



GPCards: An integrated database of genotype–phenotype correlations in human genetic diseases



Bin Li^{a,c,e}, Zheng Wang^a, Qian Chen^a, Kuokuo Li^b, Xiaomeng Wang^b, Yijing Wang^b, Qian Zeng^c, Ying Han^b, Bin Lu^d, Yuwen Zhao^c, Rui Zhang^c, Li Jiang^c, Hongxu Pan^c, Tengfei Luo^b, Yi Zhang^a, Zhenghuan Fang^b, Xuwen Xiao^c, Xun Zhou^c, Rui Wang^b, Lu Zhou^c, Yige Wang^c, Zhenhua Yuan^c, Lu Xia^b, Jifeng Guo^c, Beisha Tang^{a,c}, Kun Xia^b, Guihu Zhao^{a,c,*}, Jinchun Li^{a,b,c,*}

^a National Clinical Research Center for Geriatric Disorders, Department of Geriatrics, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China

^b Center for Medical Genetics & Hunan Key Laboratory, School of Life Sciences, Central South University, Changsha Hunan 410008, China

^c Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China

^d Department of Pathogen Biology, School of Basic Medical Sciences, Central South University, Changsha, Hunan 410008, China

^e Mobile Health Ministry of Education - China Mobile Joint Laboratory, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China

ARTICLE INFO

Article history:

Received 19 November 2020

Received in revised form 28 February 2021

Accepted 10 March 2021

Available online 22 March 2021

Keywords:

GPCards
Phenotype
Genotype
Variant

ABSTRACT

Genotype–phenotype correlations are the basis of precision medicine of human genetic diseases. However, it remains a challenge for clinicians and researchers to conveniently access detailed individual-level clinical phenotypic features of patients with various genetic variants. To address this urgent need, we manually searched for genetic studies in PubMed and catalogued 8,309 genetic variants in 1,288 genes from 17,738 patients with detailed clinical phenotypic features from 1,855 publications. Based on genotype–phenotype correlations in this dataset, we developed an user-friendly online database called GPCards (<http://genemed.tech/gpcards/>), which not only provided the association between genetic diseases and disease genes, but also the prevalence of various clinical phenotypes related to disease genes and the patient-level mapping between these clinical phenotypes and genetic variants. To accelerate the interpretation of genetic variants, we integrated 62 well-known variant-level and gene-level genomic data sources, including functional predictions, allele frequencies in different populations, and disease-related information. Furthermore, GPCards enables automatic analyses of users' own genetic data, comprehensive annotation, prioritization of candidate functional variants, and identification of genotype–phenotype correlations using custom parameters. In conclusion, GPCards is expected to accelerate the interpretation of genotype–phenotype correlations, subtype classification, and candidate gene prioritisation in human genetic diseases.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Extraordinary advances in sequencing technology have resulted in major scientific breakthroughs in human genetics [1,2]. In particular, next-generation sequencing (NGS) technologies, especially whole-exome sequencing and whole-genome sequencing, have accelerated the identification of pathogenic variants and disease-causing genes in human genetic diseases [3]. NGS technologies have been effectively applied to biomedical genetics and clinical

genetics [1,3], and revolutionised the way researchers and clinicians prioritise disease-causing genes in Mendelian disorders and other human complex diseases [4]. Moreover, medical genetics still play a huge role in the diagnosis of rare diseases and promote personalised diagnosis and treatment. Experienced clinicians now combine clinical phenotypic features with molecular genetics in disease diagnosis and treatment [5].

Since the correlation between genotype and phenotype in genetic diseases was first reported decades ago [6], increasing evidence has demonstrated that patients carrying pathogenic variants in some disease-causing genes presented clinically recognisable phenotypes and accompanying syndromes [7]. Meanwhile, researchers and clinicians turned their attentions to the molecular subtypes classification of the disease based on the

* Corresponding authors at: National Clinical Research Center for Geriatric Disorders, Department of Geriatrics, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China.

E-mail addresses: ghzhao@csu.edu.cn (G. Zhao), lijinchen@csu.edu.cn (J. Li).

genotypes of patients [7]. Although amounts of variants have been discovered, the speed of interpretation lags far behind, scientists are not yet able to decipher the correlations between most variants and diseases. Many phenotypic features caused by genetic variants cannot be used for accurate clinical diagnosis and treatment. A better understanding of correlations between genotypes and phenotypes will revolutionise clinical diagnosis and treatment in patients with genetic diseases [8]. Nevertheless, data for genotype–phenotype correlations are distributed across a massive number of published studies and are therefore difficult to access and utilize. To address this problem, the appropriate integration of these distributed data is necessary, and the development of a database with aggregated information of genotype–phenotype correlations and detailed individual-level clinical phenotype with genetic variants is a key goal [9].

Several databases, such as Online Mendelian Inheritance in Man (OMIM) [10], Human Phenotype Ontology (HPO) [11], ClinVar [12], MalaCards [13], DisGeNET [Pinero, 2020 #98], Monarch [Shefchek, 2020 #99], and CentoMD [4], have been developed to catalogue disease-associated genes. However, there is no open-access database with detailed individual-level clinical phenotypic features related to genetic variants. Accordingly, we developed a comprehensive, global, open-access database of genotype–phenotype correlations, named GPCards (<http://www.genemed.tech/gpcards>). In GPCards, detailed information about genotype–phenotype correlations for individual patients with genetic variants is presented with a user-friendly interface and does not require registration. Moreover, the most well-known annotated information at the gene and variant levels is provided by easily operated links. GPCards provides an important resource for genetic counselling and disease diagnosis and treatment.

