Briefings in Bioinformatics, 00(00), 2018, 1–18

doi: 10.1093/bib/bby071
Advance Access Publication Date: 8 August 2018
Review article

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Computational resources associating diseases with genotypes, phenotypes and exposures

Wenliang Zhang, Haiyue Zhang, Huan Yang, Miaoxin Li, Zhi Xie and Weizhong Li

Corresponding authors: Weizhong Li, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China. Tel.: +86 20 85295864; E-mail: liweizhong@mail.sysu.edu.cn

Abstract

The causes of a disease and its therapies are not only related to genotypes, but also associated with other factors, including phenotypes, environmental exposures, drugs and chemical molecules. Distinguishing disease-related factors from many neutral factors is critical as well as difficult. Over the past two decades, bioinformaticians have developed many computational resources to integrate the omics data and discover associations among these factors. However, researchers and clinicians are experiencing difficulties in choosing appropriate resources from hundreds of relevant databases and software tools. Here, in order to assist the researchers and clinicians, we systematically review the public computational resources of human diseases related to genotypes, phenotypes, environment factors, drugs and chemical exposures. We briefly describe the development history of these computational resources, followed by the details of the relevant databases and software tools. We finally conclude with a discussion of current challenges and future opportunities as well as prospects on this topic.

Key words: disease phenotype; genotype; environmental exposure; database; software tool; web platform

Introduction

As the advance of sequencing and other high-throughput technologies are producing big omics data for medical research, how to utilize and analyze these data to understand human diseases has become increasingly challenging. Whole exome sequencing or whole genome sequencing could unravel hundreds of thousands to even millions of variants, of which only a few may

Wenliang Zhang is a PhD student in Zhongshan School of Medicine at Sun Yat-sen University. His research is focused on the interpretation of human genotypes and phenotypes.
Haiyue Zhang is a research assistant in Zhongshan School of Medicine at Sun Yat-sen University. Her research is focused on building and applying ontologies for phenotypes and diseases.
Huan Yang is a PhD student in Zhongshan School of Medicine at Sun Yat-sen University. Her research is focused on the integrative analysis of multi-omics data.
Miaoxin Li (PhD) is a professor of Bioinformatics in Zhongshan School of Medicine at Sun Yat-sen University. He is interested in discovering novel genomic variations associated with human diseases.
Zhi Xie (MD, PhD) is a professor of Bioinformatics in Zhongshan Ophthalmic Center at Sun Yat-sen University. He is interested in understanding transcriptional and translational regulation through the integrative analysis of multi-omics data.
Weizhong Li (PhD) is a professor of Bioinformatics in Zhongshan School of Medicine at Sun Yat-sen University. He is interested in understanding and interpreting the relationships between genomic factors and disease phenotypes through computational approaches.
Submitted: 31 May 2018; **Received (in revised form):** 1 July 2018

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

be disease-causative or related [1–4], thus identifying disease-causing genes and pathogenic variants is critical in human genetics studies. The focus of the genetics field is shifted from the production of genotypic data to the annotation and interpretation of analysis results.

The causes of a disease and its therapies are not only related to genotypes but also associated with other factors, such as phenotypes, environmental exposures, drugs and chemical molecules, etc. Distinguishing disease-related factors from many neutral factors is critical as well as difficult. Misleading assignment of pathogenicity to factors may result in inaccurate disease-risk assessments and diagnoses along with unsuitable treatments. Individual phenotype, broadly defined as any observable characteristics of an individual [5], arises from complex interactions between the above multiple factors. Correct and accurate interpretation of the relationships between these factors is fundamentally important for the investigation of human disease mechanisms.

Over the past two decades, bioinformaticians have developed more than 100 computational resources to integrate the omics data and discover associations among genotypes, phenotypes, environmental exposures, drugs and chemical molecules. These computational resources, including databases such as Online Mendelian Inheritance in Man (OMIM) [6], ClinVar [7] and dbGAP [8–10], software tools such as Polyphen [11], ANNOVAR [12], Eigen [13], DeepSea [14] and PhenIX [15] and web platforms such as Open Targets [16] and DisGeNet [17], offer online and standalone applications to prioritize genotype–phenotype associations (GPAs), phenotype-drug/chemical-target associations and other associations. Undoubtedly, these computational resources have facilitated the research in life sciences and greatly supported the development of precision clinical medicine. However, researchers and clinicians are experiencing difficulties in choosing appropriate resources from hundreds of relevant databases and software tools. Therefore, it is imperative to critically review the disease-related databases and tools, not only for life scientists, but also for medical researchers and clinicians.

Here we systematically review the public computational resources of human diseases related to genotypes, phenotypes, environment factors, drugs and chemical exposures. We begin with the history of development of computational resources for human diseases, followed by the description of the relevant databases and the comparison of their scales of data and scopes of usage. Then we summarize and compare the software tools and the web platforms for the deeper understanding of associations between multiple disease-related factors. Finally, we conclude with a discussion of current challenges and future opportunities as well as prospects on this topic.

Development of the computational resources

Disease-related data, including phenotypes, genotypes, environment factors and drug/chemical exposures, were mainly generated by a range of international projects or research programs and have been stored and integrated in different public computational resources, freely available to the public (Figure 1 and Supplementary S-Table 1).

OMIM is the first established database to provide a catalog of human genes and genetic disorders [6], followed by the starting of the Human Genome Project in 1990. Five years later, the Human Gene Mutation Database (HGMD) was published to handle the data of human gene mutations [18, 19], followed by the construction of dbSNP [20] and Orphanet [21] in the late 1990s to integrate data of single nucleotide polymorphisms (SNPs) and rare diseases based on protein-coding genes. Since the year of 2000, several organism models have been developed and the databases of these model species are available not only for life science studies but also for medical research, e.g. Mouse Genome Database (MGD) [22] and MouseNet [23], Rat Genome Database (RGD) [24] and Zebrafish Model Organism Database (ZFIN) [25]. In 2000s, the databases of drug targets and chemical molecules were established to accelerate the development of molecular drugs, such as PharmGKB [26], DrugBank [27] and PubChem [28]. Since the late 2000s, noncoding RNAs have been found important in the development of diseases [29–33], and thus databases have been constructed to classify relationships between noncoding RNAs and human diseases, for example, NONCODE [34], miR2Disease [35] and LncRNADisease [36]. At the same time, the international projects and research programs of population genomics, including 1000Genomes [37], TCGA [38, 39], ICGC [40] and UK10K [41], have produced biomedical big data for the communities of life and medical sciences to share, analyse and utilize. Environmental factors (EFs), drugs and chemicals also play critical roles in the development of diseases, such as the Comparative Toxicogenomics Database (CTD) [42], LncREnvironmentDB [43] and Exposome-Explorer [44].

With the rapid growth of data in these databases, data mining and analysis have become another challenge. Since 2003, at least 30 tools have been developed to annotate, predict and prioritize functional effects of genomic variants, as well as to identify genomic variants of uncertain significance (Figure 1 and Supplementary S-Table 1), e.g. SIFT [45], PolyPhen [11], ANNOVAR [12], VASST [46] and GWAVA [47]. Additionally, several ontology-driven computational tools have been developed to facilitate clinical interpretation of genomic variants based on functional prediction of genomic variants and deep phenotype annotations, such as PhenIX [15] and Phevor [48]. Moreover, machine learning technologies (including deep learning) have recently been implemented to predict variations and their biological effects, for example, CADD [49], Eigen [13], DeepSea [14] and DeepVariant [50]. Furthermore, several web platforms, such as DisGeNET [17], MalaCards [51], Monarch Initiative [52] and Open Targets Platform [16], have been established to comprehensively integrate a variety of disease-related data sources with computational tools, allowing easy and simultaneous data access and analysis.

Databases

Dozens of public databases have been developed to store, retrieve and manage disease-related data. According to scopes and data associations, the databases can be categorized into seven groups (Table 1). The database group for coding genes includes data resources that primarily provide association information between protein-coding genes and phenotypes of human diseases, while the group for ncRNAs contains noncoding RNAs information associated with diseases. The group for genomic variations associates genomic variant information with phenotypes of disease. The group for population genomic data focuses on the worldwide clinical genomic variation and allele frequencies in various populations. The group of genetical organism models stores association information between genotypes and phenotypes/diseases of laboratory organisms. The group of environment exposures offers toxicogenomic relationships relevant to exposed factors, genes, proteins, phenotypes or diseases. The treatment group provides information that involves target drugs, drug resistance mutations, disease and their associations. All of these databases offer internet access of data through web browsers, and some of them also

