DNMR-average show high correlation (Pearson coefficient ≥ 0.9) with number of rare SNVs (AF < 0.01) in ExAC, with DNMR-average the most highly correlated (Pearson coefficient $= 0.935$), which was set as default (Supplementary Figure S2).

Variant frequencies in human genetic variation databases, including ExAC (27), ESP6500 (28), UK10K (29), 1000G (30,31) and dbSNP (32), were included in mirDNMR (Figure 1). The Exome Aggregation Consortium (ExAC, version r0.3.1) contains a wide variety of large-scale sequencing data from 60 706 unrelated individuals, including African American, East Asian, Finnish, Non-Finnish European, South Asian and others. The NHLBI GO Exome Sequencing Project (ESP, version ESP6500SI-V2) supplied exome sequencing for 6503 samples, including European Americans and African Americans. The UK10K project identified rare genetic variants in 10 000 samples, including 4000 whole-genome cohorts, 3000 neurodevelopment sample sets, 2000 obesity sample sets and 1000 rare diseases sample sets. The 1000 Genomes (1000G, Phase 3) performed whole genome sequencing of 2504 individuals from 26 different populations. In addition, germline variations from the dbSNP database (build 147) were also included in mirDNMR.

Variant annotation

The genetic variants were annotated using the ANNOVAR program (33) as for gene region, variant effect, amino acid change, cytoband, etc. To determine the damaging effect, all variants were classified into three categories: (1) Loss-of-Function (LoF) variants including frameshift INDELs, splicing SNVs, stop-gain or stop-loss variant that could lead to protein function disruption; (2) tolerant variants including synonymous SNVs and non-frameshift INDELs (3); missense SNVs that was difficult to determine their relationship to protein function and accounted for the greatest proportion of all variants. In the mirDNMR database, 14 computational methods to predict damaging effects, including SIFT, Polyphen2_hvar, Polyphen2_hdiv, GERP++, PhyloP, SiPhy, RadialSVM, MetaLR, MutationTaster, MutationAssessor, LRT, VEST3, CADD and FATHMM were incorporated to predict the severity of missense SNVs. The mutation was finally classified as damaging only if at least nine results achieved good agreement.

Prioritize candidate genes based on DNM burden

To identify potential candidate genes based on over-occurrence of DNMs, three different statistical methods incorporated in mirDNMR for this purpose were: TADA that was published (22), Binomial test, and Poisson test. TADA

