BMC Genomics

SOFTWARE Open Access

Cruxome: a powerful tool for annotating, interpreting and reporting genetic variants

Qingmei Han[1†], Ying Yang[2†], Shengyang Wu[1], Yingchun Liao[1], Shuang Zhang[1], Hongbin Liang[1], David S. Cram[1*] and Yu Zhang[1*]

Abstract

Background: Next-generation sequencing (NGS) is an efficient tool used for identifying pathogenic variants that cause Mendelian disorders. However, the lack of bioinformatics training of researchers makes the interpretation of identified variants a challenge in terms of precision and efficiency. In addition, the non-standardized phenotypic description of human diseases also makes it difficult to establish an integrated analysis pathway for variant annotation and interpretation. Solutions to these bottlenecks are urgently needed.

Results: We develop a tool named "Cruxome" to automatically annotate and interpret single nucleotide variants (SNVs) and small insertions and deletions (InDels). Our approach greatly simplifies the current burdensome task of clinical geneticists and scientists to identify the causative pathogenic variants and build personal knowledge reference bases. The integrated architecture of Cruxome offers key advantages such as an interactive and user-friendly interface and the assimilation of electronic health records of the patient. By combining a natural language processing algorithm, Cruxome can efficiently process the clinical description of diseases to HPO standardized vocabularies. By using machine learning, in silico predictive algorithms, integrated multiple databases and supplementary tools, Cruxome can automatically process SNVs and InDels variants (trio-family or proband-only cases) and clinical diagnosis records, then annotate, score, identify and interpret pathogenic variants to finally generate a standardized clinical report following American College of Medical Genetics and Genomics/ Association for Molecular Pathology (ACMG/AMP) guidelines. Cruxome also provides supplementary tools to examine and visualize the genes or variations in historical cases, which can help to better understand the genetic basis of the disease.

Conclusions: Cruxome is an efficient tool for annotation and interpretation of variations and dramatically reduces the workload for clinical geneticists and researchers to interpret NGS results, simplifying their decision-making processes. We present an online version of Cruxome, which is freely available to academics and clinical researchers. The site is accessible at http://114.251.61.49:10024/cruxome/.

Keywords: Cruxome, Next Generation Sequencing, Mendelian disorders, Variant annotation, Variant interpretation, Whole Exome Sequencing, Natural language processing

* Correspondence: david.cram@berrygenomics.com; zhangyu001@berrygenomics.com
†Qingmei Han and Ying Yang contributed equally to this work.
[1]Berry Genomics Company Limited, Building 5, Courtyard 4, Shengmingyuan Road, ZGC Life Science Park, Changping District, 102200 Beijing, China
Full list of author information is available at the end of the article

Background

Genetic diseases that follow an autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive or mitochondrial pattern of inheritance are known as Mendelian disorders. [1–3]. Currently, in the order of 7,000–9,600 Mendelian disorders have been recorded by Global Genes (https://globalgenes.org/), Online Mendelian Inheritance in Man (OMIM, https://omim.org/) and Orphadata (http://www.orphadata.org/) databases and approximately 300 new Mendelian phenotypes are updated each year [4]. Of all the Mendelian disorders, approximately 80 % now have a defined genetic cause [5, 6] whereas for the remaining 20 %, the genes and genetic lesions remain unknown [7–9]. Thus, clinical research is ongoing to fully characterize the causative genes, develop a better understanding of the underlying disease mechanisms and, explore potential treatment options [10].

Next-generation sequencing (NGS) has emerged as an innovative tool for medical genetics, and has led to a paradigm shift in medical research and clinical practice [11–13]. With the decreasing cost of sequencing, methods such as whole exome sequencing (WES) have become affordable and are widely used for the diagnosis of Mendelian disorders, with typical positive diagnostic yields of 25–40 % [14, 15]. With the fast development of different NGS techniques, the gap between data yield, quality and gene coverage between platforms is rapidly closing. The challenge now is the ability to systematically analyze the hundreds of thousands of high-quality variant calls (including single nucleotide variants, SNVs, short insertions or deletions, InDels and large copy number variants, CNVs) that are revealed in WES sequencing files [16–19]. Even after rigorous filtering, there are still tens to hundreds of candidate causal variants to be considered [19–22]. Thus, an important step is to choose the appropriate analysis tools to efficiently and precisely mine the causative variants, especially when the analysis team lacks training in the use of sophisticated bioinformatic programs. In addition, secondary confirmatory analyses are also required for verification or support when candidates of causative variation are related to the phenotype.

Several open-source analysis tools for variant annotation and functional effect prediction have been reported including spliceAI [20], ANNOVAR [21], SnpEff [23], PolyPhen-2 [24], CADD [25] and InterVar [26]. For example, SpliceAI is a deep learning-based tool specifically designed to identify splice variants. Combined Annotation Dependent Depletion (CADD) is used to score the deleteriousness of SNV as well as InDel variants in the human genome. Alternatively, InterVar can be used for clinical interpretation of genetic variants using the ACMG/AMP 2015 guidelines [26]. However, almost all of these tools are command line tools that have an unfriendly user interface and require a strong background in bioinformatics to comprehensively analyze the data.