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

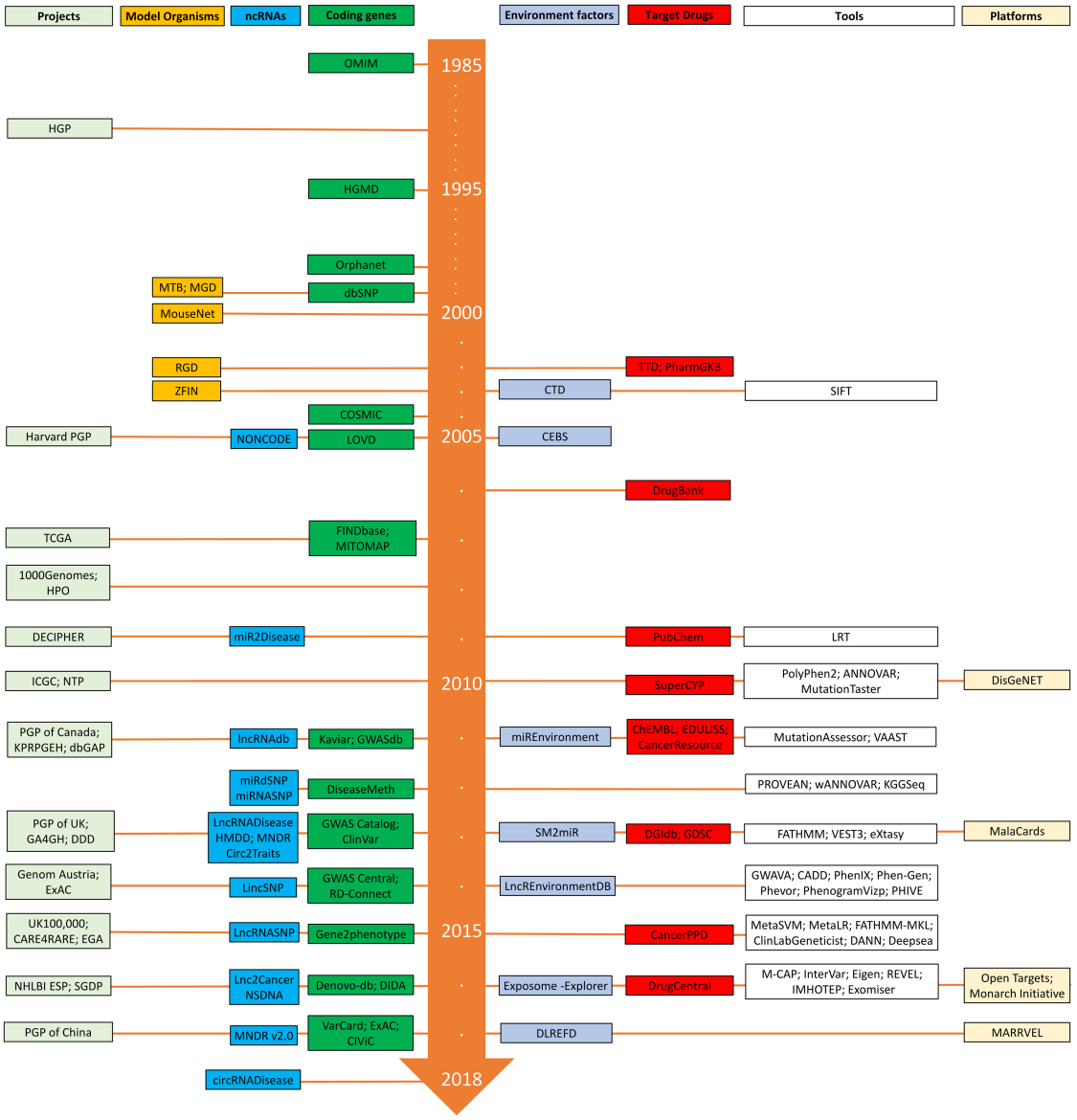

Figure 1. Development history of disease-related computational resources. The development of disease-related databases, software tools and web platforms, is depicted over the timeline. According to scopes and applications, the computational resources are classified into different groups.

offer web Application Programming Interfaces (APIs). Table 1 summarizes the groups according to their scopes and associations. Table 2 states the current status of the databases and Supplementary S-Table 2 states the data standards of nomenclature. The URLs of the databases can also be found in the supplementary file.

Coding genes

Approximately 50 databases provide disease-related phenotype information associated with genotypes. Several of them focus on depicting the association between protein coding gene and phenotypes (Table 1). One of the most widely used databases is OMIM, which is manually collated and integrated from numerous peer-reviewed literature and other medical information, offering broad and powerful compilations of knowledge about human genes, genetic phenotypes and the relationships between them [6]. The latest OMIM database contains 15 919 gene descriptions, 8670 phenotypes and 3928 genes with association to 1 or more phenotype(s) [6] (Table 2). Another similar example is Orphanet [53]. Instead of targeting on Mendelian disorders, Orphanet focuses on easy access to accurate and specific recommendations for the management of rare diseases. It establishes the relationships between classification of rare diseases, textual data and the appropriate services for patients and healthcare professionals.

It has been debated that many diseases classically considered monogenic may be better described as more complex inheritance, such as oligogenic mechanisms [101]. Gazzo *et al.* published the DIDA database as a *Nucleic Acids Research* breakthrough article in 2016 to offer the first-time detailed information on genes and associated genetic variants involved in digenic disorders, the simplest form of oligogenic inheritance [55]. The current DIDA database includes 213 digenic combination-disease associations involved in 44 digenic diseases (Table 2).

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 1. Comparison of different disease-related data resources

Data resource name	Phenotype/Disease			Genotype			Environmental factors	Drugs/ chemicals	Association	
	Mendelian and Rare	Complex and Trait	Organism model of disorder	Coding	Non-coding	Function annotation of variant			Types	Score
Coding genes										
OMIM	√(M)	√(F)	√	√(M)	√(F)				GPAs	
Orphanet	√			√(M)	√(F)			√	GPAs, PDAs	
DIDA	√			√(digenic)					GPAs	
DiseaseMeth		√(Cancer)		√					GPAs	
Noncoding RNAs										
miR2Disease		√			√(miR)				GPAs	
HMDD v2.0		√			√(miR)				GPAs	
NONCODE		√			√(lnc)				GPAs	
LncRNADisease		√			√(lnc)				GPAs	
Lnc2Cancer		√(Cancer)			√(lnc)				GPAs	
NSDNA		√(NSDs)			√(ncR)				GPAs	
circRNADisease		√			√(circ)				GPAs	
MNDR	√(F)	√(M)			√(ncR)				GPAs	√
Genomic variants										
HGMD	√	√		√(M)	√(F)	√			GPAs	√
ClinVar	√	√		√(M)	√(F)	√			GPAs	√
VarCards	√	√		√(M)	√(F)	√			GPAs	√
GWAS Catalog		√		√(M)	√(F)	√			GPAs	√
GWAS Central		√		√(M)	√(F)	√			GPAs	√
GWASdb		√		√(M)	√(F)	√		√	GPAs, GDAs	√
COSMIC		√(Cancer)		√(M)	√(F)	√		√	GPAs, GDAs	√
CIViC		√(Cancer)		√		√		√	GPAs, GDAs	√
Denovo-db		√(NSDs)		√(M)	√(F)	√			GPAs	√
miRdSNP		√			√(miR)	√			GPAs	
LincSNP		√			√(lnc)	√			GPAs	√
LncRNASNP		√			√(lnc)	√			GPAs	√
Population genomic data										
dbSNP				√	√	√				√
ESP	√	√		√		√			GPAs	
ExAC				√		√				
1000Genome				√	√	√				
Kaviar				√	√	√				
FINDbase				√	√	√				
Genetical organism models										
MGD	√	√	√(Mouse)	√					GPAs	
MTB		√(Cancer)	√(Mouse)	√					GPAs	
RGD	√	√	√(Rat)	√					GPAs	
ZFIN	√	√	√(zebra fish)	√				√	GPAs	
Environmental exposures										
CTD	√	√		√			√		GPAs, GEFAs, PEFAs	
miREnvironment		√			√(miR)		√		GPEFAs	
SM2miR		√			√(miR)		√	√	GEFAs	
LncEnvironmentDB					√(lnc)		√		GEFAs	√
DLREFD		√			√(lnc)		√	√	GPEFAs	

Continued

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 1. (continued)

Data resource name	Phenotype/Disease			Genotype			Environmental factors	Drugs/ chemicals	Association	
	Mendelian and Rare	Complex and Trait	Organism model of disorder	Coding	Non-coding	Function annotation of variant			Types	Score
Treatments (drugs and their targets)										
ChEMBL				√(Target)		√(F)		√	PDTAs	
DrugBank	√	√		√(Target)		√(F)		√	PDTAs	
DrugCentral	√	√		√(Target)				√	PDTAs	
TTD	√	√		√(Target)		√		√	PDTAs	
PharmGKB	√	√		√(Target)		√		√	GDAs	√
DGIdb				√(Target)				√	DTAs	√
CancerPPD		√(Cancer)		√(Target)				√(Peptides)	DTAs	

According to scopes and data associations, the databases can be categorised into major groups, but some of them could be included in multiple groups. The symbol '√' indicates the relevant information provided in each database. The following are the name abbreviations: NSDs: nervous system diseases; M: majority; F: few; lnc: lncRNA; mi: miRNA; circ: circRNA; ncR: ncRNAs, including lncRNA, miRNA, piRNA, siRNA and snoRNA etc.; GPAs: genotype–phenotype associations; GDAs: genotype-drug associations; PDAs: phenotype-drug associations; GPEFAs: genotype–phenotype-EF associations; GEFAs: genotype-environmental gactor associations; PDTAs: phenotype-drug-target associations; DTAs: drug-target associations.

The publication of DIDA may initiate further data annotation and tool development for deciphering more complex inheritance, such as polygenic disorders.

Complex diseases generally involve multiple levels of alterations, such as epigenetics and transcriptomic alterations [102, 103]. The human disease methylation database (DiseaseMeth), first published in 2012 [104], associates aberrant DNA methylation with human diseases, especially various cancers. Data in DiseaseMeth are manually or computationally extracted from experimental studies and high-throughput methylome data. The current DiseaseMeth [56] database contains over 679 000 aberrant DNA methylation-disease associations across 88 diseases (Table 2). To identify correlations between DNA methylation and RNA expression, another methylation-related database, called MethHC, provides a large collection of DNA methylation data combined with mRNA/microRNA expression profiles in human cancer [105]. These resources provide coding gene-disease associations that are a great utility in different research and clinical purposes, including the investigation of causes of specific human diseases and the interpretation of clinical significance of genetic dysfunctions in coding genes. Researchers are recommended to use OMIM for studies in Mendelian inheritance, Ophanet for rare disorders, DIDA for digenic disorders and DiseaseMeth for disease-related methylation.

Noncoding RNAs

A large portion of human genome is transcribed into noncoding RNAs (ncRNAs), particularly long-noncoding RNAs (lncRNAs), micro RNAs (miRNAs) and circular RNA (circRNA), potentially representing another layer of epigenetic regulation [33, 106]. Accumulative investigations have shown that ncRNAs play critical roles in many important biological processes [32] and its deregulations could be related to a broad spectrum of diseases [29–33]. Evidently, ncRNAs have become a novel class of potential biomarkers and targets for disease diagnosis, therapy and prognosis. Due to their functional and clinical significance, several databases have been established since 2005, including miRbase [107] for miRNAs, NONCODE [57], LNCipedia [108] and lncRNAdb [109] for lncRNAs. These databases connect ncRNA to diseases and also integrate annotation data of sequences, functions, expressions, related targets and cellular locations. For example, the latest NONCODE [57] has annotated 167 150 human lncRNA sequences, of which 1110 are associated with 284 diseases [36] (Table 2).

Several databases target on the association between ncRNA dysregulation and human diseases (Table 1 and Table 2). For example, miR2Disease [35] and Human MicroRNA Disease Database (HMDD) [58] provide miRNA dysregulation-human disease associations and miRNA-target associations. The current release of HMDD has integrated 10 368 associations between 572 miRNAs and 378 diseases. Similarly, LncRNADisease [36] and Lnc2Cancer [59] contain manually curated entries of experimentally supported lncRNA-disease associations and lncRNA-target associations, and the latter focuses on association data for cancer research. Unlike LncRNADisease and Lnc2Cancer, the Nervous System Disease NcRNAome Atlas (NSDNA) [60] aims to offer a comprehensive, quality and special resource of NSD-related ncRNA dysregulation. It manually collects experimentally supported associations between nervous system diseases (NSDs) and different types of ncRNAs, including miRNAs, lncRNAs, piRNAs, siRNAs and snoRNAs. The latest [60] NSDNA contained 26 128 associations between 8736 ncRNAs and 144 NSDs (Table 2). The MNDR database [110] integrates experimentally supported and predicted ncRNA-disease associations from 14 resources such as HMDD [58], Lnc2Cancer [59], NSDNA and LncDisease [111].