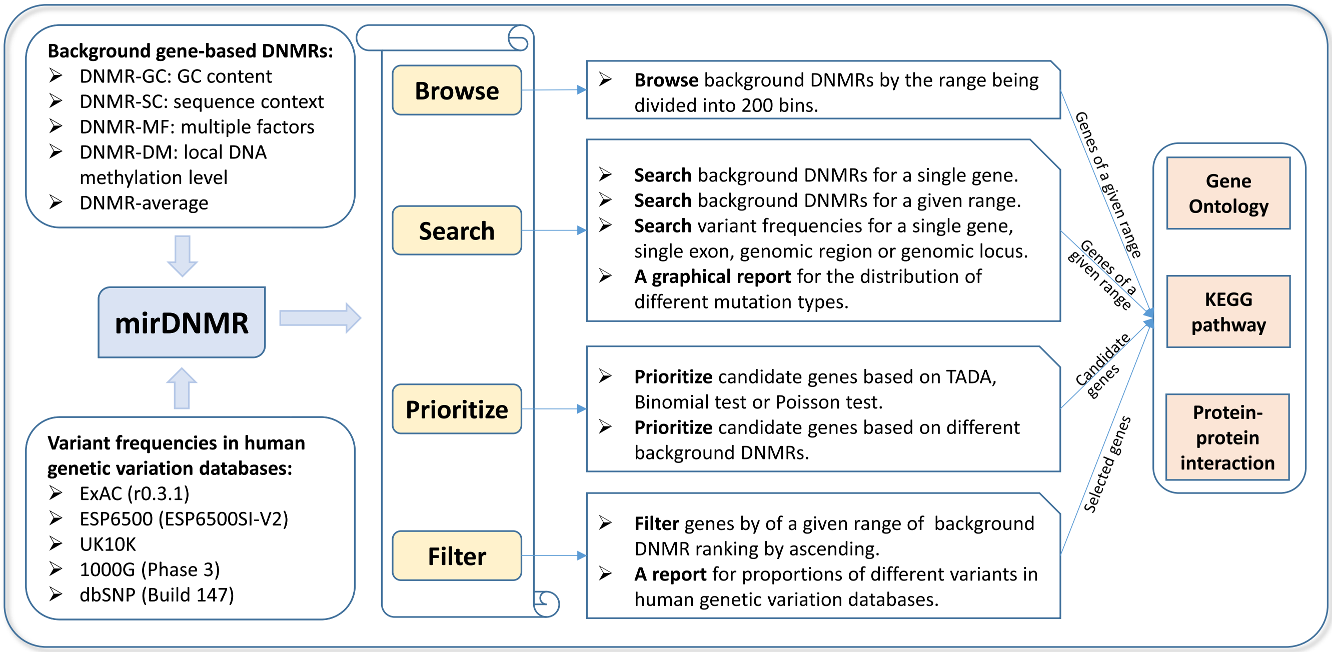

Background gene-based DNMRs:
DNMR-GC: GC content
DNMR-SC: sequence context
DNMR-MF: multiple factors
DNMR-DM: local DNA methylation level
DNMR-average
mirDNMR
Variant frequencies in human genetic variation databases:
ExAC (r0.3.1)
ESP6500 (ESP6500SI-V2)
UK10K
1000G (Phase 3)
dbSNP (Build 147)
Browse
Search
Prioritize
Filter
Browse background DNMRs by the range being divided into 200 bins.
Search background DNMRs for a single gene.
Search background DNMRs for a given range.
Search variant frequencies for a single gene, single exon, genomic region or genomic locus.
A graphical report for the distribution of different mutation types.
Prioritize candidate genes based on TADA, Binomial test or Poisson test.
Prioritize candidate genes based on different background DNMRs.
Filter genes by of a given range of background DNMR ranking by ascending.
A report for proportions of different variants in human genetic variation databases.
Genes of a given range
Genes of a given range
Candidate genes
Selected genes
Gene Ontology
KEGG pathway
Protein-protein interaction

Figure 1. The flowchart of mirDNMR. mirDNMR is a gene-centered database incorporating four different background DNMRs and variant frequencies in five human genetic variation databases. Four functions are in this database, which allow users to retrieve background DNMRs and variant frequencies in normal populations and to prioritize candidate genes. GO, KEGG pathway and PPI analysis are also provided for annotation of candidate genes.

prioritizes candidate genes using a Bayesian model to effectively combine the *de novo* LoF variants and the *de novo* damaging missense variants (predicted as 'deleterious' or 'conserved' by at least 9 of the 14 methods) and compared the observed DNMs with the background DNMRs. The Binomial test used an R function 'binom.test(x, n, p, alternative = 'greater'),' where 'x' refers to the number of *de novo* variants in each case, 'n' refers to '(number of trios) × 2,' and 'p' refers to background DNMRs specified by users. The Poisson test used an R function 'poisson.test(x, n, p, alternative = 'greater'),' where 'x,' 'n' and 'p' of the function are the same as those of the Binomial test. For the three methods, *P*-values were adjusted by the FDR approach to obtain the q values.

GO, KEGG pathway and PPI analysis

To further explore the biological functions and interactions among candidate genes, we incorporated Gene Ontology (GO) (34), KEGG pathway (35) and protein–protein interaction (PPI) into mirDNMR to allow for enrichment analysis (Figure 1). GO and KEGG pathway enrichment were run using a bioconductor package named 'GOstats' in an R environment. GO and KEGG pathway databases were originated from the 'org.Hs.eg.db' package. The GO term and KEGG pathway enrichment tests used a hypergeometric test, and *P* values were adjusted by the FDR approach (*q* value <0.05 by default). For PPI analysis, interactions between two candidate genes were retrieved from the BioGRID database (Version 3.4.138) (36).

Database construction

The web interface of mirDNMR is an Apache environment based on a CentOS release 6.5 Linux operating system. All the data were managed using the MySQL database and the web interface for database browsing. The result pages were generated using PHP scripts. To ensure data security, the web interface used a secure https protocol. The database has been successfully tested with Microsoft Internet Explorer 11.0, Firefox 38, Google Chrome 45 and Safari 5.1.