When a set of candidate variants are identified, the aim of follow-up analysis is to establish a strong relationship between the candidate genes and known diseases by using information in the published literature and databases. However, this information is sometimes incomplete or fragmented and distributed differently across many databases, which makes this step very time-consuming and inefficient. There are several reported tools that integrate the various databases and simplify the search. These tools include IPAD (integrated pathway analysis database for systematic enrichment analysis) [27], SIDD (semantically integrated database towards a global view of human disease) [28], VariED (integrated database of gene annotation and expression profiles for variants related to human diseases) [29], DisGeNET (integrated information on human disease-associated genes and variants) [30] and Human Disease Insight (integrated knowledge-based platform for disease-gene-drug information) [31]. However, while useful, these tools only focus on specific applications. Thus, comprehensive integration of different databases for relevant knowledge is urgently needed to increase the yield of positive diagnoses.

Electronic health records (EHRs) have been widely implemented by clinical geneticists and include the patient's information such as name, age, gender, laboratory test results, phenotypic description, diagnosis and medication details. Almost all tools or databases adopt Human Phenotype Ontology (HPO) as the reference. HPO uses standardized vocabulary for describing phenotypic abnormalities in human disease, drawing on over 13,000 terms and over 156,000 annotations to hereditary diseases (https://hpo.jax.org/) [29, 32]. For clinical geneticists, it is almost impossible to accurately describe all of patient's phenotype using standard terms, and often the diagnosis records are more colloquial and not directly computationally useful [32, 33]. Benefiting from the development of big data techniques, large-scale EHR data mining has become widely used in data-driven medical studies, clinical decision making, and health management [34–36]. Since the phenotypic description of patients is a critical factor for precise variant interpretation, it is urgent to develop new algorithms to efficiently and accurately transform colloquial descriptions to more standardized vocabulary.

Based on these challenges, we develop Cruxome, an automated and user-friendly tool for variant interpretation which is designed to efficiently and precisely handle the Variant Call Format (VCF) file (either from WES or gene panel data) and generate standardized clinical reports. By mining the hundreds of thousands of literature accounts and integrating appropriate databases, Cruxome harbors a comprehensive and regularly updated biomedical knowledge base to keep pace with precise variant interpretation. Cruxome uses a natural language processing algorithm (NER) to transform colloquial descriptions of phenotype to standard HPO vocabulary. Cruxome also supports building a personal knowledge base to efficiently manage patient's information and interpret results with traceable evidence record of interpretation decisions. Above all, Cruxome provides an overall solution for variant interpretation, dramatically reducing workload and facilitating better decision-making processes.

Implementation

Construction of Cruxome and main features

Cruxome was designed with a user-friendly interface and developed based on a Browser/Server style to facilitate easy access and to minimize incompatibility with different computer operating systems. Cruxome runs in Docker mode (https://www.docker.com/) which is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. Thus, Cruxome can easily be deployed on either a cloud server (for example Amazon Web Services, Microsoft Azure) or on a local server. To enhance the functionality of Cruxome, improve efficiency and simplify code maintenance, a layered pattern was used in the basic architecture of Cruxome (Fig. 1). Cruxome consists of six sublayers: a user interface layer (UIL), a model layer (ML), a controller layer (CL), a support layer (SL), a data exchange layer (DEL) and a data storage layer (DSL). UIL, ML and CL provide the interactive and data presentations to users; SL provides support to CL; DEL provides compatibility to various database types and a connection to laboratory information management systems (LIMS) and other software and DSL is responsible for read/write data from database (MySQL as default, https://www.mysql.com/) and for storage of the information.

Minimum requirements of Cruxome (available on all modern computers):

- A modern browser (Chrome, FireFox, Safari or Edge).
- A 24-core server with 64G memory, 1T hard disk.
- An internet or intranet connection of 10Mbit.

Results

Cruxome pipeline

The overall workflow of Cruxome is shown in Fig. 2. The workflow of Cruxome commences with uploading of VCF files listing the genetic variants identified from gene panel or WES data and, uploading of the phenotypic records of each patient. Next, Cruxome performs variant annotation, phenotype processing and interpretation, and then generates a standardized report summarizing the candidate genetic variants, and provides conclusions and