2. Material and methods

2.1. Data collection and quality control

Genotype–phenotype correlations were retrieved by manual searches of each human gene against PubMed using the search strategy “*gene symbol [Title/Abstract] AND (mutation [Title/Abstract] OR variant [Title/Abstract])*” (Fig. 1). Though all human genes were searched, only effective genetic studies were obtained according to the following inclusion criteria: (i) no fewer than three patients with detailed data for genotype–phenotype correlations and (ii) within the top five studies with respect to level of detail for genotype–phenotype correlations, if there are more than five eligible studies for a human gene. Exclusion criteria were as follows: (i) studies that focused on molecular mechanisms, rather than genetic studies; (ii) studies reporting fewer than three patients; and (iii) studies without original data for genotype–phenotype correlations, or without original phenotypic details of patients, which may cite from other published studies. We get rid of unsuitable studies by reading abstracts of the searched publications which were retrieved from PubMed according with the exclusion criteria. After that, we screened out effective genetic studies from the rest literature according with the inclusion criteria. The data collectors collected genotype and phenotype information in the literature. At last, a geneticist was assigned to reviewed and curate the genetic and phenotypic data and to confirm the accuracy of the collected data. All data collectors, who were rigorously trained to ensure the consistency of collected data, were researchers or PhD students with a strong background in clinical genetics.

For each genetic study meeting the quality control criteria described above, two types of information were collected. First, we catalogued the phenotypes associated with each disease-causing gene, including the PubMed ID, gene symbol, diagnostic

diseases, total number of patients with genetic variants in a given gene, and number of patients with each clinical phenotype or symptom (Fig. 1). Second, we catalogued the detailed phenotypic features and genotypes of each patient, including the PubMed ID, sample ID, Mendelian inheritance (recessive or dominant), genomic position of each variant, nucleotide change, amino acid change, origin of variants (*de novo* or inherited), types of variants (homozygous or heterozygous), gender, and status of each phenotypic features or symptoms (Fig. 1). LiftOver was employed to convert the genomic position from one genome assembly (hg18 or hg38) to the genome assembly hg19. If the genomic positions of genetic variants were not available in the original studies, VarCards [14] was used to match the genomic positions based on definitions of transcripts from RefSeq.

2.2. Variant annotation and integration

ANNOVAR was used for the comprehensive annotation of genetic variants in each study (Fig. 1). The allele frequencies in different populations were extracted from various human genetic variation databases, such as gnomAD (release 2.1.1) [15,16], ExAC (release 1.0) [15,17], ESP6500 (release ESP6500SI-V2) [18]; 1000 Genomes Project (final phase of the project) [19]; Kaviar genomic variant database (version 160204-Public) [20], and Haplotype Reference Consortium (HRC) (15). The predicted pathogenicity of missense variants was also evaluated using 24 widely accepted algorithms, including ReVe [21], REVEL [22], SIFT [23,24], PolyPhen2 HDIV [25], PolyPhen2 HVAR [25], LRT [26], MutationTaster [27], MutationAssessor [28], FATHMM [29], PROVEAN [30], VEST 3.0 [31], MetaSVM [32], MetaLR [32], M-CAP [33], CADD [34], DANN [35], FATHMM MKL [36], Eigen [37], GenoCanyon [37], fit-Cons [38], GERP++ [39], PhyloP [40], PhastCons [41], and SiPhy [42]. Some disease-related information for variants was also annotated, including InterVar [43] (103), COSMIC [44], ICGC [45], nci60, InterPro [46], dbSNP v150 [47], and ClinVar [12].

Comprehensive annotations were also performed at the gene level, as described in our previous studies [14,48], including six data types: basic information, gene function, phenotype and disease, gene expression, variants in different populations, and drug–gene interactions (Fig. 1). In the panel of basic information, core gene-level information was extracted from NCBI Gene [49], Gene Ontology (GO; V1.4) [50] (113), and InBio Map (release 20160912) [51]. Data were obtained for the intolerance score (RVIS) [52], novel gene intolerance ranking system LoFtool [53], heptanucleotide context intolerance score [54], gene damage index (GDI) [55], Episcore [56], and the probability of loss of function intolerance score [15]. In the gene function panel, core information from UniProt (release 201902) [57], InterPro [46], InBio Map (release 20160912) [51], and NCBI BioSystems (release 20170421) [58] was integrated. In the phenotype and disease-related information panel, the gene-level information from OMIM [10], ClinVar [12], Gene4Denovo [48], MGI [59], and HPO [51] was catalogued. In the gene expression panel, data were sourced from BrainSpan [60], GTEx [61], and the Human Protein Atlas [62]. The final panel included drug–gene interaction data and gene druggability in the drug–gene interaction from DGIdb [63].

2.3. Database construction and interfaces

Integrating all of the information for genotype–phenotype correlations and comprehensive annotations described above, GPCards (<http://genemed.tech/gpcards/>) was developed by combining Vue with a PHP-based web framework Laravel to construct a user-friendly web interface (Fig. 1). The front and back models were separated for construction. The UI Toolkit Element, supporting most modern browsers across platforms, such as Microsoft