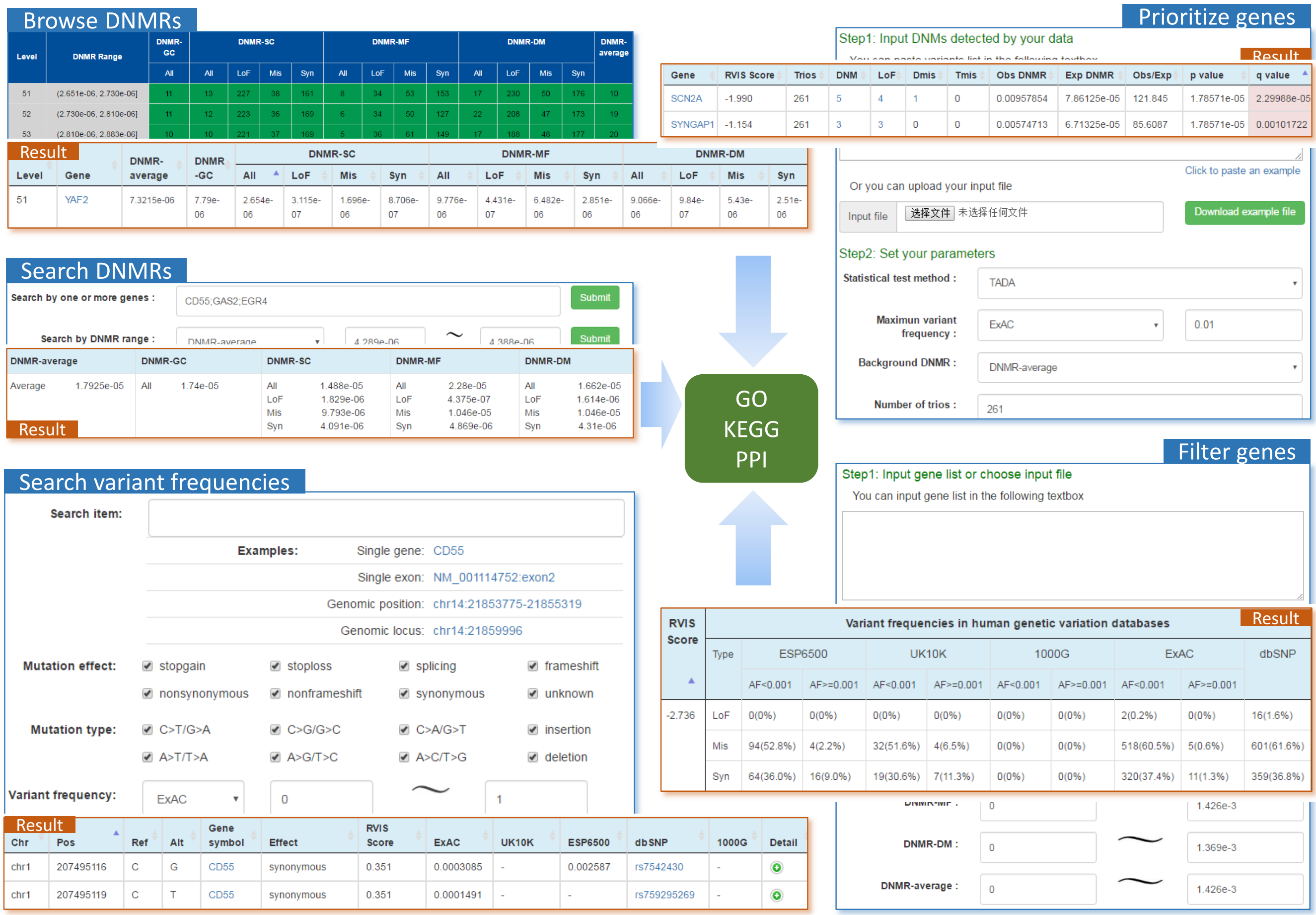
WEB INTERFACE

Browse background DNMRs

For users to easily browse background DNMRs, the four background DNMRs of specified variant types (LoF, missense, synonymous and combined) and their averages (DNMR-average) for 20 524 RefSeq genes are presented as a table on the web page. All genes were divided into 200 sets based on the DNMR range from 0 to 1.426 × 10^{-3}, and assigned names as level 1 to level 200. Counts of gene within a given DNMR range were filled in a specified cell (Figure 2). Users can view detailed gene information in a given DNMR range by clicking on the number in a specified cell. Detailed information includes a bar plot with the DNMR tendency and a table listing four types of DNMRs in ascending order for all genes at this level.

Search background DNMRs and variant frequencies in normal populations

In mirDNMR, users can search for the background DNMRs for each gene by gene name or DNMR range (Figure



Browse DNMRs

Level	DNMR Range	DNMR-GC	DNMR-SC				DNMR-MF				DNMR-DM				DNMR-average
		All	All	LoF	Mis	Syn	All	LoF	Mis	Syn	All	LoF	Mis	Syn	
51	(2.651e-06, 2.730e-06]	11	13	227	38	161	8	34	53	153	17	230	50	176	10
52	(2.730e-06, 2.810e-06]	11	12	223	36	169	6	34	50	127	22	208	47	173	19
53	(2.810e-06, 2.883e-06]	10	10	221	37	169	5	36	61	149	17	188	48	177	20

Result

Level	Gene	DNMR-average	DNMR-GC	DNMR-SC				DNMR-MF				DNMR-DM			
				All	LoF	Mis	Syn	All	LoF	Mis	Syn	All	LoF	Mis	Syn
51	YAF2	7.3215e-06	7.79e-06	2.654e-06	3.115e-07	1.696e-06	8.706e-07	9.776e-06	4.431e-07	6.482e-06	2.851e-06	9.066e-06	9.84e-07	5.43e-06	2.51e-06

Search DNMRs

Search by one or more genes : CD55;GAS2;EGR4 Submit

Search by DNMR range : DNMR-average ▾ 4.289e-06 ~ 4.388e-06 Submit

DNMR-average		DNMR-GC		DNMR-SC		DNMR-MF		DNMR-DM	
Average	1.7925e-05	All	1.74e-05	All	1.488e-05	All	2.28e-05	All	1.662e-05
				LoF	1.829e-06	LoF	4.375e-07	LoF	1.614e-06
				Mis	9.793e-06	Mis	1.046e-05	Mis	1.046e-05
				Syn	4.091e-06	Syn	4.869e-06	Syn	4.31e-06