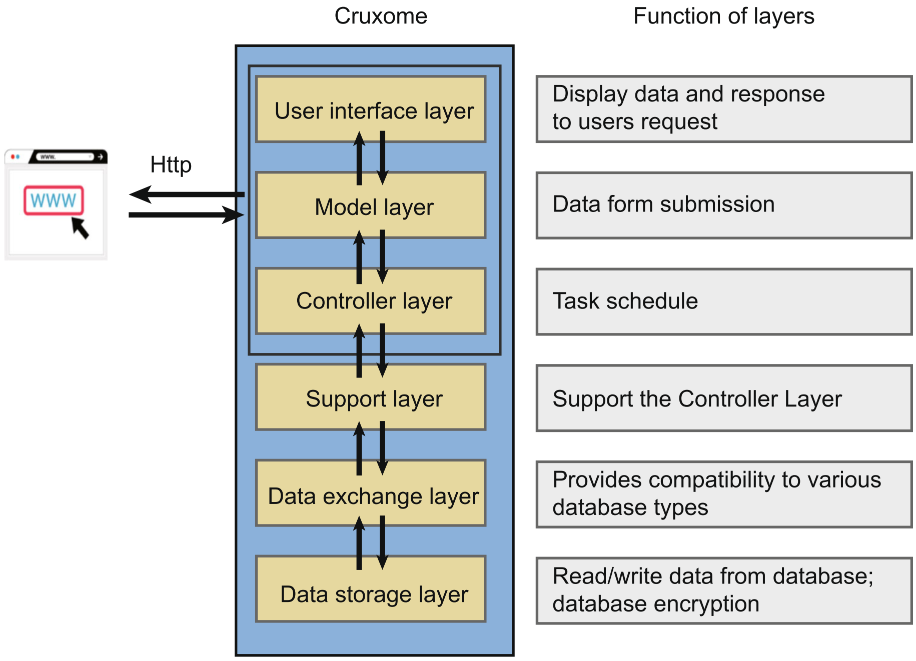

Cruxome
Function of layers
Http
www
User interface layer
Model layer
Controller layer
Support layer
Data exchange layer
Data storage layer
Display data and response to users request
Data form submission
Task schedule
Support the Controller Layer
Provides compatibility to various database types
Read/write data from database; database encryption

Fig. 1 Architecture of Cruxome. The six interactive layers of Cruxome and their function are shown. Users can access Cruxome via a modern browser. The User Interface Layer, Model Layer and Controller Layer are responsible for interactive presentation to users; Data Exchange Layer provides compatibility to various database types and software, and the Data Storage Layer is responsible for data read/write operation and data security

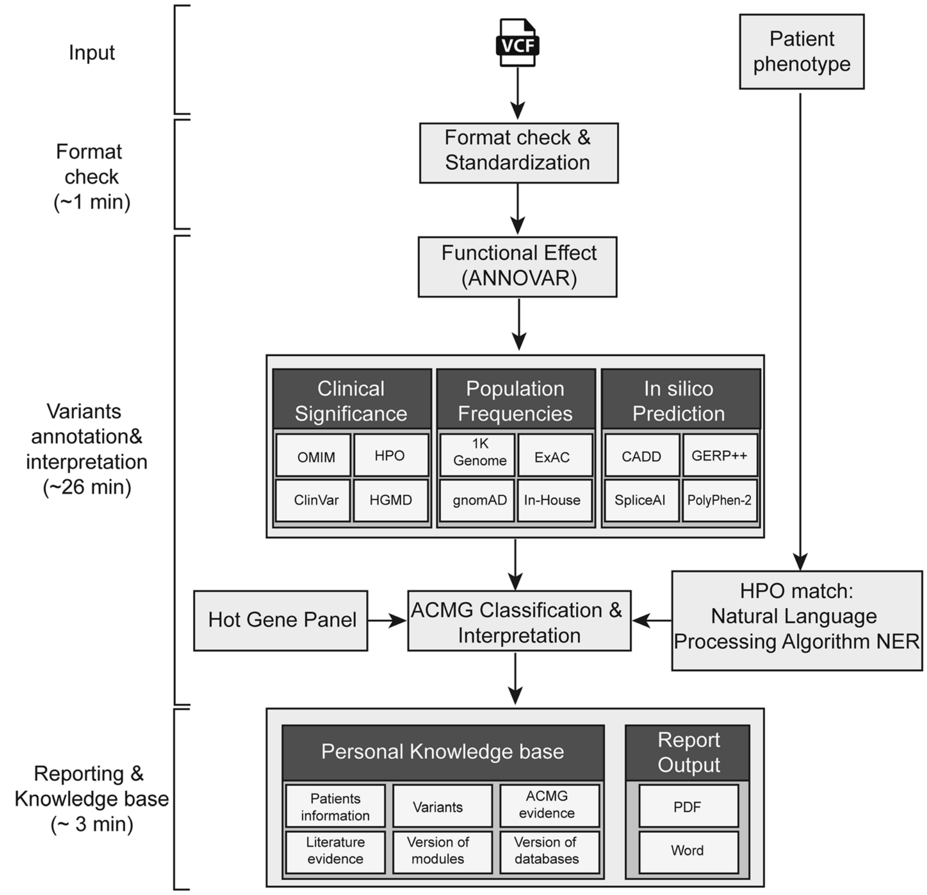


Fig. 2 Schematic representation of Cruxome workflow. Typical Cruxome workflow contains four steps (input, format check, variants annotation and interpretation, and reporting and knowledge base), and time consumption to perform each step is indicated. After uploading a VCF file, Cruxome executes the VCF file format check in "Format check" step. Sequential "Variants annotation and interpretation" step is then launched to annotate and interpret the variants using various tools and databases. By using Natural Language Processing Algorithm NER, the hot gene panel and multiple databases, Cruxome performs integrated analyses that interpret and score variants according to ACMG guidelines. The report of the analysis is then exported in a PDF or Word format, and the personal knowledge base is automatically updated to store all the interpretation information. From sample input to report, the whole process takes approximately 30 min. OMIM: Online Mendelian Inheritance in Man, a comprehensive, authoritative compendium of human genes and genetic phenotypes; HPO: Human Phenotype Ontology, provides standardized vocabulary of phenotypic abnormalities encountered in human disease; ClinVar: a publicly available database that aggregates information about sequence variation and its relationship to human health; HGMD: Human Gene Mutation Database, represents all known (published) gene lesions responsible for human inherited disease; 1 K Genome: data from 1000 Genomes Project; ExAC: Exome Aggregation Consortium; gnomeAD: Genome Aggregation Database; In-House: genomic database of Berry Genomics Co., Ltd.; CADD: Combined Annotation Dependent Depletion, integrates multiple annotations into one metric; GERP++: Genomic Evolutionary Rate Profiling; SpliceAI: A deep learning-based tool to identify splice variants; PolyPhen-2: predicts possible impact of an amino acid substitution on the structure and function of a human protein

relevant references (PDF or Word format). A user manual file for step by step instruction of how to use the Cruxome software is available for download on the Cruxome website.