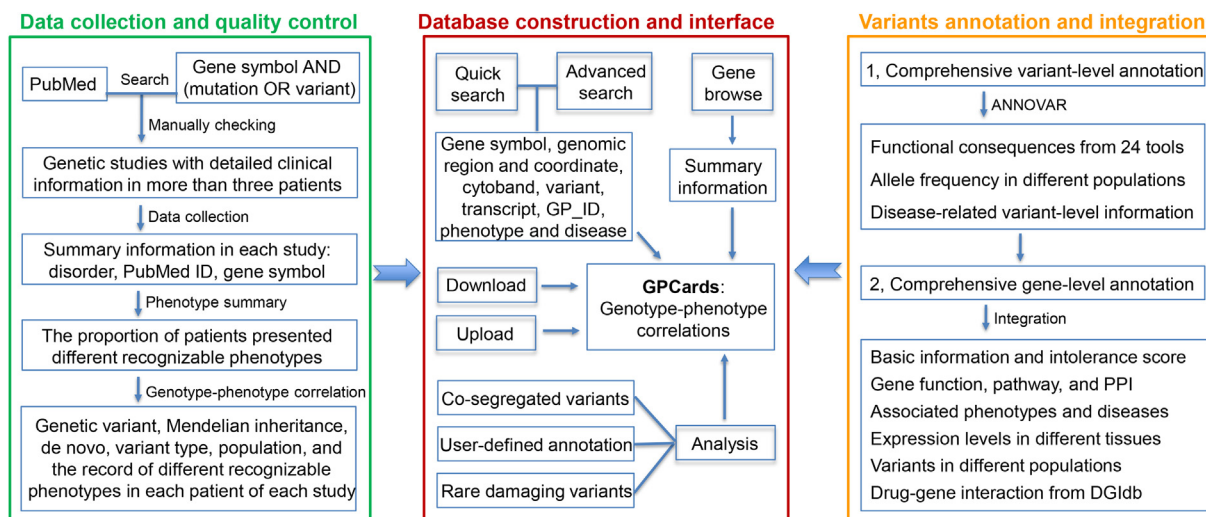


Fig. 1. A general workflow of GPCards. Data collection and quality control information were showed in green box; Variants annotation and integration flow chart was listed in yellow box; and database construction and interface were exhibited in red box. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Edge, Google Chrome, and Safari, was used. The back-end was developed using Laravel, a common PHP web framework. GPCards could run smoothly and is compatible with multiple operating systems, including Windows, Mac, and Linux. Finally, all genotype-phenotype data and annotation data were stored in the MySQL database.

2.4. Database update

The data in GPCards will be updated semi-annually, by manually searching the genotype-phenotype correlations in PubMed with search strategy “(phenotype [Title/Abstract] or clinical feature [Title/Abstract]) AND (mutation [Title/Abstract] OR variant [Title/Abstract])” accompany with published data limited from the latest update time. We also encourage users upload original phenotype-genotype data, which have been anonymized in Upload section of GPCards.

3. Results and web interface

3.1. Summary of catalogued data for genotype-phenotype correlations

We reviewed more than 20,000 studies from PubMed and 1,855 genetic studies with detailed phenotype information satisfying the quality control requirements were finally integrated. In total, 8,309 nonredundant genetic variants in 1,288 genes from 17,738 patients with formatted and detailed clinical phenotypic features were integrated into the GPCards database. For these 1,288 disease-associated genes with individual-level detailed phenotypic features, 119 (10.9%), 92 (8.4%), 59 (5.4%), and 436 (39.9%) were reported to carry three, four, five and no fewer than six genetic variants, respectively (Fig. 2A). Among the 1,855 studies, 129 (7.9%), 175 (10.8%), 131 (8.1%), and 1,070 (65.9%) described two, three, four, and no fewer than five phenotypic features for each patient, respectively (Fig. 2B). Moreover, 220 (13.4%), 185 (11.3%), 190 (11.6%) and 1,049 (63.8%) studies described three, four, five and no fewer than six patients, respectively (Fig. 2C). Furthermore, for 17,738 patients with clinical information, we found that 1,200 (8.0%), 1,795 (12.0%), 1,253 (8.4%), and 9,619 (64.3%) patients had two, three, four, and no fewer than five clinical phenotypic features, respectively (Fig. 2D).

3.2. Search modules of GPCards

To facilitate the mining and application of genotype-phenotype correlation data, the GPCards database was developed with a user-friendly query interface. It provides an overview of the individual-level genotype-phenotype correlation with comprehensive annotation information. A quick search bar was set up with various types of searches as prompts in a prominent position on the home page of GPCards (Fig. 3). This quick search panel could automatically recognise a variety of key terms related to phenotype or phenotypic features information, including gene symbols, genomic regions, cytoband, genetic variants, gene transcripts, genomic coordinates, disease symbols, phenotype keywords, and identifiers of GPCards (GP_ID). In addition, GPCards provided an advanced search function, by which users can conveniently search for the catalogued genotype-phenotype correlation data in batches (<http://www.genemed.tech/gpcards/search>) (Fig. 3). Examples of different types of search items are presented in this panel. Another key feature of the advanced search of GPCards is that it allows users to assign annotation information presented in the search results, including pathogenicity information based on 24 predictive tools, population-specific allele frequencies, and data from established disease- and phenotype-related databases (Fig. 3). To avoid excessive data, users can select any of these data sources, as needed, in the advanced search panel. For example, users could select gnomAD datasets [15,16] only in the allele frequency section, which is considered the most comprehensive and ethnically diverse allele frequency database, as shown on the search results page.

3.3. Genotype-phenotype correlations in GPCards

The results of the quick search and advanced search are presented as tables that summarise the basic information for disease-associated genes, including the PubMed ID, gene symbol, disorder name, number of variants, patients, and phenotypes in each study (Fig. 4). Notably, the genotype-phenotype correlations of GPCards were specific to each original study and not aggregated across different studies which reported phenotypic features with different manners and vocabularies. When users click “phenotype summary and genotype-phenotype correlation”, a new and clear