Result

Search variant frequencies

Search item:

Examples: Single gene: CD55

Single exon: NM_001114752:exon2

Genomic position: chr14:21853775-21855319

Genomic locus: chr14:21859996

Mutation effect: ☑ stopgain ☑ stoploss ☑ splicing ☑ frameshift ☑ nonsynonymous ☑ nonframeshift ☑ synonymous ☑ unknown

Mutation type: ☑ C>T/G>A ☑ C>G/G>C ☑ C>A/G>T ☑ insertion ☑ A>T/T>A ☑ A>G/T>C ☑ A>C/T>G ☑ deletion

Variant frequency: ExAC ▾ 0 ~ 1

Result

Chr	Pos	Ref	Alt	Gene symbol	Effect	RVIS Score	ExAC	UK10K	ESP6500	dbSNP	1000G	Detail
chr1	207495116	C	G	CD55	synonymous	0.351	0.0003085	-	0.002587	rs7542430	-	⊕
chr1	207495119	C	T	CD55	synonymous	0.351	0.0001491	-	-	rs759295269	-	⊕

Prioritize genes

Step1: Input DNMs detected by your data

You can paste variants list in the following textbox

Result

Gene	RVIS Score	Trios	DNM	LoF	Dmis	Tmis	Obs DNMR	Exp DNMR	Obs/Exp	p value	q value
SCN2A	-1.990	261	5	4	1	0	0.00957854	7.86125e-05	121.845	1.78571e-05	2.29988e-05
SYNGAP1	-1.154	261	3	3	0	0	0.00574713	6.71325e-05	85.6087	1.78571e-05	0.00101722

Click to paste an example

Or you can upload your input file

Input file 选择文件 未选择任何文件 Download example file

Step2: Set your parameters

Statistical test method : TADA ▾

Maximun variant frequency : ExAC ▾ 0.01

Background DNMR : DNMR-average ▾

Number of trios : 261

Filter genes

Step1: Input gene list or choose input file

You can input gene list in the following textbox

Result

RVIS Score	Variant frequencies in human genetic variation databases									
	Type	ESP6500		UK10K		1000G		ExAC		dbSNP
		AF<0.001	AF>=0.001	AF<0.001	AF>=0.001	AF<0.001	AF>=0.001	AF<0.001	AF>=0.001	
-2.736	LoF	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	2(0.2%)	0(0%)	16(1.6%)
	Mis	94(52.8%)	4(2.2%)	32(51.6%)	4(6.5%)	0(0%)	0(0%)	518(60.5%)	5(0.6%)	601(61.6%)
	Syn	64(36.0%)	16(9.0%)	19(30.6%)	7(11.3%)	0(0%)	0(0%)	320(37.4%)	11(1.3%)	359(36.8%)

DNMR-MF : 0 ~ 1.426e-3

DNMR-DM : 0 ~ 1.369e-3

DNMR-average : 0 ~ 1.426e-3

Figure 2. An example of mirDNMR use. For the 'Browse' function, all four background DNMRs and DNMR-average were divided into 200 bins based on magnitude ranging from 0 to 1.426e-03. Users can search background DNMRs by gene or DNMR range. Users can also search variant frequencies in human genetic variation databases for a gene, exon, genomic region, or locus. With an input DNM list, users can prioritize candidate genes based on TADA, Poisson test, or Binomial test using one of the four background DNMRs. Using a gene list, users can prioritize candidate genes based on background DNMRs, RVIS score or the distribution of different variant types in human genetic variation databases. For a given gene list generated by these functions, GO, KEGG pathway and PPI annotations can also be performed in mirDNMR.

2). For the search of gene name, one or more genes (separated by semicolon) are acceptable for input. In a gene search result, four background DNMRs of specified variant type (LoF, missense, synonymous and combined) and DNMR-average are displayed in a table. Meanwhile, a jitter plot is also provided to show the global distribution of DNMRs of the given gene. For a search using DNMR range, a table listing all genes within this range is shown.

mirDNMR supports different input for searching variant frequencies in the five human genetic variation databases (ExAC, ESP6500, UK10K, 1000G, dbSNP) (Figure 2). Users can perform a search by inputting a single gene, single exon, genomic region or locus. The search result can be filtered by selecting different variant effects (including stop-gain, stop-loss, splicing site, frameshift, non-synonymous, non-frameshift, synonymous and unknown), variant types (including C>T/G>A, C>G/G>C, C>A/G>T, A>T/T>A, A>G/T>C, A>C/T>G, insertion and deletion), and using a custom range of variant frequency of a given database. The search result contains three blocks: a summary of the variant counts in the five human genetic variation databases, a pie chart for the distribution of variant effects and variant types, and a table listing all variants that passed the filtering process. For each variant in the table, detailed information is retrieved by clicking the '+' in the last column, including the variant frequency of different population in each database and the damaging prediction result from the 14 software.