Typical application scenario

After login to Cruxome, the user first creates the patient's record with detailed information about phenotype, age, family relationship or directly imports the records from existing patient databases. Secondly, the VCF files are uploaded. Cruxome supports all of the VCF file formats that meet VCF 4.2 standard or higher, and supports both the GRCh37 (hg19) and GRCh38 reference genomes. Cruxome then automatically checks file formats and standardizes the files.

After checking the patient information and VCF file format, Cruxome then launches its annotation module. For the most comprehensive evaluation and interpretation of variants, Cruxome integrates multiple databases, including sequence databases for gene functional information, population databases for calculation of variants allele frequencies and disease databases to define clinical significance and phenotype relationships relevant to disease phenotype. Multiple tools are then applied to evaluate the effect of variants on protein function and to finally generate a variant score [37].

Next, Cruxome uses a natural language processing algorithm NER to transform clinical diagnosis records to HPO standard format (Fig. 2). By using our newly developed algorithm, Cruxome automatically performs the variant interpretation and clinical classification by combining the phenotypic diagnosis description and the hot gene panel according to American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines [38]. At the end of the process, Cruxome generates outputs of variant interpretation results with corresponding evidence ordered by pathogenic criterion. Users can further review the interpreted variants, combine more clinical information if required and then generate a clinical report summarizing the relevant genetic variants, conclusions and references (PDF or Word format) (Fig. 2).

Management of your own knowledge base

Once variant interpretation is complete, Cruxome automatically updates your personal knowledge base and stores all the information generated during the interpretation process, including candidate variants, ACMG evidences, literature and versions of modules and databases, thus making the interpretation decisions traceable (Fig. 2). Personal knowledge base dramatically facilitates data tracking, data management and re-interpretation variants using updated databases. Users can also manually modify or update literature records in their own knowledge base, including the clinical level of variants or other fields. When the same variants are again found following the analysis of new samples, the variants are automatically highlighted showing the information from previous records and thus provides users greater confidence with the case at hand.

User case demonstration

A representative proband-only case is presented to demonstrate the functionality of the Cruxome pipeline (Fig. 3). The clinical diagnosis of the six-month-old proband was “decreased fetal movement in the prenatal period and increased head circumference (45.7 cm), global developmental delay, periventricular leukomalacia, hip dysplasia, motor deterioration and impaired pursuit initiation and maintenance post birth”. After login to Cruxome, the home page is loaded (Fig. 3A). The left panel of home page shows the modules of Cruxome whereas the right panel shows the list of patient records. After clicking the “Add” button in Sample Management module, patient’s information such as name, gender, age, clinical phenotype needs to be entered into the pop-up window (Fig. 3B). The VCF file is then uploaded (click “import” button), and Cruxome automatically performs variant interpretation. The progress of VCF uploading, analyzing and interpretation can be visualized in real time by the progress bar on the home page (Fig. 3A). The final interpretation results can be accessed in the “Sample Interpretation” module (Fig. 3C). Supporting information about variants or interpretation can be examined or reviewed by clicking the corresponding button. If candidate pathogenic gene variants are found (*AHDC1* gene in this case), users should mark the corresponding variants as “Positive” in the conclusion column (Fig. 3C). By clicking the “Generate Report” icon in upper-right of interpretation results (Fig. 3B), the generate report page will be loaded (Fig. 3D). By choosing the “Positive conclusion” or “Negative conclusion” selection box, variants, references and an automated conclusion will be displayed in corresponding section (Fig. 3D). In this example case, Cruxome successfully identified a pathogenic variant (NM_001029882.3: c.2773 C > T: p.R925*) in the *AHDC1* gene, which has been reported to be responsible for autosomal dominant Xia-Gibbs syndrome [39]. Users can simply export a standardized clinical report by clicking the “Generate Report” button below (Fig. 3D). The new report can be accessed in Report Management module.