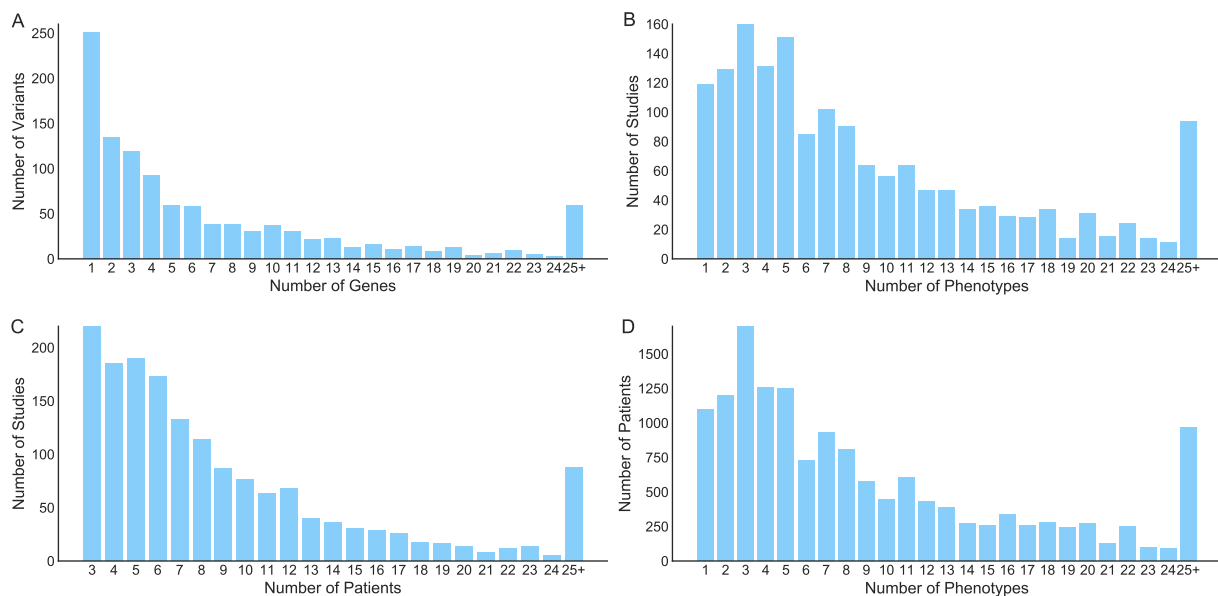


Fig. 2. Summary of catalogued genotype–phenotype correlation data. (A) The distribution of disease-associated genes with different number of genetic variants. (B) The distribution of studies with different number of clinical phenotypes. (C) The distribution of studies with different number of patients. (D) The distribution of patients with different number of clinical phenotypes.

interface is presented with two sections: phenotype summary and genotype–phenotype correlation. The phenotype summary section shows the frequencies of various clinical phenotypic features or symptoms of disease-causing genes in a given study (Fig. 4). For each clinical phenotypic feature, users would obtain the total number of patients examined, the number of patients that presented this phenotypic feature and the prevalence of this phenotypic features in the study. For example, by searching *JAG1*, users could conveniently learn that one study reported Alagille syndrome associated with *JAG1* in 70 patients. Furthermore, 17 patients (24.29%) show an interlobular bile duct paucity, 64 (91.43%) have a cardiac murmur, and 57 (81.43%) have characteristic facial features, in addition to other summarised phenotypes.

In the section on genotype–phenotype correlations, users can conveniently browse the detailed clinical phenotypic feature and genotype information as well as comprehensive annotations for genetic variants (Fig. 4). For the genotype information, users could obtain the genomic position, reference allele, alternative allele, Mendelian inheritance (recessive or dominant), origin of variants (*de novo* or inherited), variant type (homozygous or heterozygous), functional effects (stop-gain, frameshift, nonsynonymous, or splicing), and functional consequences predicted by several tools. For phenotype information, users could learn whether a patient with a specific genotype presents specific phenotypic features. For the annotation information, users could evaluate pathogenicity based on 24 predictive tools, allele frequencies in different populations, and whether the variant has been catalogued in other well-known disease- and phenotype-related databases. For example, a patient with Alagille syndrome carries a heterozygous *de novo* nonsynonymous variant (c.550C > T, p.R184C) in *JAG1* [64], with clinical phenotypic features of interlobular bile duct paucity, cholestasis, cardiac murmur, skeletal abnormalities, characteristic facial features, and posterior embryotoxon and without interlobular bile duct paucity and kidney abnormalities which phenotypic features may be presented in other patients with different genetic variants of *JAG1*. By clicking “Detailed Annotation”, users can learn that this variant was predicted to be deleterious or conserved by all 24 predictive tools, has not been reported in any population in gnomAD,

ExAC, and other population databases, is catalogued as likely pathogenic variant in InterVar, and is reported as pathogenic in the ClinVar database (Fig. 4). Notably, by clicking the *JAG1* gene symbol, users could also obtain comprehensive gene-level information (http://genemed.tech/gpcards/geneDetail/main?gene_symbol=JAG1), as mentioned in the Material and Methods section, similar to the Gene4denovo database (48) previously developed by our group.

3.4. Other functions in GPCards

GPCards support an analysis service that is freely available to all users on the Analysis page (<http://www.genemed.tech/gpcards/analysis>). Users are able to analyse genetic data by uploading the anonymized patient data files in VCF4 format and inputting their E-mail address. If users choose the “Trio” option for uploading a VCF file, they should select the sample IDs of the proband, unaffected father, and mother, and GPCards would automatically identify *de novo* mutations, homozygous variants, compound heterozygous variants, and the inherited hemizygous. If users choose the “Non-Trio” option and set the genotype (heterozygous, homozygous, or wild) of each sample, GPCards would automatically identify the co-segregated genetic variants that meet users’ requirements. With informed patient consent, GPCards would link the anonymized genetic variants to genotype–phenotype correlations. If GPCards identified a variant that has been catalogued, it would provide the detailed phenotypes of patients carrying the same variants. If GPCards prioritised a gene that has been catalogued, it would provide gene-level summary information for genotype–phenotype correlations. In addition, GPCards provides several parameters for quality control and detection of co-segregating rare damaging variants.

There are also some additional useful sections in GPCards. In the download section, users are allowed to freely download all of the genotype–phenotype correlation data compiled by about 20 professionals over several months (<http://genemed.tech/gpcards/download>). In the upload section, users could upload anonymized genotype–phenotype data, which would be helpful