Moreover, several databases store predicted circRNA-disease associations such as Circ2Traits [112] and manually curated circRNA-disease associations from peer review papers such as circRNADisease [61]. Currently, circRNADisease provides 354 curated associations between 330 circRNAs and 48 diseases including cancers, neurodegeneration and cerebrovascular diseases [61]. Each association has comprehensive annotation information such as circRNA name, expression pattern, associated partners, associated diseases, experimental detection techniques and publication reference.

The above resources of ncRNA-disease relationships can be used conjunctively to discover and predict associations between novel ncRNAs and diseases, and to facilitate the interpretation of clinical significance of dysfunctions in ncRNAs. Lnc2Cancer is preferable for studying cancer-related lncRNAs, and NSDNA for NSD-related lncRNAs.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 2. Summary of disease-related databases

Database	Scope and scale	Date of statistic
Coding genes		
OMIM [6]	15 919 gene descriptions, 8670 phenotypes and 3928 genes with association to 1 or more phenotype(s)	22 June 2018[w]
Orphanet [53]	6949 associations between genes and rare diseases	Aug 2016[w]
Gene2phenotype [54]	2285 GPAs in developmental disorders	Oct 2017[w]
DIDA [55]	213 digenic combination-disease associations in 44 digenic diseases	Oct 2015[p]
DiseaseMeth v2.0 [56]	679 602 aberrant DNA methylation-disease associations in 88 diseases, especially in various cancer	Nov 2016[p]
Noncoding RNAs		
NONCODE [57]	1110 lncRNAs associated with 284 diseases	Nov 2016[p]
miR2Disease [35]	3273 associations between 349 miRNAs and 169 diseases	Jun 2018[w]
HMDD v2.0 [58]	10 368 associations between 572 miRNAs and 378 diseases	Jun 2013[p]
LncRNADisease [36]	3000 association between 914 lncRNAs and 329 diseases	July 2017[w]
Lnc2Cancer [59]	1488 associations between 666 lncRNAs and 97 cancers	July 2016[w]
NSDNA [60]	26 128 associations between 8736 ncRNAs and 144 nervous system diseases	May 2017[w]
circRNADisease [61]	354 associations between 330 circRNAs and 48 diseases	Apr 2018[p]
MNDR v2.0 [62]	8824 lncRNA-disease, 70 381 miRNA-disease, 118 piRNA-disease and 67 snoRNA-disease experimental associations across 6 mammals	Nov 2017[p]
Genomic variants and population genomics		
Clinvar [7]	428 435 genomic variant-disease associations across 30 181 genes	Jun 2018[w]
HGMD [63]	224 642 disease related variants on 8784 genes	Jan 2018[w]
Denovo-db [64]	(July 2016)[p]: 32 991 de novo genetic variants in neurodevelopmental disorders	
VarCards [65]	110 154 363 artificially generated SNVs and 1 223 370-reported indels in coding region and splicing sites	Oct 2017[p]
LOVD 2.0 [66]	3 334 104 (2 400 084 unique) variants in 248 807 individuals in 86 LOVD installations	Dec 2015[p]
MITOMAP [67]	1746 variants on mitochondrial DNA	Dec 2015[p]
COSMIC [68]	208 368 associations between somatic mutations and cancer	Nov 2016[p]
CIViC [69]	1678 interpretations of clinical relevance for 713 variants affecting 283 genes associated with 209 cancer subtypes and 291 drugs	Feb 2017[p]
GWAS Catalog v2 [70]	~60 000 associations between SNPs and traits/diseases	Apr 2018[w]
GWASdb v2.0 [71]	252 530 associations between SNPs and traits/diseases	Nov 2015[p]
GWAS Central [72]	69 986 326 associations between 2 974 961 SNPs and 829 traits/diseases	Nov 2017[w]
LincSNP2.0 [73]	371 647 associations between lncRNA SNPs and diseases, and 1 266 485 Linkage disequilibrium (LD)-SNPs	Oct 2016[p]
LncRNASNP2 [74]	697 lncRNA-Disease associations; 602 GWAS-SNPs and 2 859 147 SNPs in LD regions	Oct 217[p]
miRdSNP [75]	786 associations between 630 unique disease-associated SNPs and 204 disease types	2012[p]
miRNASNP [76]	2257 SNPs in 1596 human pre-miRNAs;706 SNPs in miRNA mature regions and 227 SNPs in miRNA seed regions	Jan 2015[p]
dbSNP [77]	A genomic variation database including 660 773 127 SNPs of *Homo sapiens*.	Mar 2018[w]
ExAC [78]	Variations from 130 000 subject exome sequencing data from a wide variety of large-scale sequencing projects	Aug 2016[p]
ESP [79]	1 788 563 variants of 6700 exome sequencing data from heart-, lung- and blood-related diseases and traits	Oct 2016[p]
1000Genome [80-82]	Over 88 million variants of 2504 whole genome sequencing data from 26 populations	Oct 2015[p]
Kaviar [83]	Over 162 million variants from 35 projects encompassing 13 200 genomes and 64 600 exomes	Feb 2016[w]
Genetically modified organism models		
MGD [84]	5021 associations between mouse genetic models and human diseases	Nov 2016[p]
MouseNet v2 [85]	788 080 functional gene network associations for laboratory mouse and eight other model vertebrates	Nov 2015[p]
MTB [86]	6057 associations between mouse genetic models-human cancer; 2288 associations between specific genes-cancers	Oct 2014[p]
RGD [87]	2998 associations between rat genetic models-human diseases	Nov 2016[p]
ZFIN [88]	11 348 associations between zebrafish genetic models-human diseases	Nov 2016[p]
Environmental exposures		
CTD [89]	1 379 105 chemical-gene associations, 202 085 chemical-disease associations and 33 583 gene-disease associations	Sep 2016[p]

Continued

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 2. (continued)

Database	Scope and scale	Date of statistic
ExposomeExplorer [44]	8034 concentrations correspond to dietary biomarkers (488) for 50 foods and 78 food compounds	Oct 2016[p]
CEBS [90]	Over 11 000 exposure agents and over 8000 exposure studies	Nov 2016[p]
SM2miR [91]	5161 associations between 1681 miRNAs and 255 small molecules	Apr 2015[p]
miREnvironment [92]	3857 associations between 1242 miRNAs, EFs and 305 phenotypes	Sep 2012[w]
DLREFD [93]	835 associations between 475 LncRNAs, 153 EFs and 124 phenotypes	Oct 2016[p]
Drug/chemical exposures		
ChEMBL [94]	Over 1.6 million distinct compound structures and 14 million activity values from over 1.2 million assays; ~11 000 drug targets including 9052 proteins	Nov 2016[p]
DrugBank 4.0 [95]	2037 FDA-approved small molecule drugs and 241 FDA-approved biotech (protein/peptide) drugs; over 6000 experimental drugs and over 201 SNP-associated drug effects, and 4661 drug targets	Nov 2013[p]
DrugCentral [96]	2021 FDA drugs, 2423 drugs approved outside US, 3799 small molecules, 239 peptides, 294 other drugs; 10 427 human protein targets including 837 drug efficacy targets	Oct 2016[p]
TTD [97]	2071 approved drugs, 7291 clinical trial drugs, 357 preclinical drugs, 17 803 experimental drugs397 successful targets, 723 clinical trial targets, 1469 research targets	Nov 2015[p]
PharmGKB [98]	20 017 associations between SNPs and drugs, and 65 important pharmacogenes	Jun 2018[w]
DGIdb [99]	40 017 mining clinically associations between 2644 genes and 11 215 drugs	Nov 2015[p]
CancerPPD [100]	3491 Experimentally verified anticancer peptides and 121 proteins spanning in 21 tissues	Sep 2014[p]

Scope refers to the major focus of the databases. The number of associations or items currently provided in the database is given. In the date of statistic, p indicates the Month-Year of statistic from journal publications; w refers to the Month-Year of statistic from official websites.

Genomic variations

Many genetic and complex diseases are associated with genomic variations and thus many genotype–phenotype databases store and curate genomic coverage of germline and somatic variations in single genes across the majority of genetic diseases, including Mendelian disorders, rare diseases and complex traits (Table 2). HGMD [63] is a representative repository for the clinical annotation of genetic mutations manually curated from more than 2600 peer-reviewed journals. HGMD has two types of version: the public version is freely available to users from academic institutions and non-profit organizations while the subscription version is available to all users under a commercial license provided by QIAGEN Inc. Another representative repository is ClinVar [7], which provides clinical annotation of genomic variation data. Data in ClinVar are submitted by clinical laboratory users and integrated from a variety of curated resources, including HGMD. Compared to HGMD, the freely available database LOVD provides not only the gene-centric collection and web search of nuclear DNA variations, but also the patient-centric data storage and storage of NGS data, even of variants outside of genes [66]. Moreover, MITOMAP reports 1746 human mitochondrial variants associated with diseases [67].

To provide standardization of annotation and improve accessibility of genomic variants, Li *et al.* developed VarCards to artificially generate all possible human single nucleotide variants (SNVs) in coding regions and splicing sites, and to classify all reported insertions and deletions (indels) [65]. VarCards has annotated variants from more than 60 genomic data sources, including disease-associated knowledge, functional effects, drug–gene interactions, predicted consequences through different *in silico* algorithms and allele frequencies in different population [65]. VarCards currently maintain over 110 million possible SNVs and more than 1.2 million reported indels (Table 2). Additionally, several other databases also cover genomic variations in genome-wide association studies (GWASs), such as GWAS Catalog [70], GWASdb [71], GWAS Central [72] and somatic variations in cancer, such as Catalogue of Somatic Mutations in Cancer (COSMIC) [68].

During recent years, abundant *de novo* variants and non-coding variants have been discovered in studies of complex diseases [64]. Novel variants of an individual not presented in either of his/her parents are termed *de novo* [113]. To facilitate better usages of the data of *de novo* variants, many databases have been established to integrate, characterize and annotate disease-related human *de novo* variants, including Denovo-db [64], NPdenovo [114] and Developmental Brain Disorder [115]. On the other hand, a few other databases focus on the disease/trait-related variants in human ncRNAs, ncRegion or their transcript factor binding sites (TFBSs), e.g. lncRNASNP [74], SNP2TFBS [116], miRdSNP [75], miRNASNP [117] and LincSNP 2.0 [73]. LincSNP specifically integrates and annotates disease-associated SNPs in human lncRNAs and TFBSs [73]. Similarly, miRNASNP [117] collects polymorphisms altering miRNA target sites, in order to identify miRNA-related SNPs in GWAS SNPs and eQTLs. The current miRNASNP [76] has integrated multiple filters to prioritize functional SNPs and experimentally supported miRNA-mRNA, as well as provided expression level annotation and correlation of miRNAs and target genes in various tissues.