Prioritize candidate genes

mirDNMR allows users to prioritize candidate genes based on DNM burden of observed DNMR and background DNMR (Figure 2). After DNM list uploaded, users should define the statistical methods (TADA, Binomial test, and Poisson test) and background DNMR for analysis. TADA prioritizes candidate genes by integrating the counts of LoF and damaging missense variants that can be determined by the 14 software methods. Weights for LoF and missense variants could be freely set by users according to different datasets. Of note, users can select their interested software and threshold of the number of damaging predictions to determine whether a missense variant is damaging. For the three statistical methods, P values are adjusted by the FDR approach to obtain the q values. In the gene prioritization result, a table of candidate genes is shown, containing their RVIS score (37), counts of different types of variants, observed DNMR, background DNMR, P value, and q value in ascending order. Users can get the detailed list of DNMs by clicking the number of DNMs in the table. Users can also run GO, KEGG pathway, and PPI annotation for their candidate genes by clicking the button on the top of the results page (Figure 2).

Filter genes with custom range

mirDNMR provides users the ability to filter genes based on the range of background DNMRs, RVIS score and variant frequencies in human genetic variation databases (Figure 2). Unlike the 'Prioritize' function, this utility prioritizes genes from a gene list. On the results page, genes remaining after this filtering process are listed by ascending order of DNMRs. Meanwhile, the RVIS scores and the distribution of the counts of LoF, missense, and synonymous variants in human genetic variation databases in both rare ($AF< 0.001$) and common ($AF \geq 0.001$) forms are shown.

Case study

To demonstrate the power of mirDNMR in gene prioritization, we used DNMs from WES/WGS of 1,031 trios affected with intellectual disability and 982 normal control trios from the NPdenovo database (38) to prioritize candidate genes as a case study (Supplementary Table S1). Based on the TADA method, genes were prioritized using the four background DNMRs for intellectual disability and controls. We defined the genes with a q value <0.1 which were jointly identified based on the four background DNMRs as candidate genes. In the result, we identified 46 candidate genes for intellectual disability but no candidate gene for control (Figure 3A and B, Supplementary Tables S2 and S3). Among the 46 candidate genes, *POGZ, SCN2A, CTNNB1, GATAD2B, TCF20* and *SYNGAP1* are the most significant ones (q value < 0.0001 based on the four background DNMRs), and they are strongly correlated with intellectual disability based on previous studies (Figure 3B) (39–42). Several other genes, such as *SETBP1, TBR1, WAC, STXBP1* and *MED13L*, were also found to underlie the pathology of intellectual disability (43). Several genes, such as *SCN2A, SYNGAP1, DLG4, STXBP1* and *GRIN2A*, were not only correlated with intellectual disability, but also involved in other neuropsychiatric disorders, such as autism spectrum disorder and epileptic encephalopathy (Supplementary Table S2).

To further explore the functional relevance of the 46 candidate genes, we performed GO term enrichment annotation (Figure 3C, Supplementary Table S4). As a result, for biological processes, 18 of the 46 genes (*CTNNB1, DLG4, DYRK1A, FOXG1, GRIN2A, KDM2B, PPP2R5D, PURA, SATB2, SCN2A, SCN8A, SLC6A1, SMARCA4, SOX5, STXBP1, SYNGAP1, TBR1, TCF4*) are enriched in four GO terms (GO:0007399, GO:0031644, GO:0051969, GO:0048169), suggesting that these genes are important in the formation of neurons and synaptic transmission processes. For molecular function, 26 of the 46 genes (*AHDC1, CTCF, CTNNB1, FOXG1, FOXP1, GATAD2B, GRIN2A, KCNH1, KDM2B, MED13L, POGZ, PURA, SATB2, SCN2A, SCN8A, SETBP1, SMARCA4, SON, SOX5, SRCAP, TBR1, TCF20, TCF4, TRIP12, USP7, WAC*) are enriched in five GO terms (GO:0003677, GO:0008134, GO:0003712, GO:0005244, GO:0003682), indicating that these genes are involved in gated channel activity and gene transcription. For cellular components, eight of the 46 genes (*CTNNB1, DLG4, GRIN2A, PURA, SCN2A, SCN8A, SLC6A1, SYNGAP1*) are enriched in three GO terms (GO:0043005, GO:0030424, GO:0030425), indicating that these genes are involved in the composition of neuron. Overall, we could see that the candidate genes identified based on the mirDNMR database are quite relevant to the studied individuals' phenotypes, which prove its efficiency.