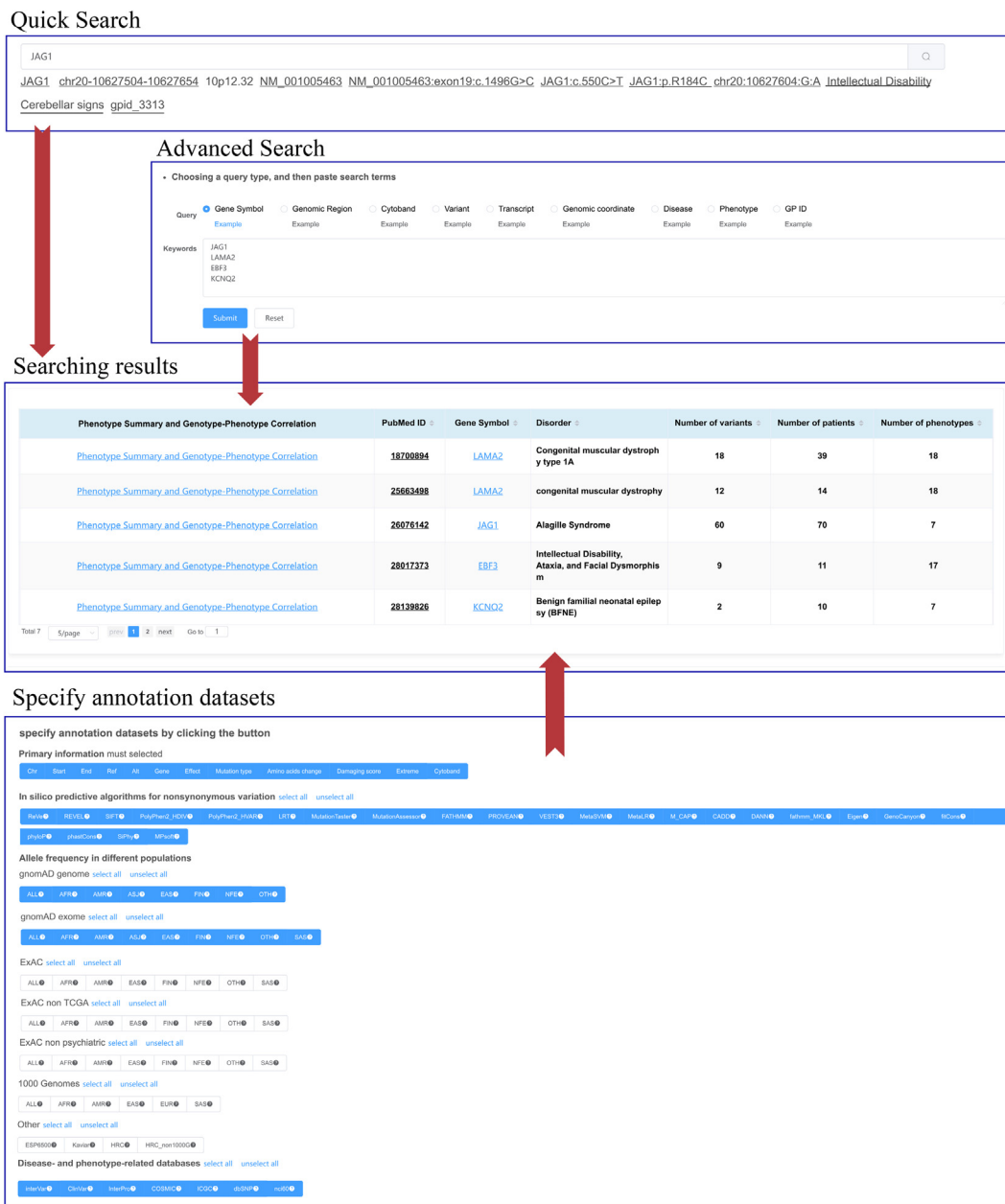


Fig. 3. Snapshot of search modules in GPCards. The quick search bar is set with 11 types of searches prompts as the example of *JAG1*. The advanced search could be used to conveniently search in batches with nine type of search prompts. The searching results would show PubMed ID, gene symbol, disorder name, number of variants, patients, and phenotypes. “Specify annotation datasets” is a selectable panel with 24 predictive tools, population-specific allele frequencies, and data from established disease- and phenotype-related databases, allowing users to assign annotation information presented in the panel of searching results.

for enriching the database (<http://genemed.tech/gpcards/upload>). After receiving the data uploaded by users, we will connect users to inform the consent, and perform de-identification before public release in GPCards database. Users could also access information for genotype–phenotype correlations by the browse function, which lists all catalogued genes and the total number variants, patients, and phenotypes in each study (<http://genemed.tech/gpcards/browse>). Moreover, in the browse section, users can efficiently access phenotypic data by choosing the first letter of the gene symbol. In the data source section, all integrated databases or algorithms are listed with summary information (<http://genemed.tech/gpcards/source>). We also supply an instruction manual on the Tutorial page, with detailed information about how to get started (<http://genemed.tech/gpcards/tutorial>).

4. Discussion

With the exponential growth of genetic data, especially in view of the extensive application of NGS technologies in the past decade, increasing disease-associated genetic variants have been discovered and implemented in diagnostic settings in medical genetics [1,2,5]. However, the overall diagnostic yield still lags behind the discovery of disease-associated genes [1,5]. Owing to the amount of data, it is increasingly difficult for clinical investigators and geneticists to extract relevant genotype–phenotype information from various literatures. To resolve this issue, we developed the GPCards database, which enables users to conveniently access information about genotype–phenotype correlations without requiring registration or payment. By using the GPCards database,

A: Searching results of genotype-phenotype correlations

Phenotype Summary and Genotype-Phenotype Correlation	PubMed ID	Gene Symbol	Disorder	Number of variants	Number of patients	Number of phenotypes
Phenotype Summary and Genotype-Phenotype Correlation	26076142	JAG1	Alagille Syndrome	69	79	7

B: Phenotype summary and genotype-phenotype correlation

Phenotype Summary and Genotype-Phenotype Correlation of JAG1 (26076142)

Phenotype Summary

PubMed ID	Gene Symbol	Disorder	Interlobular bile duct paucity (%)	Cholestasis (%)	Cardiac murmur (%)	Skeletal abnormalities (%)	Characteristic face (%)	Posterior embryotoxon (%)	Kidney abnormalities (%)
26076142	JAG1	Alagille Syndrome	17/70 (24.29%)	70/70 (100.00%)	64/70 (91.43%)	57/70 (81.43%)	57/70 (81.43%)	25/70 (35.71%)	9/70 (12.86%)

Genotype-Phenotype Correlation

Functional variants: All | Extreme: All


Detail Annotation	Functional variants	Amino acids change	Extreme	Chr	Start	End	Ref	Mut	Mendelian inheritance	De novo (Y/N)	Homozygous/heterozygous	Sample ID	Sex	Nucle
Detail Annotation	stopgain	JAG1_NM_000214...	Y	chr20	10622485	10622485	C	T	recessive	Y	Heterozygous	6	Female	c.2
Detail Annotation	splicing	NM_000214:exon3...	Y	chr20	10644609	10644609	A	AA	recessive	Y	Heterozygous	7	Female	c.43
Detail Annotation	nonsynonymous SNV	JAG1_NM_000214...	Y	chr20	10639260	10639260	G	A	recessive	Y	Heterozygous	8	Male	c.1
Detail Annotation	splicing	NM_000214:exon2...	Y	chr20	10623135	10623135	C	A	recessive	N	Heterozygous	9	Male	c.25
Detail Annotation	frameshift deletion	JAG1_NM_000214...	Y	chr20	10632796	10632805	CCTGAATACC	-	recessive	N	Heterozygous	10	Female	c.900