These above resources often have a limitation that there is no mechanism for rapid improvement of the content and annotation of genomic variants. To address this, Griffith *et al.* have recently developed the CIViC knowledgebase for biocurators to annotate the clinical interpretation of variants in cancer which involves in the susceptible, diagnostic, therapeutic and prognostic relevance of somatic and germline variants of all types [69]. CIViC currently provides 1678 interpretations of clinical relevance for 713 variants affecting 283 genes associated with 209 cancer subtypes and 291 drugs. The variants in CIViC are annotated by provenance of supporting evidence and allowed users to transparently generate current and accurate variant interpretations [69].

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Altogether, these comprehensive resources of genomic variants with disease-related annotations are not only valuable for investigating the functions and mechanisms of coding genes and ncRNAs in human diseases, but also helpful for developing computational tools to functionally predict and interpret clinical significance of genomic variants in exome and genome sequencing data. According to the maturity and the annotation quality, HGMD, ClinVar, CIViC and COSMIC are highly recommended in this category.

Population genomic data

Population genomics examines genomic variations within and among various populations. NCBI's dbSNP is the first published population genomic database [20], which deposits SNPs and other classes of minor genetic variation including indels, copy number variations (CNVs) and structure variations from multiple resources [77]. With the NGS technology being widely adopted, several international projects have been launched to construct and integrate large number of genomic databases associated with populational phenotypes and features. These projects include National Heart, Lung and Blood Institute Exome Sequencing Project (NHLBI ESP), Exome Aggregation Consortium (ExAC), 1000 Genome and Kaviar (Table 2). NHLBI ESP [79] has offered an unprecedented depth to identify rare variants located in protein coding regions from about 6500 individuals who have been clinically diagnosed with heart, lung and blood disorders. Similarly, ExAC [118] has discovered rare variants from over 130 000 subjects whose exomes have been sequenced as part of various disease-specific and population genetic studies. Compared to NHLBI ESP and ExAC, the 1000 Genomes project provides a comprehensive resource for over 88 million human genomic variants in 2504 individuals from 26 populations [80–82]. 1000 Genomes also offers freely available RNA expression data from RNA sequencing and expression arrays, which can be explored to determine whether the genomic variants are associated with the changes of gene expression in RNA level [119]. Another consolidated database for allele frequencies is Kaviar [83] that contains genotype information of over 162 million variants from 35 projects, encompassing 13 200 genomes and 64 600 exomes. dbSNP is recommended for its quality annotation and maturity, Kaviar is recommended for its large scale of data in both genomes and exomes and 1000 Genomes is preferable for studying diseases associated with different populations.

Genetical organism models

Despite the recent success in identifying causative associations between genetic alterations and disorders, GPAs remain uncovered for many diseases. For example, almost half of the known genetic disorders recorded in the OMIM knowledgebase are still unclear for causative genes [120]. With the advanced technology of gene modifying and gene editing such as RNAi, Zinc-Finger Nuclease, TALENs and CRISPR/Cas system, a number of genetic modified organism models have been constructed to investigate genetic mechanisms in human diseases and to identify GPAs. The disease-associated information of genetically modified organism models is annotated and available from different databases, such as MGD [84], MouseNet [85], Mouse Tumor Biology (MTB) [86], RGD [87] and ZFIN [88, 121] (Table 2).

MGD is a highly integrated and curated database, housing comprehensive knowledge about mouse genes, genetic markers and genomic features as well as associations to various human diseases [84]. MGD also provides a portal of the Human-Mouse Disease Connection to facilitate the investigation of phenotypic similarity between mouse models and human patients. Similarly, RGD is a comprehensive data repository for laboratory rat, involving genomic and genetic variants as well as disease data [87]. The various disease portals at RGD are entry points of data and tools related to 12 classes of diseases, including cancer, diabetes, aging and cardiovascular disease. Compared to MGD, MTB is a database for mining data on tumor development and patterns of metastases [86]. It can facilitate the selection of strains in cancer research. In addition, Zebrafish (*Danio rerio*) is another useful model organism to investigate human disease, especially in developmental disorders. ZFIN is a central resource for zebrafish genomic, genetic, phenotypic and developmental data [88]. MGD, MTB, TGD and ZFIN house thousands of disease associations between the model species and human beings, involving cancer, mutation, congenic and transgenic constructions, etc. Other special organism model resources for rhesus monkey [122], dog [123], chicken [124], *Drosophila* [125] and *Caenorhabditis elegans* [126, 127] have also integrated confirmed association information between genetic makers and disorders. Thus, genetical organism models associated with diseases are useful resources for demonstrating and identifying the relationships between genetic alterations and phenotypes of human diseases.

Environmental exposures

Except for genetic factors, accumulative evidence has suggested that EFs have a great contribution to the development of many diseases, especially in complex disorders such as cancer and cardiovascular diseases [128–131]. Moreover, complex interaction between genetic factors and environmental exposures plays critical roles in developing the phenotypes of diseases. Several databases have been established to associate environment factors with protein coding genes and phenotypes of diseases [44, 90, 132–134] (Table 2). For example, the CTD [89] is a comprehensive repository of interactions between chemicals and gene products, as well as their relationships to diseases. The latest CTD contains over 30.5 million toxicogenomic relationships for the interactions of chemical-gene, chemical-disease and gene-disease [89]. Different from CTD, the Exposome Explorer database focuses on annotating biomarkers of exposure to environmental risk factors and dietary [44].

Recently, like other genetic factors, it has been suggested that miRNAs, lncRNAs and other type of ncRNAs also have complex interactions with a wide spectrum of exposure factors such as drugs [135], stress [136], alcohol [137], cigarette [138], virus [139], radiation [140], air pollution [141] and diet [142] in the development of diseases. With the rapid growth of interaction data between ncRNAs, environmental exposures and development of diseases, a number of databases have been generated to describe their relationships, such as SM2miR [91], miREnvironment [92], DLREFD [93] and LncEnvironmentDB [43] (Table 2). SM2miR is the first established database to provide experimentally validated effects of small molecules on miRNA expression and hosts manually curated association data between miRNAs and small molecules across 17 species [91]. Compared to SM2miR, miREnvironment not only provides manually curated information on environmental exposures and miRNA expression, but also offers phenotypes associated with miRNAs and EFs [92]. Different from SM2miR and miREnvironment for miRNAs, DLREFD [93] and LncEnvironmentDB [43] focus on the lncRNAs that are exper-

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

imentally or computationally associated with environmental exposures and disease-related phenotypes.

These environment-related databases (Table 2) are valuable data resources for investigating the impacts of EFs on the development of human diseases at the molecular level as well as at the network level. Due to the large numbers of associations, CTD is highly recommended for coding genes associated with environmental and chemical exposures in this category.

Drugs and their targets

To facilitate successful medicine research with comprehensive information across drug discovery and development process, several public repositories have been established to dedicate associations across phenotypes, drugs, chemicals and their targets (Table 2). Therapeutic Target Database (TTD) is the earliest repository [143] to provide information about drugs, targets and their associations with specific pathways. DrugBank [95] and DrugCentral [96] are the other two main databases, hosting comprehensive drug-target interactions and drug action information captured and integrated from online non-commercial resources, e.g. US Food and Drug Administration (FDA), European Medicines Agency and Japan Pharmaceutical and Medical Devices Agency, as well as curated data from published research articles and drug labels. DrugBank and DrugCentral have become the referential drug data source for a number of well-known public databases such as PubChem [144], ChEMBL [94], PharmGKB [98], UniProt [145]and SuperTarget [146]. Moreover, TTD, DrugBank and DrugCentral link to targets and pathways to *in silico* drug discovery efforts. Other notable databases include PharmGKB [98] for impact of human genetic variations on drug responses, and the Drug-Gene Interaction Database (DGIdb) [99] for drug–gene interactions and gene druggability. Moreover, several databases have integrated drug-target information with special medical indications, such as cancer [100, 147, 148], side effects [149], pharmacophores [150] and special metabolic pathways [151]. The data resources of drugs with diseases enable the investigations of drug effects in specific genetic contexts and provide new insights in drug action at the molecular level. Due to the maturity and the data quality, ChEMBL and DrugBank are recommended for drug annotation in this category. On the other hand, PharmGKB is recommended for the interpretation of impact of human genetic variations on drug responses.

Software tools and web platforms

Software tools and web platforms are another type of computational resources, accelerating deeper understanding associations between multiple disease-related factors. Most of the available public software tools used to bridge the gaps between biology, medicine and clinic are driven by either genomic features or ontologies. These tools can be downloaded and used to analyze data in a standalone computer. To analyze online, several web platforms have been constructed to include interactive applications that comprehensively integrate a variety of disease-related data sources and software tools to prioritize disease-related associations spanning genotypes, phenotypes and treatments.

Genomic feature-driven tools

To facilitate clinical interpretation of genetic and genomic factors, many computational tools have been developed based on various features including evolutionary conservation, sequence homology and genomic and epigenetic annotations (Table 3). These computational tools have been widely used to annotate, predict and prioritize functional effects of varieties of genomic variants from high-throughput sequencing data, including KGGSeq [152, 153], ANNOVAR [12] and wANNOVAR [154] for functional annotation of genetic variants, VEST3 [155] and REVEL [156] for prioritization of rare missense variants, GWAVA [47] and Deepsea [14] for prioritization of noncoding variants, MutationTaster [157], VAAST [46], CADD [49], DANN [158], FATHMM-MKL [159] and Eigen [13] for prediction of the functional consequences of both coding and non-coding variants (Table 3). Some past research attempted to compare the usage and performance of these tools. It has been shown that Eigen has better discriminatory ability than CADD using disease-related variants and putatively benign variants in both noncoding and coding regions [13]. Moreover, M-CAP [160] and InterVar [161] were developed to eliminate the majority of variants of uncertain significance and facilitate interpretation of clinical significance of variants (Table 3). Furthermore, SIFT [45], LRT [162], PolyPhen2 [11], MutationAssessor [163], PROVEAN [164], FATHMM [165], MetaSVM [166] and IMHOTEP [167] have been developed to predict functional impacts of amino acid substitutions (Table 3). On predictions of polymorphisms and mutations with variants causing single amino acid substitutions, MutationTaster2 [168] had the highest accuracy compared to SIFT, PolyPhen-2 and PROVEAN. Different from all the above tools, ClinLabGeneticist [169] was established to manage clinical genetic variants from whole exome sequencing based on extensive variants annotation data (Table 3). ClinLabGeneticist contains information of data entry, distribution of work assignments and selection of variants for validation, report generation and communications between various personnel, and the entire workflow of ClinLabGeneticist has been integrated into a single data management platform.