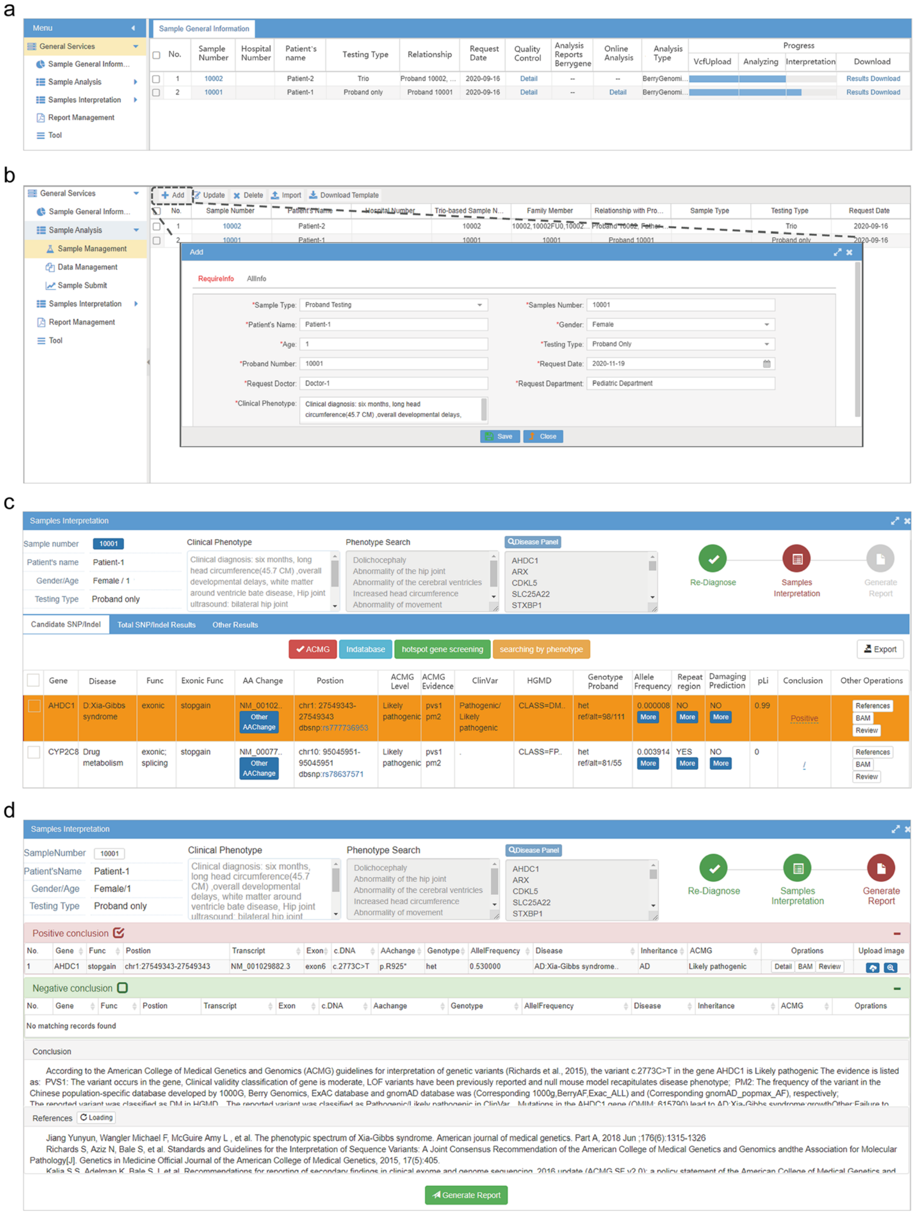
In another representative trio-family case, the clinical diagnosis of the proband was hyperhomocystinemia, methylmalonic acidemia, anemia, megaloblastic anemia, proteinuria, occult blood, feeding difficulties. Cruxome successfully identified a likely pathogenic (NM_015506.2: c.80 A > G: p.Q27R) and a pathogenic (NM_015506.2: c.217 C > T: p.R73X) variant in the *MMACHC* gene, which is responsible for methylmalonic aciduria and homocystinuria [40] (Supplemental Table 1). The proband was a compound heterozygote for variants p.Q27R and p.R73X whereas the father and mother were confirmed to be heterozygous for the respective variants.

Extra Tools

Cruxome also provides other useful tools to help clinical geneticists visualize their data. First, the “getting sequence” tool can display DNA sequence of a given region (Fig. 4A). Second, the “examine bam file” tool can be used to schematically display NGS reads that aligned on the reference genome (Fig. 4B). Third, the “locus searching” tool can be used to calculate frequency of variants in all samples in the personal knowledge base (Fig. 4C). Lastly, the “gene coverage and depth” tool can search coverage, depth and the number of variants of a given gene in all samples (Fig. 4D).

Update and version options

Cruxome is frequently updated to incorporate the latest clinical genetic research findings with options for adding new algorithms and new annotation sources and analysis modules. Benefitting from version updates, Cruxome



□	No.	Sample Number	Hospital Number	Patient's name	Testing Type	Relationship	Request Date	Quality Control	Analysis Reports Berrygene	Online Analysis	Analysis Type	Progress VcfUpload	Analyzing	Interpretation	Download
□	1	10002		Patient-2	Trio	Proband 10002, ...	2020-09-16	Detail	--	--	BerryGenomi...				Results Download
□	2	10001		Patient-1	Proband only	Proband 10001	2020-09-16	Detail	--	Detail	BerryGenomi...				Results Download

□	No.	Sample Number	Patient's Name	Hospital Number	Trio-based Sample N...	Family Member	Relationship with Pro...	Sample Type	Testing Type	Request Date
□	1	10002	Patient-2		10002	10002,10002FU0,10002...	Proband 10002, Father...		Trio	2020-09-16
□	2	10001	Patient-1		10001	10001	Proband 10001		Proband only	2020-09-16