Next 66 | Skip page | prev 1 | 3 4 5 6 | 14 next | On 10 | 2

Functional variants: All | Extreme: All

De novo (Y/N)	Homozygous/heterozygous	Sample ID	Sex	Nucleotide change	AA Alteration	Interlobular bile duct paucity	Cholestasis	Cardiac murmur	Skeletal abnormalities	Characteristic face	Posterior embryotoxon	Kidney abnormalities
Y	Heterozygous	6	Female	c.2629D>A	p.W67E	+	+	+	+	+	+	-
Y	Heterozygous	7	Female	c.439>delupT	-	+	+	+	+	+	+	-
Y	Heterozygous	8	Male	c.550C>T	p.R184C	-	+	+	+	+	+	-
N	Heterozygous	9	Male	c.2572+10>T	-	+	+	+	+	+	+	-
N	Heterozygous	10	Female	c.900_989delG...	p.G327DfX32	NA	+	+	+	+	+	-

Next 66 | Skip page | prev 1 | 3 4 5 6 | 14 next | On 10 | 2



C: Variants-level implication

Detail Annotation	Functional variants	Amino acids change	Extreme	Chr	Start	End	Ref	Mut	Mendelian inheritance	De novo (Y/N)	Homozygous/heterozygous	Sample ID	Sex	Nucle
Detail Annotation	stopgain	JAG1_NM_000214...	Y	chr20	10622485	10622485	C	T	recessive	Y	Heterozygous	6	Female	c.2
Detail Annotation	splicing	NM_000214:exon3...	Y	chr20	10644609	10644609	A	AA	recessive	Y	Heterozygous	7	Female	c.43
Detail Annotation	nonsynonymous SNV	JAG1_NM_000214...	Y	chr20	10639260	10639260	G	A	recessive	Y	Heterozygous	8	Male	c.1

In silico missense prediction

Algorithm	Score	Prediction
ReVe	0.996	Damaging
SIFT	0.0	Damaging
LRT	0.000	Deleterious
MutationTaster	1	Disease causing automatic
MutationAssessor	3.495	Damaging
FATHMM	-4.21	Damaging
PROVEAN	-7.87	Damaging
VEST3	0.97	Damaging
MetaSVM	1.097	Damaging
MetaLR	0.960	Damaging
M-CAP	0.620	Damaging
CADD	35	Damaging
DANN	0.999	Damaging
Eigen	0.913	Damaging
GenoCanyon	1.000	Damaging
FitCons	0.719	Damaging
GERP++	5.43	Conserved
phyloP	8.950	Conserved
phastCons	1.000	Conserved
SiPhy	19.231	Conserved
REVEL	0.976	Damaging

Allele frequency in population

Dataset	Population	Allele frequency
gnomAD_exome	ALL	-
gnomAD_exome	African American	-
gnomAD_exome	Latino	-
gnomAD_exome	Ashkenazi Jewish	-
gnomAD_exome	East Asian	-
gnomAD_exome	Finnish	-
gnomAD_exome	Non-Finnish European	-
gnomAD_exome	Other	-
gnomAD_exome	South Asian	-
gnomAD_genome	ALL	-
gnomAD_genome	African American	-
gnomAD_genome	Latino	-
gnomAD_genome	Ashkenazi Jewish	-
gnomAD_genome	East Asian	-
gnomAD_genome	Finnish	-
gnomAD_genome	Non-Finnish European	-
gnomAD_genome	Other	-

Disease-related information

Database	Information
InterVar	Likely pathogenic
COSMIC70	-
ICGC	ID= OCCURRENCE=
ncBI	-
dbSNP	rs121918350
ClinVar	CLNDEN: Alagille_syndrome_1 CLNACC: RCV00000858.3 CLNDSDB: MedGen:OMM CLNDSDBID: C1956125:118450 Clinical significance: Pathogenic
InterPro	Delta/Serratelag-2 (DSL) protein

Fig. 4. Snapshot of genotype-phenotype correlations in GPCards. In “Phenotype Summary and Genotype-Phenotype Correlation” panel, the basic information of the searched genes was presented. The frequencies of various clinical phenotypes or symptoms of disease-causing genes is exhibited in the “Phenotype Summary” panel. The detailed individual-level phenotypes and genotypes were present in “Genotype-Phenotype Correlation” panel. Moreover, comprehensive variant-level annotations of each genetic variant were also present in this panel.

clinicians could classify complex diseases and syndromes into “molecular subtypes”, which would improve diagnostic accuracy and therapeutic efficacy. Clinicians could also conveniently identify genes or variants related to a phenotype using the disease name or phenotypic feature as a search term. This database is expected to substantially improve the application of genetic data to clinical diagnosis and treatment.

It is a complex, laborious, and expensive task to archive genotype–phenotype data from a large number of published studies to construct a useful database [4,65]. Owing to the substantial input of expertise, resources, and time, the newly developed GPCards database is practical and highly integrative. This database includes patients from a wide range of ethnic groups and geographical locations worldwide. All data were screened by professionals following a strict quality control system. Furthermore, we annotated all variants and genes using 62 well-established genetic or clinical data sources, providing a convenient one-stop database for the interpretation of pathogenicity of genetic variants. A quick search model and advanced search model provide easy operation interfaces with simple and easy-to-understand tips for users with a wide range of expertise, from beginners to scientists. GPCards is the first freely available database combining detailed individual-level information for genotype–phenotype correlations in human genetic diseases. Users can effectively simplify genotype–phenotype correlation data by utilising different functions of GPCards with personal needs, such as quick search, advanced search, browse, analysis, download and upload.