Ontology-driven tools

The ontology databases in life science, such as Human Phenotype Ontology (HPO) [170–174], Mammalian Phenotype Ontology [175], Disease Ontology [176], Gene Ontology (GO) [177] and Experimental Factor Ontology (EFO) [178], provide standard terminologies and controlled vocabularies to describe and classify molecules, diseases, genotypic and phenotypic features, etc. The ontologies can be utilized to support computational tools that allow sophisticated search and analysis routines. For example, HPO offers standard terminologies for phenotypic features and diseases, to bridge the gap between genome biology and clinical medicine [179]. Several tools use phenotypic ontologies to enable deep interpretation for the analysis results of NGS data, including eXtasy [180], PhenIX [15], Exomiser [181], Phen-Gen [182], Phevor [48] and PhenogramViz [183] (Table 4). eXtasy, the earliest tool of them, ranks the damaging impacts of nonsynonymous single-nucleotide variants (nSNVs) by genomic data fusion. PhenIX evaluates and prioritizes impacts of SNVs, splice sites and short indels in the exome sequencing data of Mendelian diseases based on pathogenicity of variants and semantic similarity of HPO-based phenotypes [15]. Compared to PhenIX, Phen-Gen implements an exome-centric approach to rank the impacts of coding mutations, and a genome-wide approach to prioritize pathogenicity of non-coding variants (Table 4). Similar to Phen-Gen, the recently developed tool Exomiser [184] integrates a number of algorithms, including HiPHIVE [185], PHIVE [186], ExomeWalker [187] and OWLSim [188], to enable the clinical interpretation of variants in exome

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 3. Genomic feature-driven tools for annotation and evaluation of clinical significance of variants

Application	Year of first deployment: tool name	Regular update	Based on
Functional annotation of genomic variants	2010: ANNOVAR [12]	Yes, annually since 2015	Functional annotation of genetic variants from high-throughput sequencing data
	2012: wANNOVAR [154]	Yes	Functional annotation of genetic variants from high-throughput sequencing data
	2012: KGGSeq [152, 153]	Yes, bugs fixed monthly	Three different levels: genetic level, variant-gene level and knowledge level
Prediction of functional impact of amino acid substitutions	2003: SIFT [45]	Last update in Aug 2011	Sequence homology based on PSI-BLAST
	2009: LRT [162]	Last update in Nov 2009	Sequence homology
	2010: PolyPhen2 [11]	Last update in 2016	Eight sequence-based and three structure-based predictive features
	2011: MutationAssessor [163]	Last update in Dec 2015	Sequence homology of protein families and subfamilies between species
	2012: PROVEAN [164]	Last update in Jan 2015	Sequence homology
	2013: FATHMM [165]	Last update in May 2015	Sequence homology
	2015: MetaSVM [166]	Last update in 2016	9 prediction scores and allele frequencies in 1000Genomes
	2017: IMHOTEP [167]	Unknown	9 popular predicted tools
Prioritization of rare missense variants	2013: VEST3 [155]	Yes, quarterly	86 sequence features
	2016: REVEL [156]	Last update in 2016	13 popular predicted tools
	2016: M-CAP [160]	Last update in 2016	Pathogenicity likelihood scores and direct measures of evolutionary, conservation, the cross-species analog to frequency within the human population
Prioritization of noncoding variants	2014: GWAVA [47]	Last update in 2014	Various genomic and epigenomic annotations
	2015: DeepSEA [14]	Yes, annually	Regulatory sequence code
Prediction of functional consequences for both coding and non-coding variants	2010: MutationTaster [157]	Yes	Conservation, splice site, mRNA features, protein features and regulatory features
	2011: VAAST [46]	Last update in Sep 2016	Variant frequency data with AAS effect information on a feature-by-feature basis
	2014: CADD [49]	Last update in Apr 2018	63 annotations including 949 sequence features
	2015: DANN [158]	Last update in 2015	63 annotations including 949 sequence features that is same to CADD
	2015: FATHMM-MKL [159]	Last update in 2015	1281 sequence features
	2016: Eigen [13]	Last update in 2016	Functional, evolutionary conservation and regulatory annotations
Interpretation of clinical significance of variants	2017: InterVar [161]	Yes, last update in Jan. 2018	The-2015-ACMG-AMP-Guidelines
	2015: ClinLabGeneticist [169]	Last update in 2014	Extensive variant annotation data source and prioritization of variants

The tools are classified into different categories according to their uses.

and genome sequencing data. Instead of postulating a set of fixed associations between genes, diseases and phenotypes, Phevor dynamically integrates various knowledge of multiple biomedical ontologies into the variant-ranking process [48]. This enables Phevor to improve its accuracy not only of established gene-disease-phenotype associations but also of previously atypical and undescribed disease statements. PhenogramViz focuses on the interpretation of candidate CNVs and their pathogenicity prioritization from the data analyses of array comparative genome hybridization (aCGH) and NGS [183].

In the performance aspect of causal gene identification, previous researches indicate that Phen-Gen gains 13∼58% improvement in sensitivity over eXtasy, Phevor, PHIVE and the earlier version of Exomiser [182]. Bone *et al.* [181] suggest that Exomiser is slightly favorable compared to Phen-Gen in the causal gene identification for autosomal dominant disorders and autosomal recessive disorders as well as the detection of novel variant-disease associations [181]. Moreover, Exomiser can analyse multiple samples or families per run for both Mendelian and multigenic disorders, while Phen-Gen can only handle single sample or family per run for Mendelian disorders (Table 4).

eXtasy and Phen-Gen have both online and standalone versions of programs. The standalone eXtasy has many library dependencies of bioinformatics, statistics and machine learning algorithms (Table 4). Exomiser has the standalone version only, while PhenIX and Phevor have online versions instead. PhenogramViz can be downloaded, installed as an application in Cytoscape [189], and used through the Cytoscape interface. The standalone tools can be installed locally and run within hospital firewalls, thereby relieving the concerns of privacy and security for the information of patients. On the other hand, the online version tools are more acceptable and useable for many biologic researchers and clinicians, who lack bioinformatic and computing skills. In the timing aspect, the online eXtasy

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 4. Comparison of phenotype-driven tools for interpretation of clinical significance of variants

Year: tool	Availability	Operation System	Requirements	Algorithms implemented	Input data and parameter	Application scopes
2013: eXtasy [185]	Online & Standalone	Linux	Ruby; Tabix; Bedtools; R Statistical Framework with randomForest; RobustRankAggreg libraries	Random Forests; Phenomizer	VCF file; TSV for HPO term(s)	Mendelian and oligogenic disorders; nSNVs; Exome analysis; (Only 1 sample per run)
2014: Phen-Gen [187]	Online & Standalone	Linux (Ubuntu, CentOS, & RHEL)	Perl	Bayesian framework; Random walk–with–restart; Variant-predicted pathogenicity score; Phenomizer	VCF file; text file for HPO term(s); Pedigree(PED) file; Inheritance models; Type of prediction-genomic or coding; Discard de novo and Stringency	Rare disorders; nSNVs, splice-sites and short indels and non-coding variants; Genome and Exome analysis; (Only 1 family or 1 sample per run)
2014: PhenIX [15]	Online	-	-	Semantic similarity score; Variant-frequency score; Variant-predicted pathogenicity score	VCF file; HPO term(s); Inheritance modes; Frequency sources; Number of candidates to show	Mendelian diseases; SNVs, splice-sites and short indels; Exome analysis; (Only 1 sample per run)
2014: Phevor [54]	Online	-	-	Disease-gene association score; Variant-prioritization score	VAAST simple or Table for variants; Ontology Term(s); Ontologies to link to HPO	Rare disorders; SNVs; Exome analysis; (Only 1 sample per run)
2016: Exomiser [186]	Standalone	Linux; Mac OS X; Windows	~4GB RAM for an exome analysis and ~12GB RAM for a genome analysis; >3 GB free RAM (8 GB preferred); Java 8 or above	HiPHIVE; PHIVE; PhenIX; Exome Walker; OWLSim; Logistic regression	YML file that include VCF file name; HPO term(s); PED file name; inheritance modes, Probands; Frequency sources; Pathogenicity sources and other alterative parameters	Mendelian, oligogenic and multigenic disorders; SNVs, splice-sites, short indels and non-coding variants; Genome and Exome analysis; (Multiple samples or families per run)
2014: Phenogram Viz [188]	Cytoscape app	Windows	Cytoscape Version 3.1.0. and above	Phenogram-score (PHS); NAG, OBE, OPA, HI score	Enter symptom(s) directly for symptoms or create file with HPO term(s); Lists of CNVs (include types, Chromosome, Start, End); Lists of genes	Mendelian disorders; CNVs; aCGH and exome analysis; (Only 1 sample per run)

The availabilities, the requirements and the use of these tools are detailed in the table.

takes about 15 min to analyze a whole exome data sample with ~82 000 variants, while the online PhenIX takes about 100 s to complete the same analysis, much faster than eXtasy. Exomiser [184] consumes about 10~15 min to analyze an exome and genome sample or family, approximately 5–15 min faster than the online Pen-Gen (http://54.173.20.191). Moreover, Exomiser [184] produces HTML, tab-delimited and VCF format files that can be incorporated into many bioinformatic workflows.

Taken together, the standalone versions of Phen-Gen and Exomiser are recommended to skilled bioinformaticians for the interpretation of SNVs, splice-sites, short indels and non-coding variants from data of exomes and genomes. Exomiser is also suggested for the analysis of multiple samples or families. Phevor is recommended for the prioritization of variants pathogenicity related to previously atypical and undescribed disease statements, and PhenogramViz for the interpretation of CNVs pathogenicity.