*Sample Type:	Proband Testing	*Samples Number:	10001
*Patient's Name:	Patient-1	*Gender:	Female
*Age:	1	*Testing Type:	Proband Only
*Proband Number:	10001	*Request Date:	2020-11-19
*Request Doctor:	Doctor-1	*Request Department:	Pediatric Department
*Clinical Phenotype:	Clinical diagnosis: six months, long head circumference(45.7 CM) ,overall developmental delays,		

Sample number	10001
Patient's name	Patient-1
Gender/Age	Female / 1
Testing Type	Proband only

Clinical Phenotype

Clinical diagnosis: six months, long head circumference(45.7 CM) ,overall developmental delays, white matter around ventricle bate disease, Hip joint ultrasound: bilateral hip joint

Phenotype Search

Dolichocephaly
Abnormality of the hip joint
Abnormality of the cerebral ventricles
Increased head circumference
Abnormality of movement

AHDC1
ARX
CDKL5
SLC25A22
STXBP1

□	Gene	Disease	Func	Exonic Func	AA Change	Postion	ACMG Level	ACMG Evidence	ClinVar	HGMD	Genotype Proband	Allele Frequency	Repeat region	Damaging Prediction	pLi	Conclusion	Other Operations
□	AHDC1	D:Xia-Gibbs syndrome	exonic	stopgain	NM_00102.. Other AAChange	chr1: 27549343-27549343 dbsnp:rs777736953	Likely pathogenic	pvs1 pm2	Pathogenic/ Likely pathogenic	CLASS=DM..	het ref/alt=98/111	0.000008 More	NO More	NO More	0.99	Positive	References BAM Review
□	CYP2C8	Drug metabolism	exonic; splicing	stopgain	NM_00077.. Other AAChange	chr10: 95045951-95045951 dbsnp:rs78637571	Likely pathogenic	pvs1 pm2	.	CLASS=FP..	het ref/alt=81/55	0.003914 More	YES More	NO More	0	/	References BAM Review

SampleNumber	10001
Patient'sName	Patient-1
Gender/Age	Female/1
Testing Type	Proband only

Clinical Phenotype

Clinical diagnosis: six months, long head circumference(45.7 CM) ,overall developmental delays, white matter around ventricle bate disease, Hip joint ultrasound: bilateral hip joint

Phenotype Search

Dolichocephaly
Abnormality of the hip joint
Abnormality of the cerebral ventricles
Increased head circumference
Abnormality of movement

AHDC1
ARX
CDKL5
SLC25A22
STXBP1

Positive conclusion ☑

No.	Gene	Func	Postion	Transcript	Exon	c.DNA	AAchange	Genotype	AllelFrequency	Disease	Inheritance	ACMG	Oprations	Upload image
1	AHDC1	stopgain	chr1:27549343-27549343	NM_001029882.3	exon6	c.2773C>T	p.R925*	het	0.530000	AD:Xia-Gibbs syndrome..	AD	Likely pathogenic	Detail BAM Review	

Negative conclusion □

No.	Gene	Func	Postion	Transcript	Exon	c.DNA	Aachange	Genotype	AllelFrequency	Disease	Inheritance	ACMG	Oprations
No matching records found													

Conclusion

According to the American College of Medical Genetics and Genomics (ACMG) guidelines for interpretation of genetic variants (Richards et al., 2015), the variant c.2773C>T in the gene AHDC1 is Likely pathogenic The evidence is listed as: PVS1: The variant occurs in the gene, Clinical validity classification of gene is moderate, LOF variants have been previously reported and null mouse model recapitulates disease phenotype; PM2: The frequency of the variant in the Chinese population-specific database developed by 1000G, Berry Genomics, ExAC database and gnomAD database was (Corresponding 1000g,BerryAF,Exac_ALL) and (Corresponding gnomAD_popmax_AF), respectively;
The reported variant was classified as DM in HGMD . The reported variant was classified as Pathogenic/Likely pathogenic in ClinVar . Mutations in the AHDC1 gene (OMIM: 615790) lead to AD:Xia-Gibbs syndrome:growthOther:Failure to