GPCards is a practical and highly integrative database aimed at aiding geneticists and clinicians. It can be used to prioritise novel candidate genes, for example. Different categories of human diseases may share extensive phenotypic features and therefore may be caused by mutations in the same genes, such as *de novo* mutations (DNMs) in *SCN2A* were reported to be associated different neuropsychiatric disorders we previously reported [66]. Therefore, a single gene may be associated with two correlated diseases. If the phenotype information indicated this gene is associated to a given disease, we can infer that this gene may be associated with another disease which share the similar clinical features, based on the genotype–phenotype association. For example, previous studies demonstrated that DNMs in *CHD8* were associated with autism spectrum disorder [67], a disease usually accompanied with intellectual disability, suggesting that *CHD8* is a candidate gene for intellectual disability.

In the past decades, many disease-related databases have been developed, such as OMIM [10], CentoMD [4], HGMD [68], HPO [11], ClinVar [12], DECIPHER [69], and MalaCards [13], as well as PhenoTips [70], Phenopolis [71], RD-Connect [72] and Patient Archive [73]. Both OMIM and HPO were database of describing human genes and associated diseases/phenotypes without enough variant-level information. In addition, CentoMD, PhenoTips, and HGMD were all pay-per-use databases with genetic and clinical information from HPO and OMIM, users have to pay for the query services. Meanwhile, ClinVar database was well known for the variant-level information and associated disease, but lacks the detailed individual-level phenotypic information, as well as other listed databases above. DECIPHER was used by the clinical community to share and compare phenotypic and genotypic data. MalaCards listed the known aliases, as well as inter-disease connections, consolidated from 74 sources. There are also some workflows, which can be adapted to any set of patients for which phenomic and genomic data is available, such as PhenCo {Diaz-Santiago, 2020 #97}, were reported recently. Furthermore, GWASkb [74], GWAS Central [75], GWAS Catalogue [76,77], PhenoScanner [78] and GRASP [79] focused on the relationship between different human traits and common SNPs instead of pathogenic variants. Compared to these databases, GPCards was

an open accessed database which integrated peer-reviewed patient-level genotype–phenotype associations of genetic diseases and provided one-stop service for researchers and clinicals to interpret the pathogenicity of genetic variants.

Furthermore, most of the existing genotype and phenotype databases do not supply analysis service, especially the free analysis function. However, GPCards features a free analysis service that allows users to easily complete a preliminary analysis of genotype data, annotated and prioritised genes with valuable information in gene level associated with phenotypes. This free analysis service will be groundbreaking in providing convenience to users and advancing the development of genotype and phenotype data analysis. Meanwhile, the download section and upload section are other highlights of GPCards. Based on the concept of maximum openness, users can upload anonymous genotype–phenotype information, which is necessary for patients’ data protection, and can also download the data collected by GPCards for re-analysis and re-mining. Thus, GPCards provides a platform for researchers to jointly promote the development of genotypic and phenotypic correlation research. GPCards will provide more accurate and comprehensive information regarding to genotype–phenotype correlations in more patients with the development of medical genetics. The current version of GPCards is the beginning and attempt to decipher the genotype and phenotype correlation and will be widely concerned by researchers and clinicians. However, there are some limitations in the present study. First, a large number of genotype–phenotype correlation data have been reported in thousands of literatures, but the format and standards of these information were differed widely. We try our best to search genetics studies of each human gene and found that the clinical phenotypic features were not available for most studies, leading to some genetic variants were missed in GPCards. We suggested that phenotypic data and corresponding variant data should be recorded in as much detail as possible in future publications. Meanwhile, although the continuous updating of the database is costly in terms of both in terms of money and time, we firmly believe in the potential utility of the database, so we will keep to update it semi-annually. Furthermore, we also encourage users to upload anonymized genotype–phenotype data to GPCards. Second, genotype–phenotype correlations in GPCards could be used to assist in diagnosis but not as diagnostic criteria, due to the following three points: (i) pathogenic variants may later be identified as non-pathogenic, as previous reported [80]; (ii) some of the reported pathogenic have not been functionally validated in cell and animal experiments; (iii) many genes and variants may present incomplete penetrance.

In conclusion, GPCards offers extensive information about patient-level genotype–phenotype correlations in a user-friendly open-access interface, without requiring registration. GPCards also provides comprehensive gene- and variant-level annotations to facilitate the interpretation of the pathogenesis of genetic variants. We expect GPCards to be helpful for the prioritisation of novel candidate genes, genetic counselling and diagnosis, and development of appropriate treatment strategies.

5. Availability of data and materials

The datasets analyzed during the current study are available in the Gene4PD repository (<http://genemed.tech/gene4pd/>).

6. Contribution

BL, GZ, and JL were involved in study conception and design; WZ, QC, KL, XW, YW, QZ, YH, Blu, YZ, RZ, LJ, HP, TL, YZ, ZF, XX, XZ, RW, LZ, YW, ZY, LX, JG, BT, KX collected the data. GZ and QZ build this online platform. BL, GZ and JL wrote the manuscript.

All authors contributed to the preparation of the manuscript and read and approved the final manuscript.

Funding

This work was supported by The National Natural Science Foundation of China (81801133 to JCL, 82001362 to BL), Young Elite Scientist Sponsorship Program by CAST (2018QNRC001 to JCL), Innovation-Driven Project of Central South University (20180033040004 to JCL), and Natural Science Foundation for Young Scientists of Hunan Province, China (2019JJ50974 to GHZ), Hunan Natural Science Foundation Outstanding Youth Fund (2020JJ3059 to JCL), Changsha Municipal Natural Science Foundation (kq2014278 to BL).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