Interactive platforms

To tackle the hurdles in utilising disease-related data resources, several web platforms have implemented a number of analysis software tools to allow users to search, analyze and visualize the resources through web interface and APIs (Table 5). Most of these platforms, such as DisGeNET [190], Open Targets [16], Monarch Initiative [52] and MalaCards [51], target on human Mendelian and complex diseases, involving data of genotypes,

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

Table 5. Summary of different biomedical data and analysis web platforms

Name	Scope and scale (Date of statistic)	Applications/Tools Available	Sources
DisGeNET [190]	**GPAs** (May 2017)[w]: 429 036 associations between 17 381 genes and 15 093 human diseases; 72 870 associations between 46 589 SNPs and 6356 human diseases/phenotypes	Web interface, DisGeNET Cytoscape plugin, Disgenet2r R package, DisGeNET-RDF	UniProt, dbSNP, GDA, CTD, MGD, OMIM, Clinvar, RGD, GWAS Catalog, Orphaned, HPO, UMLS, MeSH, DO, ICD9-CM, HGNC, dbSNP, CTD in total 22 resources
Monarch Initiative [52]	**Genetically modified model support GPA** (Nov 2016)[p]: 237 531 gene-phenotype associations in human; 1 489 573 variant-phenotype associations in human; 19 783 disease models	Web interface, Phenotypes Analyzer, PhenoGrid, Text annotator, Exomiser	ClinVar, CTD, GeneReviews, OMIM, HPO, Orphanet, GWAS Catalog, MGI, ZFIN, NCBI, UCSC, HGNC, MeSH, OMIM, ORDO, HPO, EFO, UMLS in total 53 resources
Open Targets Platform [16]	**Genotype–phenotype-drug association** (Apr 2018)[w]: 2 336 807 associations between genes/variants/drugs and diseases/phenotypes/targets	Web interface, Phylogenetic tree and HEART, Application programming interface	GWAS Catalog, UniProt, Expression Atlas, ChEMBL, Reactome, PhenoDigm, UMLS, MeSH, GO, ECO, HPO, MP, OMIM, ICD9-CM in total 21 resources
MalaCards [51]	**Genotype–phenotype-drug association** (Nov 2016)[p]: 10 198 genes associated with 13 619 disease entries; 966 338 associations between 8005 distinct diseases and 3017 distinct drugs	Web interface, Tgex, GeneAnalytics, VarElect GeneALaCart, PathCards	Clinvar, Cosmic, dbSNP, DGIdb, DrugBank, FDA, HGMD, OMIM, PharmGKB, ICD10, MeSH, MGI, UMLS, UniProt in total 68 resources
MARRVEL [191]	**GPA** (June 2017)[p]: 12.3 million variants; 6.95 million genotype–phenotype relationships	Web interface, Mutalyzer Position Converter, OMIM API, DIOPT, GTEx	ExAC, gnomAD, IMPC, Monarch, ClinVar, Geno2MP, DGV, DECIPHER, DIOPT, Mutalyzer, SGD, PomBase, WormBase, FlyBase, ZFin, MGI and RGD in total 17 resources

Scope refers to the major focus of the web platform. Scale is the number of associations and items currently provided in the platform. Each platform has integrated multiple tools/applications. Sources refer to the original data resources that have been integrated in the platform. In the date of statistic, p indicates the Month-Year of statistic from journal publications; w refers to the Month-Year of statistic from official websites.

phenotypes, genetically organism models, drugs targets and chemical molecules.

The distinctions between different platforms are reflected in their different focuses and different applications. DisGeNET [190] is designed to collate GPAs and to offer tool applications for medical and biological research. It can be plugged into Cytoscape to visualise and explore gene-disease associations in bipartite networks [17] (Table 5). Open Targets and MalaCards not only integrate GPA information from OMIM, GWAS Catalog, ClinVar, UniProtKB and disease model databases, but also offer information of target-diseases related to approved drugs, clinical candidates, biological pathways and RNA expressions (Table 5). Due to their comprehensive knowledgebases, sophisticated web technologies as well as User Experience designs, Open Targets and MalaCards have been considered as effective platforms for medicine research. For instance, Open Targets provide two types of workflows to enable effective applications for different destinations which are as follows: the disease-centric workflow to identify targets (such as genes, variants, proteins and chemicals) associated with a specific disease, and the target-centric workflow to identify diseases associated with a specific target [16]. Moreover, Monarch Initiative semantically integrates genotype–phenotype resources from many species for exploring their relationships across species [52]. Based on its broad genotype–phenotype information, many tool applications have been developed on Monarch Initiative, including Phenogrid for phenotype analysis [52], text annotators [52] for text annotation of genes, diseases and phenotypes, Exomiser [181] for inferring causative variants (Table 5). MARRVEL [191] is another publicly available platform integrating multiple model organism resources for rare variant exploration. It improves accessibility of data collection and facilitates analysis of human genes and variants by aggregating about 18 million public data records (Table 5).

Altogether, these platforms have not only facilitated the research in life sciences, but also greatly supported the development of precision clinical medicine. They can be used for the investigation of causes of specific human diseases and their comorbidities, the discovery of therapeutic action and adverse effects, the validation of computationally predicted phenotypes and genotypes and the evaluation of text-mining methods performance.

Discussion

The computational resources have facilitated deeper understanding of disease mechanisms, easier assessment of disease risks and more accurate diagnoses, and also helped to guide clinical therapies as well as to evaluate prognosis. However, challenges remain in many aspects, such as building complex networks of associations, database design for bigger data, data analysis with more effective tools and platforms, data interpretation in consistent and standard manners, result representation with user friendly interfaces and so on.

Phenotype plays a central role in connection with other disease-related factors in the current network (Figure 2). The focus of software and database development is being shifted from the connection between genotypes and phenotypes to the association among multiple factors. As wider collaborations have been made to establish interoperable systems across international projects, much bigger data are being generated by many complete genomes of whole populations. Difficulties exist in connecting much more complex and multi-dimensional data.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

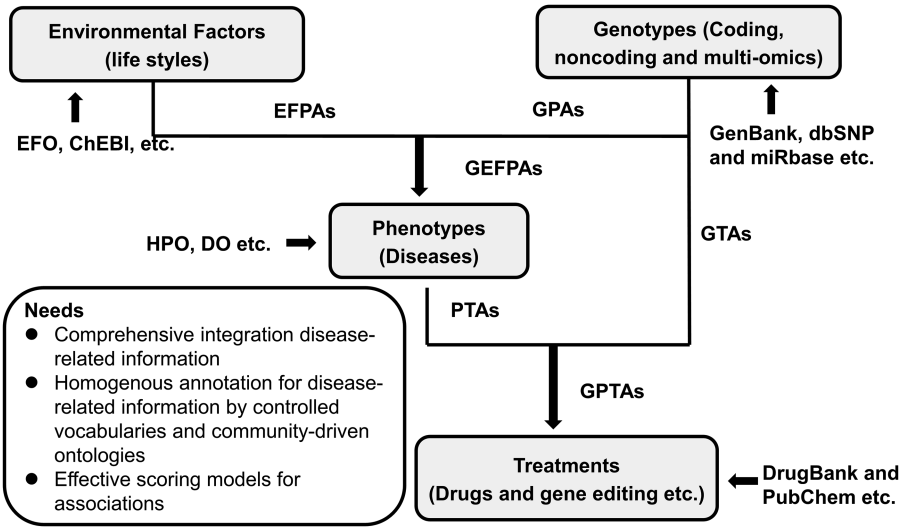

Figure 2. Framework of a comprehensive web platform. A comprehensive web platform should integrate various disease-related information including genotypes, phenotypes, environmental factors, life styles and so on. The available information in the platform should be homogenously annotated by controlled vocabularies and community-driven ontologies, such as GenBank, dbSNP and miRbase for genotypes, HPO and DO for phenotypes, EFO and ChEBI for environmental factors and life styles, DrugBank and PubChem for drugs. Moreover, the platform should have solid scoring models to prioritize associations between different factors, such as genotype-phenotype associations (GPAs), environmental factor-phenotype associations (EFPAs), genotype-environmental factor-phenotype associations (GEFPAs), phenotype-treatment associations (PTAs), genotype-treatment associations (GTAs) and genotype-phenotype-treatment associations (GPTAs).

Moreover, additional data types including multi-omics results, extensive environmental contexts and life styles of patients are necessary to integrate and associated in the current network. Obviously, more effective algorithms and software tools are greatly needed to take more related factors, additional data types and bigger size of data into account.

Although the approaches of deep phenotyping are helpful for clinical diagnosis in Mendelian disorders and rare diseases, patients with similar features or at a same stage of illness often have various clinical outcomes in cancer and many complex diseases [2]. Existing spectrum of phenotype states is not optimally captured by current phenotypic ontology systems. Therefore, substantial efforts are required to better integrate the ontologies and enable the full interpretation of clinical outcomes of genetic mutations that may lead to the precision management of diseases.

Currently, there are abundant biomedical resources that cover disease information involving in genotypes, phenotypes, environmental exposures and their associations. However, most of the popular resources only represent a fraction of available information. Therefore, more comprehensive platforms are needed to integrate other ever-growing biomedical information, such as noncoding genetic factors, multi-omics and extensive environmental contexts and life styles (Figure 2). In addition, these platforms should integrate clinical, environmental contexts and life styles of patients to enable reliable and useful diagnoses and discoveries, and also make data fully accessible and easily interpreted through with highly graphical representation. Moreover, the available information in majority of databases is represented and annotated by heterogenous vocabularies (Supplementary S-Table 2). Thus, better platforms are needed to comprehensively integrate the available information with controlled vocabularies and community-driven ontologies and present analysis results in a consistent and standard manner (Figure 2). Recently, MNDR has been updated to offer confidence score of each ncRNA-disease association based on a simple classification of supporting evidences [62]. However, to better support translational research and precision medicine, there is a great need to develop solid scoring models or to refine current models based on experimental evidences to assist the prioritization of associations, such as GPAs, EF-phenotype associations, genotype-EF-phenotype associations, phenotype-treatment associations, genotype-treatment associations and genotype–phenotype-treatment associations (Figure 2).

In this review, we detail the human disease-related computational resources, including databases, software tools and online platforms. These resources are classified by disparate data types with focuses on association among genotypes, phenotype, EFs, organism models, drugs and chemical molecules. We also provide some of the resulting needs and requirements that should be regarded as imperative for the development of databases, tools and platforms (Figure 2).

From the view of precision medicine, better services of computation resources and more training on these services will accelerate better medical research and clinical diagnoses as well as treatments. Life scientists, bioinformaticians and clinicians are suggested to cooperate to develop more comprehensive databases, more accurate software tools and more practical platform systems to facilitate the goals of precision medicine, enabling reliable and useful diagnoses and discoveries.

Key Points

- The present study is a comprehensive review of available computational resources of human diseases, including databases, software tools and interactive platforms to assist in the appropriate selection and use of relevant resources.
- Bioinformaticians have developed more than 100 computational resources to integrate omics data and discover associations among genotypes, phenotypes,

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

environmental exposures, drugs and chemical molecules.
- According to scopes and data associations, the databases can be categorized into seven groups, including coding genes, noncoding RNAs, genomic variations, population genomic data, genetical organism models, environment exposures and treatments.
- Most of the available public software tools used to bridge the gaps between biology, medicine and clinic are driven by either genomic features or ontologies.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the National Key R&D Program of China [2016YFC0901604]; and the National Natural Science Foundation of China [31771478].