References Loading

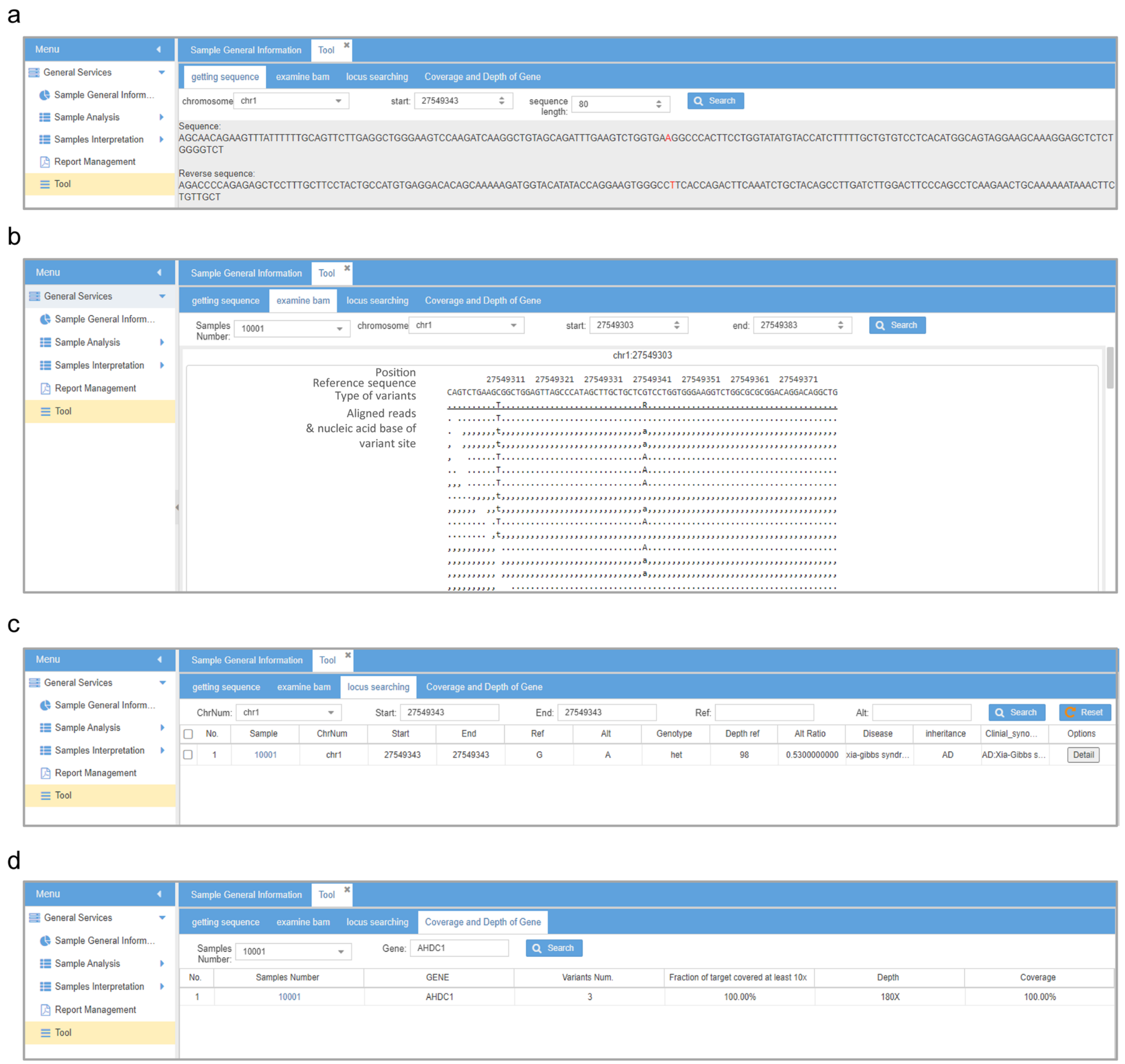
Jiang Yunyun, Wangler Michael F, McGuire Amy L , et al. The phenotypic spectrum of Xia-Gibbs syndrome. American journal of medical genetics. Part A, 2018 Jun ;176(6):1315-1326
Richards S, Aziz N, Bale S, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics andthe Association for Molecular Pathology[J]. Genetics in Medicine Official Journal of the American College of Medical Genetics, 2015, 17(5):405.
Kalia S S, Adelman K, Bale S J, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and

Fig. 3 Interface of Cruxome. **A**. Home page of Cruxome, which shows a module list (left panel) and overview of all the information of samples under user's account (right panel). Users enter the submodule by clicking the corresponding text in the left panel. **B**. To perform an interpretation, click "Add" button in "Sample Management" module, and input the patient's information in the pop-out window. **C**. After Cruxome finishes interpretation process, a detailed list of variants with annotation and ACMG classification is produced. Users can review the interpretation of variants and examine the literature and bam file by clicking corresponding button. Candidate variants could be marked as "Positive" or "Negative" by clicking "/" in conclusion column. **D**. After entering the "Generate Report" page, users can easily export a clinically standardized report with the inclusion of all supported knowledges by choosing the "Positive conclusion" or "Negative conclusion" checkbox. Report can be found in "Report Management" module

users can easily re-analyze cases stored in the knowledge base, and potentially identify novel pathogenic variants.

Comparison of Cruxome with other software

A range of commercial software has been reported to perform variant annotation and interpretation [26, 41–43]. Compared with above mentioned software, Cruxome offers unique advantages (Table 1). Firstly, it facilitates (i) transformation of colloquial description of phenotype to standard HPO vocabulary using a natural language processing algorithm (NER), (ii) automatic variant annotation and interpretation which greatly reduces the workload of users and (iii) export of a standard clinical report summarizing the relevant genetic variants, conclusions and references. However, in the current version of Cruxome, only variants from WES and panel data are supported, and file



chromosome: chr1 start: 27549343 sequence length: 80

Sequence:
AGCAACAGAAGTTTATTTTTTGCAGTTCTTGAGGCTGGGAAGTCCAAGATCAAGGCTGTAGCAGATTTGAAGTCTGGTGAAGGCCCACTTCCTGGTATATGTACCATCTTTTTGCTGTGTCCTCACATGGCAGTAGGAAGCAAAGGAGCTCTCTGGGGTCT

Reverse sequence:
AGACCCCAGAGAGCTCCTTTGCTTCCTACTGCCATGTGAGGACACAGCAAAAAGATGGTACATATACCAGGAAGTGGGCCTTCACCAGACTTCAAATCTGCTACAGCCTTGATCTTGGACTTCCCAGCCTCAAGAACTGCAAAAAATAAACTTCTGTTGCT