DISCUSSION AND PERSPECTIVES

DNMs are widely used to identify genes underlying various genetic disorders based on trio-based WES/WGS of large-scale sporadic cases (1,18,43). However, accurate identification of actual disease-causal genes using DNMs from probands is complicated and challenging. A common solution to this problem is to identify genes with significantly more DNMs than expected by chance (25). Therefore, background DNMR for each gene is required to calculate significance of DNM recurrence. In this study, we have constructed a novel database named mirDNMR, which provides gene-centered background DNMRs predicted by four different methods: DNMR-GC, DNMR-SC, DNMR-MF and DNMR-DM. Meanwhile, mirDNMR also provides population genetic variants from the five largest population variation databases to assist with gene prioritization: ExAC, ESP6500, UK10K, 1000G and dbSNP. The three DNMR prediction methods were calculated according to previous studies using different models based on GC content (DNMR-GC) (23), tri-nucleotide sequencing context (DNMR-SC) (2) and multiple factors (DNMR-MF) (24). Another DNMR prediction method: DNMR-DM is developed by us based on observed strong correlation between DNMRs and human sperm DNA methylation level. The four main functions in mirDNMR are: browse, search, prioritize and filter. Users can conveniently retrieve background DNMRs and variant frequencies in human genetic variation databases for certain genes by 'Browse' and 'Search' functions. Meanwhile, mirDNMR provides two user functions that prioritize candidate genes based on DNM burden and filters genes of interest based on background DNMR and the distribution of different variant types in human genetic variation databases. In conclusion, the mirDNMR database incorporates expected DNMRs, population genetic variation data for each gene and an interface with functions assisting with disease-causal gene prioritizations based on DNMs detected from trio-based WGS/WES data.

As an ongoing project, mirDNMR will be updated regularly and incorporate new DNMR prediction methods in the future. For the moment, only summarized background DNMRs on coding regions and for gene level are considered by mirDNMR. The declining cost of WGS will enable DNM studies on non-coding regions for larger sam-