References

- [1] Liu Z, Zhu L, Roberts R, Tong W. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we? *Trends Genet* 2019;35(11):852–67.
- [2] Levy SE, Myers RM. Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* 2016;17(1):95–115.
- [3] Fernandez-Marmiesse A, Gouveia S, Couce ML. NGS technologies as a turning point in rare disease research, diagnosis and treatment. *Curr Med Chem* 2018;25(3):404–32.
- [4] Trujillano D, Oprea G-E, Schmitz Y, Bertoli-Avella AM, Abou Jamra R, Rolfs A. A comprehensive global genotype-phenotype database for rare diseases. *Mol Genet Genomic Med* 2017;5(1):66–75.
- [5] Di Resta C, Galbiati S, Carrera P, Ferrari M. Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC* 2018;29(1):4–14.
- [6] Nussinov R, Tsai C-J, Jang H, Bravo I. Protein ensembles link genotype to phenotype. *PLoS Comput Biol* 2019;15(6):e1006648.
- [7] Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. *npj Syst Biol Appl* 2019;5:15.
- [8] Dwivedi S, Purohit P, Misra R, Pareek P, Goel A, Khattry S, et al. Diseases and molecular diagnostics: a step closer to precision medicine. *Indian J Clin Biochem* 2017;32(4):374–98.
- [9] Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum Mol Genet* 2013;22(R1):R27–31.
- [10] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;47(D1):D1038–43.
- [11] Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gouridine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019;47(D1):D1018–27.
- [12] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res* 2020;48(D1):D835–44.
- [13] Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2017;45(D1):D877–87.
- [14] Li J, Shi L, Zhang K, Zhang Y, Hu S, Zhao T, et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res* 2018;46(D1):D1039–48.
- [15] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285–91.
- [16] Scheps KG, Hasenauer MA, Parisi G, Targovnik HM, Fornasari MS. Curating the gnomAD database: Report of novel variants in the globin-coding genes and bioinformatics analysis. *Hum Mutat* 2020;41(1):81–102.
- [17] Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 2017;45(D1):D840–5.
- [18] Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013;493(7431):216–20.
- [19] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* 2015;526(7571):68–74.
- [20] Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 2011;27(22):3216–7.
- [21] Li J, Zhao T, Zhang Yi, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 2018;46(15):7793–804.
- [22] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;99(4):877–85.
- [23] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31(13):3812–4.
- [24] Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc* 2016;11(1):1–9.
- [25] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7(4):248–9.
- [26] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009;19(9):1553–61.
- [27] Schwarz JM, Rödelberger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;7(8):575–6.
- [28] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;39(17):e118.
- [29] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;34(1):57–65.
- [30] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP, de Brevern AG. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 2012;7(10):e46688.
- [31] Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;14(Suppl 3):S3.
- [32] Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;24(8):2125–37.
- [33] Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;48(12):1581–6.
- [34] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46(3):310–5.
- [35] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;31(5):761–3.
- [36] Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;31(10):1536–43.
- [37] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;48(2):214–20.
- [38] Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 2015;47(3):276–83.
- [39] Noyce AJ, Bestwick JP, Silveira-Moriyama L, Hawkes CH, Giovannoni G, Lees AJ, et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Ann Neurol* 2012;72(6):893–901.
- [40] Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. *Lect Notes Comput Sci* 2006;3909:190–205.
- [41] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;15(8):1034–50.
- [42] Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009;25(12):i54–62.
- [43] Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;100(2):267–80.
- [44] Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45(D1):D777–83.
- [45] International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. *Nature*, 2010;464(7291):993–8.
- [46] Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res* 2017;45(D1):D190–9.
- [47] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29(1):308–11.
- [48] Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, et al. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res* 2020;48(D1):D913–26.

- [49] Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 2015;43(Database issue): D36–42.
- [50] The Gene Ontology C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 2017;45(D1):D331–D8.
- [51] Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res* 2017;45(D1): D865–76.
- [52] Petrovski S, Gussow AB, Wang Q, Halvorsen M, Han Y, Weir WH, et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet* 2015;11(9):e1005492.
- [53] Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* 2017;33(4):471–4.
- [54] Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* 2016;48(4):349–55.
- [55] Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A* 2015;112(44):13615–20.
- [56] Han X, Chen S, Flynn E, Wu S, Wintner D, Shen Y. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat Commun* 2018;9(1):2138.
- [57] UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46(5):2699.
- [58] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, et al. The NCBI BioSystems database. *Nucleic Acids Res* 2010;38(suppl_1):D492–6.
- [59] Eppig JT, Smith CL, Blake JA, Ringwald M, Kadin JA, Richardson JE, et al. Mouse Genome Informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Methods Mol Biol* 2017;1488:47–73.
- [60] Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. *Nature* 2014;508(7495):199–206.
- [61] Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* 2015;13(5):307–8.
- [62] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. *Science* 2015;347(6220):1260419.
- [63] Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res* 2018;46(D1):D1068–73.
- [64] Li L, Dong J, Wang X, Guo H, Wang H, Zhao J, et al. JAG1 mutation spectrum and origin in Chinese children with clinical features of Alagille syndrome. *PLoS ONE* 2015;10(6):e0130355.
- [65] Cotton RGH, Phillips K, Horaitis O. A survey of locus-specific database curation. *Human Genome Variation Society. J Med Genet* 2007;44(4): e72.
- [66] Li J, Cai T, Jiang Yi, Chen H, He X, Chen C, et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* 2016;21(2):290–7.
- [67] Bernier R, Golzio C, Xiong Bo, Stessman H, Coe B, Penn O, et al. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 2014;158(2):263–76.
- [68] Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 2014;133(1):1–9.
- [69] Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* 2014;42(D1):D993–D1000.
- [70] Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat* 2013;34(8):1057–65.
- [71] Pontikos N, Yu J, Moghul I, Withington L, Blanco-Kelly F, Vulliamy T, et al. Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics* 2017;33(15):2421–3.
- [72] Gainotti S, Torrerri P, Wang CM, Reihls R, Mueller H, Heslop E, et al. The RD-Connect Registry & Biobank Finder: a tool for sharing aggregated data and metadata among rare disease researchers. *Eur J Hum Genet* 2018;26(5):631–43.
- [73] McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the phenotype frontier: the Monarch initiative. *Genetics* 2016;203(4):1491–5.
- [74] Kuleshov V, Ding J, Vo C, Hancock B, Ratner A, Li Y, et al. A machine-compiled database of genome-wide association studies. *Nat Commun* 2019;10(1):3341.
- [75] Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet* 2014;22(7):949–52.
- [76] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014;42(Database issue):D1001–6.
- [77] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;47(D1): D1005–12.
- [78] Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* 2019;35(22):4851–3.
- [79] Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* 2014;30(12):i185–94.
- [80] van Rooij J, Arp P, Broer L, Verlouw J, van Rooij F, Kraaij R, et al. Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time. *Genet Med* 2020.