References

1. Auton A, Brooks LD, Durbin RM, *et al*. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
2. Brookes AJ, Robinson PN. Human genotype-phenotype databases: aims, challenges and opportunities. *Nat Rev Genet* 2015;**16**:702–15.
3. Lek M, Karczewski KJ, Minikel EV, *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285–91.
4. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**:628–40.
5. Gkoutos GV, Schofield PN, Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings Bioinform* 2017.
6. Amberger JS, Bocchini CA, Schiettecatte F, *et al*. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015;**43**:D789–98.
7. Landrum MJ, Lee JM, Benson M, *et al*. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**:D862–8.
8. Wong KM, Langlais K, Tobias GS, *et al*. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res* 2017;**45**:D819–26.
9. Tryka KA, Hao L, Sturcke A, *et al*. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014;**42**:D975–9.
10. Walker L, Starks H, West KM, *et al*. dbGaP data access requests: a call for greater transparency. *Sci Transl Med* 2011;**3**:113c–34c.
11. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.
13. Ionita-Laza I, McCallum K, Xu B, *et al*. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–20.
14. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931–4.
15. Zemojtel T, Kohler S, Mackenroth L, *et al*. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 2014;**6**:123r–252r.
16. Koscielny G, An P, Carvalho-Silva D, *et al*. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* 2017;**45**:D985–94.
17. Pinero J, Bravo A, Queralt-Rosinach N, *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9.
18. Cooper DN, Krawczak M. Human Gene Mutation Database. *Hum Genet* 1996;**98**:629.
19. Krawczak M, Cooper DN. Core database. *Nature* 1995;**374**:402.
20. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 1999;**9**:677–9.
21. Ayme S, Urbero B, Oziel D, *et al*. Information on rare diseases: the Orphanet project. *Rev Med Interne* 1998;**19**(Suppl 3):376S–7S.
22. Blake JA, Richardson JE, Davisson MT, *et al*. The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. *The Mouse Genome Database Group. Nucleic Acids Res* 1999;**27**:95–8.
23. Pargent W, Heffner S, Schable KF, *et al*. MouseNet database: digital management of a large-scale mutagenesis project. *Mamm Genome* 2000;**11**:590–3.
24. Twigger S, Lu J, Shimoyama M, *et al*. Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res* 2002;**30**:125–8.
25. Sprague J, Clements D, Conlin T, *et al*. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res* 2003;**31**:241–3.
26. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002;**30**:163–5.
27. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
28. Wang Y, Xiao J, Suzek TO, *et al*. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 2009;**37**:W623–33.
29. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell* 2009;**136**:629–41.
30. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
31. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
32. Managadze D, Rogozin IB, Chernikova D, *et al*. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol* 2011;**3**:1390–404.
33. Kapranov P, Cheng J, Dike S, *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007;**316**:1484–8.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

34. Liu C, Bai B, Skogerbo G, *et al*. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res* 2005;**33**:D112–5.
35. Jiang Q, Wang Y, Hao Y, *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**:D98–104.
36. Chen G, Wang Z, Wang D, *et al*. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 2013;**41**:D983–6.
37. Abecasis GR, Altshuler D, Auton A, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
38. Weinstein JN, Collisson EA, Mills GB, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
39. Collins FS, Barker AD. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* 2007;**296**:50–7.
40. Hudson TJ, Anderson W, Artez A, *et al*. International network of cancer genome projects. *Nature* 2010;**464**:993–8.
41. Samuel GN, Farsides B. The UK's 100,000 Genomes Project: manifesting policymakers' expectations. *New Genet Soc* 2017;**36**:336–53.
42. Mattingly CJ, Colby GT, Forrest JN, *et al*. The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect* 2003;**111**:793–5.
43. Zhou M, Han L, Zhang J, *et al*. A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases. *Mol Biosyst* 2014;**10**:3264–71.
44. Neveu V, Moussy A, Rouaix H, *et al*. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res* 2017;**45**:D979–84.
45. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**: 3812–4.
46. Yandell M, Huff C, Hu H, *et al*. A probabilistic disease-gene finder for personal genomes. *Genome Res* 2011;**21**:1529–42.
47. Ritchie GR, Dunham I, Zeggini E, *et al*. Functional annotation of noncoding sequence variants. *Nat Methods* 2014; **11**:294–6.
48. Singleton MV, Guthery SL, Voelkerding KV, *et al*. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 2014;**94**: 599–610.
49. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.
50. Ryan P, Dan N, Jojo D, *et al*. *Creating a universal SNP and small indel variant caller with deep neural networks*. 2016.
51. Rappaport N, Twik M, Plaschkes I, *et al*. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2017;**45**:D877–87.
52. Mungall CJ, McMurry JA, Kohler S, *et al*. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2017;**45**:D712–22.
53. Pavan S, Rommel K, Mateo MM, *et al*. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PloS One* 2017;**12**:e170365.
54. Wright CF, Fitzgerald TW, Jones WD, *et al*. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;**385**:1305–14.
55. Gazzo AM, Daneels D, Cilia E, *et al*. DIDA: a curated and annotated digenic diseases database. *Nucleic Acids Res* 2016;**44**:D900–7.
56. Xiong Y, Wei Y, Gu Y, *et al*. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res* 2017;**45**:D888–95.
57. Zhao Y, Li H, Fang S, *et al*. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 2016;**44**:D203–8.
58. Li Y, Qiu C, Tu J, *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**:D1070–4.
59. Ning S, Zhang J, Wang P, *et al*. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res* 2016;**44**:D980–5.
60. Wang J, Cao Y, Zhang H, *et al*. NSDNA: a manually curated database of experimentally supported ncRNAs associated with nervous system diseases. *Nucleic Acids Res* 2017;**45**:D902–7.
61. Zhao Z, Wang K, Wu F, *et al*. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis* 2018;**9**:475.
62. Cui T, Zhang L, Huang Y, *et al*. MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res* 2018;**46**:D371–4.
63. Stenson PD, Mort M, Ball EV, *et al*. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;**136**:665–77.
64. Turner TN, Yi Q, Krumm N, *et al*. denovo-db: a compendium of human de novo variants. *Nucleic Acids Res* 2017;**45**: D804–11.
65. Li J, Shi L, Zhang K, *et al*. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res* 2018;**46**:D1039–48.
66. Fokkema IF, Taschner PE, Schaafsma GC, *et al*. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* 2011;**32**:557–63.
67. Ruiz-Pesini E, Lott MT, Procaccio V, *et al*. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* 2007;**35**:D823–8.
68. Forbes SA, Beare D, Boutselakis H, *et al*. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**:D777–83.
69. Griffith M, Spies NC, Krysiak K, *et al*. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**:170–4.
70. MacArthur J, Bowler E, Cerezo M, *et al*. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;**45**:D896–901.
71. Li MJ, Liu Z, Wang P, *et al*. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* 2016;**44**:D869–76.
72. Beck T, Hastings RK, Gollapudi S, *et al*. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet* 2014;**22**:949–52.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

73. Ning S, Yue M, Wang P, *et al*. LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res* 2017;**45**:D74–8.
74. Miao YR, Liu W, Zhang Q, *et al*. lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res* 2018;**46**:D276–80.
75. Bruno AE, Li L, Kalabus JL, *et al*. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. *BMC Genomics* 2012;**13**:44.
76. Gong J, Liu C, Liu W, *et al*. An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database (Oxford)* 2015;**2015**:v29.
77. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2015;**43**:D6–17.
78. Karczewski KJ, Weisburd B, Thomas B, *et al*. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 2017;**45**:D840–5.
79. Auer PL, Reiner AP, Wang G, *et al*. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. *Am J Hum Genet* 2016;**99**:791–801.
80. Sudmant PH, Rausch T, Gardner EJ, *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81.
81. Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
82. Auton A, Brooks LD, Durbin RM, *et al*. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
83. Glusman G, Caballero J, Mauldin DE, *et al*. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 2011;**27**:3216–7.
84. Eppig JT, Blake JA, Bult CJ, *et al*. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* 2015;**43**:D726–36.
85. Kim E, Hwang S, Kim H, *et al*. MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. *Nucleic Acids Res* 2016;**44**: D848–54.
86. Krupke DM, Begley DA, Sundberg JP, *et al*. The Mouse Tumor Biology Database: a comprehensive resource for mouse models of human cancer. *Cancer Res* 2017;**77**:e67–70.
87. Shimoyama M, De Pons J, Hayman GT, *et al*. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* 2015;**43**: D743–50.
88. Howe DG, Bradford YM, Eagle A, *et al*. The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. *Nucleic Acids Res* 2017;**45**:D758–68.
89. Davis AP, Grondin CJ, Johnson RJ, *et al*. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2017;**45**:D972–8.
90. Lea IA, Gong H, Paleja A, *et al*. CEBS: a comprehensive annotated database of toxicological data. *Nucleic Acids Res* 2017;**45**:D964–71.
91. Liu X, Wang S, Meng F, *et al*. SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 2013;**29**:409–11.
92. Yang Q, Qiu C, Yang J, *et al*. miREnvironment database: providing a bridge for microRNAs, environmental factors and phenotypes. *Bioinformatics* 2011;**27**:3329–30.
93. Sun YZ, Zhang DH, Ming Z, *et al*. DLREFD: a database providing associations of long non-coding RNAs, environmental factors and phenotypes. *Database (Oxford)* 2017;**2017**.
94. Gaulton A, Hersey A, Nowotka M, *et al*. The ChEMBL database in 2017. *Nucleic Acids Res* 2017;**45**:D945–54.
95. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;**42**:D1091–7.
96. Ursu O, Holmes J, Knockel J, *et al*. DrugCentral: online drug compendium. *Nucleic Acids Res* 2017; **45**:D932–9.
97. Li YH, Yu CY, Li XX, *et al*. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2018;**46**:D1121–7.
98. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012;**92**:414–7.
99. Cotto KC, Wagner AH, Feng YY, *et al*. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res* 2017.
100. Tyagi A, Tuknait A, Anand P, *et al*. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015;**43**:D837–43.
101. Schaffer AA . Digenic inheritance in medical genetics. *J Med Genet* 2013;**50**:641–52.
102. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med* 2013;**19**: 1438–49.
103. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74.
104. Lv J, Liu H, Su J, *et al*. DiseaseMeth: a human disease methylation database. *Nucleic Acids Res* 2012;**40**:D1030–5.
105. Huang WY, Hsu SD, Huang HY, *et al*. MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic Acids Res* 2015;**43**:D856–61.
106. Bertone P, Stolc V, Royce TE, *et al*. Global identification of human transcribed sequences with genome tiling arrays. *Science* 2004;**306**:2242–6.
107. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**:D68–73.
108. Volders PJ, Verheggen K, Menschaert G, *et al*. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* 2015;**43**: D174–80.
109. Quek XC, Thomson DW, Maag JL, *et al*. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* 2015;**43**:D168–73.
110. Wang Y, Chen L, Chen B, *et al*. Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis* 2013;**4**:e765.
111. Wang J, Ma R, Ma W, *et al*. LncDisease: a sequence based bioinformatics tool for predicting lncRNA-disease associations. *Nucleic Acids Res* 2016;**44**:e90.
112. Ghosal S, Das S, Sen R, *et al*. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* 2013;**4**:283.
113. Kong A, Frigge ML, Masson G, *et al*. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012;**488**:471–5.
114. Li J, Cai T, Jiang Y, *et al*. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* 2016;**21**:298.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