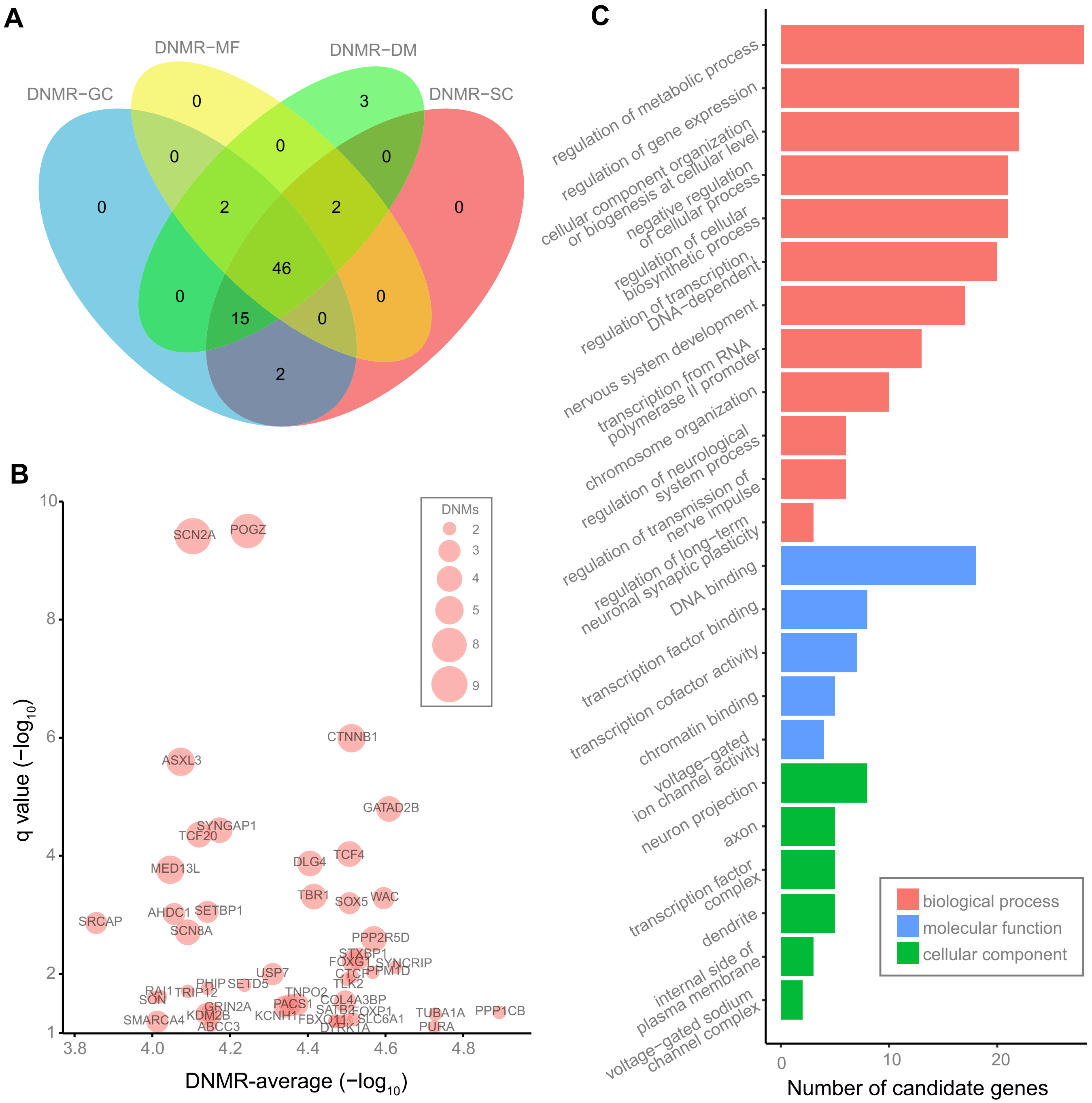


Figure 3. Prioritization of candidate genes for intellectual disability from trio-based WES/WGS. Based on the TADA method, genes were prioritized using the four background DNMRs. (**A**) Forty six genes with *q* values <0.1 were shared by the four background DNMRs in intellectual disability trios. (**B**) A scatter diagram for the 46 intellectual disability candidate genes. The size of each point indicates the total number of LoF and damaging missense DNMs for each gene. (**C**) Relative enriched GO terms (*q* value <0.05) of the 46 candidate genes for intellectual disability. Detailed information for each GO term is shown in Supplementary Table S4.

ple size. Therefore, in the future mirDNMR will provide background DNMR of non-coding regions, and enrichment analysis on nucleotide-levels. Any questions, comments, and suggestions are welcome, which will help future updates. We expect that mirDNMR will serve as a valuable resource for the research community working on identification of genetic variation underlying human diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of the Institute of Genomic Medicine of Wenzhou Medical University for their valuable discussions of this work.

FUNDING

National Natural Science Foundation of China [31571357/C060702]; National Key Basic Research Program [2013CB967502]; National Natural Science Foundation of China [81271918]. Funding for open access charge: National Natural Science Foundation of China [31571357/C060702]; National Key Basic Research Program [2013CB967502]; National Natural Science Foundation of China [81271918].
Conflict of interest statement. None declared.