115. Gonzalez-Mantilla AJ, Moreno-De-Luca A, Ledbetter DH, *et al*. A cross-disorder method to identify novel candidate genes for developmental brain disorders. *JAMA Psychiatry* 2016;**73**:275–83.
116. Kumar S, Ambrosini G, Bucher P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* 2017;**45**: D139–44.
117. Gong J, Tong Y, Zhang HM, *et al*. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. *Hum Mutat* 2012;**33**:254–63.
118. Song W, Gardner SA, Hovhannisyan H, *et al*. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet Med* 2016;**18**:850–4.
119. Kannan L, Ramos M, Re A, *et al*. Public data and open source tools for multi-assay genomic investigation of disease. *Brief Bioinform* 2016;**17**:603–15.
120. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat* 2011;**32**:564–7.
121. Bradford YM, Toro S, Ramachandran S, *et al*. Zebrafish models of human disease: gaining insight into human disease at ZFIN. *ILAR J* 2017;**58**:4–16.
122. Zhong X, Peng J, Shen QS, *et al*. RhesusBase PopGateway: genome-wide population genetics atlas in rhesus macaque. *Mol Biol Evol* 2016;**33**:1370–5.
123. Freedman AH, Schweizer RM, Ortega-Del VD, *et al*. Demographically-based evaluation of genomic regions under selection in domestic dogs. *PLoS Genet* 2016;**12**: e1005851.
124. Darnell DK, Kaur S, Stanislaw S, *et al*. GEISHA: an in situ hybridization gene expression resource for the chicken embryo. *Cytogenet Genome Res* 2007;**117**:30–5.
125. Gramates LS, Marygold SJ, Santos GD, *et al*. FlyBase at 25: looking to the future. *Nucleic Acids Res* 2017;**45**:D663–71.
126. Howe KL, Bolt BJ, Cain S, *et al*. WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res* 2016;**44**:D774–80.
127. Lee R, Howe KL, Harris TW, *et al*. WormBase 2017: molting into a new stage. *Nucleic Acids Res* 2018;**46**: D869–74.
128. Chu YH, Hsieh MJ, Chiou HL, *et al*. MicroRNA gene polymorphisms and environmental factors increase patient susceptibility to hepatocellular carcinoma. *PLoS One* 2014;**9**:e89930.
129. Kawaguchi T, Koh Y, Ando M, *et al*. Prospective analysis of oncogenic driver mutations and environmental factors: Japan Molecular Epidemiology for Lung Cancer Study. *J Clin Oncol* 2016;**34**:2247–57.
130. Lage K, Greenway SC, Rosenfeld JA, *et al*. Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proc Natl Acad Sci U S A* 2012;**109**: 14035–40.
131. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. *Nat Rev Cardiol* 2015;**12**:627–42.
132. Turner SW, Ayres JG, Macfarlane TV, *et al*. A methodology to establish a database to study gene environment interactions for childhood asthma. *BMC Med Res Method* 2010;**10**:107.
133. Kitsios GD, Zintzaras E. Synopsis and data synthesis of genetic association studies in hypertension for the adrenergic receptor family genes: the CUMAGAS-HYPERT database. *Am J Hypertens* 2010;**23**:305–13.
134. Davis AP, Grondin CJ, Johnson RJ, *et al*. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2017;**45**:D972–8.
135. Jiang YZ, Liu YR, Xu XE, *et al*. Transcriptome analysis of triple-negative breast cancer reveals an integrated mRNA-lncRNA signature with predictive and prognostic value. *Cancer Res* 2016;**76**:2105–14.
136. Pandey R, Bhattacharya A, Bhardwaj V, *et al*. Alu-miRNA interactions modulate transcript isoform diversity in stress response and reveal signatures of positive selection. *Sci Rep* 2016;**6**:32348.
137. Ladeiro Y, Couchy G, Balabaud C, *et al*. MicroRNA profiling in hepatocellular tumors is associated with clinical features and oncogene/tumor suppressor gene mutations. *Hepatology* 2008;**47**:1955–63.
138. Lu L, Luo F, Liu Y, *et al*. Posttranscriptional silencing of the lncRNA MALAT1 by miR-217 inhibits the epithelial-mesenchymal transition via enhancer of zeste homolog 2 in the malignant transformation of HBE cells induced by cigarette smoke extract. *Toxicol Appl Pharmacol* 2015;**289**: 276–85.
139. Lin Z, Flemington EK. miRNAs in the pathogenesis of oncogenic human viruses. *Cancer Lett* 2011;**305**:186–99.
140. Barjaktarovic Z, Anastasov N, Azimzadeh O, *et al*. Integrative proteomic and microRNA analysis of primary human coronary artery endothelial cells exposed to low-dose gamma radiation. *Radiat Environ Biophys* 2013;**52**:87–98.
141. Pan HL, Wen ZS, Huang YC, *et al*. Down-regulation of microRNA-144 in air pollution-related lung cancer. *Sci Rep* 2015;**5**:14331.
142. Slattery ML, Herrick JS, Mullany LE, *et al*. Diet and lifestyle factors associated with miRNA expression in colorectal tissue. *Pharmgenomics Pers Med* 2017;**10**:1–16.
143. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res* 2002;**30**:412–5.
144. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem Substance and Compound databases. *Nucleic Acids Res* 2016;**44**: D1202–13.
145. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
146. Hecker N, Ahmed J, von Eichborn J, *et al*. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res* 2012;**40**:D1113–7.
147. Ahmed J, Meinel T, Dunkel M, *et al*. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res* 2011;**39**:D960–7.
148. Yang W, Soares J, Greninger P, *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;**41**: D955–61.
149. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
150. Hsin KY, Morgan HP, Shave SR, *et al*. EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities. *Nucleic Acids Res* 2011;**39**: D1042–8.
151. Preissner S, Kroll K, Dunkel M, *et al*. SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res* 2010;**38**:D237–43.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018

152. Li MX, Gui HS, Kwan JS, *et al*. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res* 2012;**40**:e53.
153. Li M, Li J, Li MJ, *et al*. Robust and rapid algorithms facilitate large-scale whole genome sequencing downstream analysis in an integrative framework. *Nucleic Acids Res* 2017; **45**:e75.
154. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 2012;**49**:433–6.
155. Carter H, Douville C, Stenson PD, *et al*. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;**14**(Suppl 3):S3.
156. Ioannidis NM, Rothstein JH, Pejaver V, *et al*. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**:877–85.
157. Schwarz JM, Rodelsperger C, Schuelke M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6.
158. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3.
159. Shihab HA, Rogers MF, Gough J, *et al*. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**: 1536–43.
160. Jagadeesh KA, Wenger AM, Berger MJ, *et al*. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**: 1581–6.
161. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* 2017;**100**:267–80.
162. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009;**19**:1553–61.
163. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**:e118.
164. Choi Y, Sims GE, Murphy S, *et al*. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012;**7**:e46688.
165. Shihab HA, Gough J, Cooper DN, *et al*. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;**34**:57–65.
166. Dong C, Wei P, Jian X, *et al*. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;**24**:2125–37.
167. Knecht C, Mort M, Junge O, *et al*. IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. *Nucleic Acids Res* 2017;**45**:e13.
168. Schwarz JM, Cooper DN, Schuelke M, *et al*. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014;**11**:361–2.
169. Wang J, Liao J, Zhang J, *et al*. ClinLabGeneticist: a tool for clinical management of genetic variants from whole exome sequencing in clinical genetic laboratories. *Genome Medicine* 2015;**7**:77.
170. Robinson PN, Kohler S, Bauer S, *et al*. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 2008;**83**:610–5.
171. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet* 2010;**77**:525–34.
172. Kohler S, Doelken SC, Mungall CJ, *et al*. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 2014;**42**: D966–74.
173. Groza T, Kohler S, Moldenhauer D, *et al*. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet* 2015;**97**:111–24.
174. Kohler S, Vasilevsky NA, Engelstad M, *et al*. The human phenotype ontology in 2017. *Nucleic Acids Res* 2017;**45**:D865–76.
175. Smith CL, Eppig JT. The mammalian phenotype ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome* 2012;**23**:653–68.
176. Kibbe WA, Arze C, Felix V, *et al*. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**:D1071–8.
177. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;**43**:D1049–56.
178. Malone J, Holloway E, Adamusiak T, *et al*. Modeling sample variables with an experimental factor ontology. *Bioinformatics* 2010;**26**:1112–8.
179. Robinson PN, Kohler S, Bauer S, *et al*. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 2008;**83**:610–5.
180. Sifrim A, Popovic D, Tranchevent LC, *et al*. eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 2013;**10**:1083–4.
181. Robinson PN, Kohler S, Oellrich A, *et al*. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 2014;**24**:340–8.
182. Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* 2014;**11**:935–7.
183. Kohler S, Schoeneberg U, Czeschik JC, *et al*. Clinical interpretation of CNVs with cross-species phenotype data. *J Med Genet* 2014;**51**:766–72.
184. Smedley D, Jacobsen JO, Jager M, *et al*. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc* 2015;**10**:2004–15.
185. Haendel MA, Vasilevsky N, Brush M, *et al*. Disease insights through cross-species phenotype comparisons. *Mamm Genome* 2015;**26**:548–55.
186. Robinson PN, Kohler S, Oellrich A, *et al*. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 2014;**24**:340–8.
187. Smedley D, Kohler S, Czeschik JC, *et al*. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* 2014;**30**: 3215–22.
188. Washington NL, Haendel MA, Mungall CJ, *et al*. Linking human diseases to animal models using ontology-based phenotype annotation. *PloS Biol* 2009;**7**:e1000247.
189. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
190. Bauer-Mehren A, Rautschka M, Sanz F, *et al*. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Genome Res* 2010;**26**: 2924–6.
191. Wang J, Al-Ouran R, Hu Y, *et al*. MARRVEL: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *Am J Hum Genet* 2017;**100**:843–53.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby071/5067517 by Sun Yat-Sen University user on 19 October 2018