REFERENCES

1. Iossifov,I., O'Roak,B.J., Sanders,S.J., Ronemus,M., Krumm,N., Levy,D., Stessman,H.A., Witherspoon,K.T., Vives,L., Patterson,K.E. *et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**, 216–221.
2. Samocha,K.E., Robinson,E.B., Sanders,S.J., Stevens,C., Sabo,A., McGrath,L.M., Kosmicki,J.A., Rehnstrom,K., Mallick,S., Kirby,A. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.
3. 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.
4. Gilissen,C., Hehir-Kwa,J.Y., Thung,D.T., van de Vorst,M., van Bon,B.W., Willemsen,M.H., Kwint,M., Janssen,I.M., Hoischen,A., Schenck,A. *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**, 344–347.
5. Syrbe,S., Hedrich,U.B., Riesch,E., Djemie,T., Muller,S., Moller,R.S., Maher,B., Hernandez-Hernandez,L., Synofzik,M., Caglayan,H.S. *et al.* (2015) De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat. Genet.*, **47**, 393–399.
6. Epi,K.C., Allen,A.S., Berkovic,S.F., Cossette,P., Delanty,N., Dlugos,D., Eichler,E.E., Epstein,M.P., Glauser,T., Epilepsy Phenome/Genome, P. *et al.* (2013) De novo mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.
7. Nava,C., Dalle,C., Rastetter,A., Striano,P., de Kovel,C.G., Nabbout,R., Cances,C., Ville,D., Brilstra,E.H., Gobbi,G. *et al.* (2014) De novo mutations in HCN1 cause early infantile epileptic encephalopathy. *Nat. Genet.*, **46**, 640–645.
8. Xu,B., Ionita-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.
9. 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) De novo mutations in schizophrenia implicate synaptic networks. *Nature*, **506**, 179–184.
10. Gulsuner,S., Walsh,T., Watts,A.C., Lee,M.K., Thornton,A.M., Casadei,S., Rippey,C., Shahin,H., Nimgaonkar,V.L. and Go,R.C. (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, **154**, 518–529.
11. Hamdan,F.F., Srour,M., Capo-Chichi,J.M., Daoud,H., Nassif,C., Patry,L., Massicotte,C., Ambalavanan,A., Spiegelman,D., Diallo,O. *et al.* (2014) De novo mutations in moderate or severe intellectual disability. *PLoS Genet.*, **10**, e1004772.
12. 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. *Scientific Rep.*, **5**, 9124.
13. Chesi,A., Staahl,B.T., Jovicic,A., Couthouis,J., Fasolino,M., Raphael,A.R., Yamazaki,T., Elias,L., Polak,M., Kelly,C. *et al.* (2013) Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat. Neurosci.*, **16**, 851–855.
14. Zaidi,S., Choi,M., Wakimoto,H., Ma,L., Jiang,J., Overton,J.D., Romano-Adesman,A., Bjornson,R.D., Breitbart,R.E., Brown,K.K. *et al.* (2013) De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, **498**, 220–223.
15. Fang,C., Li,H., Li,X., Xiao,W., Huang,Y., Cai,W., Yang,Y. and Hu,J. (2016) De novo mutation of PHEX in a type 1 diabetes patient. *J. Pediatr. Endocrinol. Metab.: JPEM*, **29**, 621–626.
16. Jiang,S.J., Di,Z.H., Huang,D., Zhang,J.B., Zhang,Y.Y., Li,S.Q. and He,R. (2014) R75Q *de novo* dominant mutation of GJB2 in a Chinese family with hearing loss and palmoplantar keratoderma. *Int. J. Pediatr. Otorhinolaryngol.*, **78**, 1461–1466.
17. Moteki,H., Shearer,A.E., Izumi,S., Kubota,Y., Azaiez,H., Booth,K.T., Sloan,C.M., Kolbe,D.L., Smith,R.J. and Usami,S. (2015) *De novo* mutation in X-linked hearing loss-associated POU3F4 in a sporadic case of congenital hearing loss. *Ann. Otol. Rhinol. Laryngol.*, **124**(Suppl. 1), 169S–176S.
18. Veltman,J.A. and Brunner,H.G. (2012) De novo mutations in human genetic disease. *Nat. Rev. Genet.*, **13**, 565–575.
19. O'Roak,B.J., Vives,L., Fu,W., Egertson,J.D., Stanaway,I.B., Phelps,I.G., Carvill,G., Kumar,A., Lee,C., Ankenman,K. *et al.* (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**, 1619–1622.
20. O'Roak,B.J., Stessman,H.A., Boyle,E.A., Witherspoon,K.T., Martin,B., Lee,C., Vives,L., Baker,C., Hiatt,J.B., Nickerson,D.A. *et al.* (2014) Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat. Commun.*, **5**, 5595.
21. Yuen,R.K., Thiruvahindrapuram,B., Merico,D., Walker,S., Tammimies,K., Hoang,N., Chrysler,C., Nalpathamkalam,T., Pellecchia,G., Liu,Y. *et al.* (2015) Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.*, **21**, 185–191.
22. 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.
23. Sanders,S.J., Murtha,M.T., Gupta,A.R., Murdoch,J.D., Raubeson,M.J., Willsey,A.J., Ercan-Sencicek,A.G., DiLullo,N.M., Parikshak,N.N. and Stein,J.L. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237–241.
24. Francioli,L.C., Polak,P.P., Koren,A., Menelaou,A., Chun,S., Renkens,I., Genome of the Netherlands,C., van Duijn,C.M., Swertz,M., Wijmenga,C. *et al.* (2015) Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.*, **47**, 822–826.
25. De Rubeis,S., He,X., Goldberg,A.P., Poultney,C.S., Samocha,K., Cicek,A.E., Kou,Y., Liu,L., Fromer,M., Walker,S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**, 209–215.
26. Lynch,M. (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 961–968.
27. 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.
28. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
29. The UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.
30. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
31. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Hsi-Yang Fritz,M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
32. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7.

33. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
34. TheGene OntologyConsortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
35. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
36. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) TheBioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
37. Petrovski,S., Wang,Q., Heinzen,E.L., Allen,A.S. and Goldstein,D.B. (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**, e1003709.
38. 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 *de novo* mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry*, **21**, 298.
39. Berryer,M.H., Hamdan,F.F., Klitten,L.L., Moller,R.S., Carmant,L., Schwartzentruber,J., Patry,L., Dobrzeniecka,S., Rochefort,D., Neugnot-Cerioli,M. *et al.* (2013) Mutations in SYNGAP1 cause intellectual disability, autism, and a specific form of epilepsy by inducing haploinsufficiency. *Hum. Mutat.*, **34**, 385–394.
40. Hamdan,F.F., Gauthier,J., Spiegelman,D., Noreau,A., Yang,Y., Pellerin,S., Dobrzeniecka,S., Cote,M., Perreau-Linck,E., Carmant,L. *et al.* (2009) Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.*, **360**, 599–605.
41. Rauch,A., Wieczorek,D., Graf,E., Wieland,T., Endele,S., Schwarzmayr,T., Albrecht,B., Bartholdi,D., Beygo,J., Di Donato,N. *et al.* (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, **380**, 1674–1682.
42. Baasch,A.L., Huning,I., Gilissen,C., Klepper,J., Veltman,J.A., Gillessen-Kaesbach,G., Hoischen,A. and Lohmann,K. (2014) Exome sequencing identifies a de novo SCN2A mutation in a patient with intractable seizures, severe intellectual disability, optic atrophy, muscular hypotonia, and brain abnormalities. *Epilepsia*, **55**, e25–e29.
43. Vissers,L.E., Gilissen,C. and Veltman,J.A. (2016) Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.*, **17**, 9–18.