Samples Number: 10001 chromosome: chr1 start: 27549303 end: 27549383

chr1:27549303

```
                          Position         27549311  27549321  27549331  27549341  27549351  27549361  27549371
                Reference sequence  CAGTCTGAAGCGGCTGGAGTTAGCCCATAGCTTGCTGCTCGTCCTGGTGGGAAGGTCTGGCGCGCGGACAGGACAGGCTG
                  Type of variants  ..........T.............................R........................................
                                    . ........T......................................................................
                     Aligned reads  .  ,,,,,,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
           & nucleic acid base of   ,  ,,,,,,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                      variant site  ,    .....T.............................A........................................
                                    ..   .....T.............................A........................................
                                    ,,, ......T.............................A........................................
                                    ....,,,,,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                    ,,,,,,  ,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                    ........ .T.............................A........................................
                                    ........ ,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                    ,,,,,,,,,, .............................A........................................
                                    ,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                    ,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                    ,,,,,,,,,,   ....................................................................
```

ChrNum: chr1 Start: 27549343 End: 27549343 Ref: Alt:

□	No.	Sample	ChrNum	Start	End	Ref	Alt	Genotype	Depth ref	Alt Ratio	Disease	inheritance	Clinial_syno...	Options
□	1	10001	chr1	27549343	27549343	G	A	het	98	0.5300000000	xia-gibbs syndr...	AD	AD:Xia-Gibbs s...	Detail

Samples Number: 10001 Gene: AHDC1

No.	Samples Number	GENE	Variants Num.	Fraction of target covered at least 10x	Depth	Coverage
1	10001	AHDC1	3	100.00%	180X	100.00%

Fig. 4 Extra tools Interface of Cruxome. **A**. "Getting sequence" tools: displays flanking sequence of a given site in the reference genome. Nucleotides marked with red indicates the query position. **B**. "Examine bam file" tool: displays reads aligned to the reference genome. The position, reference sequence and the type of variants are shown in the above three rows. Altered nucleotides in aligned reads are shown; nucleotides in reads with no change compared with the reference sequence are indicated as "," or ".". **C**. "Locus search" tool: calculates the frequency of variants in a given region in all samples in the personal knowledge base. **D**. The "Gene coverage and depth" tool: examines the number of variants, fraction of targets covered, coverage, depth and coverage of a given gene in all samples under the users account

format of variants is restricted to standard VCF format. This limitation prevents its usages in annotating and interpretating variants from whole genome sequencing (WGS), and reduces flexibility of input files. Accordingly, further development of Cruxome is planned to include modules for annotation and interpretation of WGS variants, and modules that accept various files that contain structured records of variants (e.g. Excel or txt format) as input.

Conclusions

By using in-house algorithms and multiple databases, Cruxome can effectively perform variant annotation and interpretation. A user-friendly interface combined with a natural language processing algorithm NER makes Cruxome easy-to-use and importantly, users do not need to change their phenotype descriptions that they write in clinical diagnosis records. Although Cruxome is designed

Table 1 Functional comparison of different software

Software	Cruxome	Seqmax	QIAGEN	TGex	Seave	InterVar
Input file	VCF	Fastq	VCF	VCF	VCF	VCF
Variants	SNV, Indel	SNV, Indel	SNV, Indel	SNV	SNV, Indel, CNV, SV	SNV, Indel
Run mode	Automatic	Manual	Automatic	Manual	-	Automatic
Supports Chinese phenotype search	YES	YES	NO	YES	NO	NO
Phenotypic semantic analysis	YES	NO	NO	NO	NO	NO
Report	Variants and clinical interpretation	Variants	Variants	Variants	NO	NO
Database build	YES	NO	NO	NO	NO	NO

for users with less bioinformatics knowledge, others with a more solid grounding in bioinformatics can also use Cruxome in a more convenient and time-saving way. These features make Cruxome more versatile for use by clinical geneticists and can also provide important information to genetic counselors to discuss the results with patients. Above all, Cruxome is a powerful solution for annotating and interpreting variants and for managing personal knowledge bases and, overcomes the current bottleneck of clinical geneticists spending valuable time mining and evaluating causative variants.

Availability and requirements

Project name: Cruxome.

Project home page: http://114.251.61.49:10024/cruxome/.

Operating system(s): Platform independent.

Programming language: Java.

Other requirements: Java (version > = 1.8.1), Tomcat (version > = 8.0), Docker (version > = 18.03.1-ce), MySQL (version > = 5.7).

License: Free for academic and research use.

Abbreviations

NGS: next-generation sequencing; SNVs: single nucleotide variants; InDels: insertions or deletions; CNVs: copy number variants; HPO: human phenotype ontology; ACMG/AMP: American College of Medical Genetics and Genomics/ Association for Molecular Pathology; WES: whole exome sequencing; EHR: electronic health records; NER: natural language processing algorithm; VCF: variant call format; MAF: minor allele frequency

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07728-6.

Additional file 1.

Acknowledgements

Not applicable.

Authors' contributions

YZ designed and supervised the development of Cruxome. QH, YL and SZ developed the software. YY provided WES data and thoroughly tested Cruxome for clinical use. QH, SW and HL drafted the manuscript. DC and YZ were major contributors in writing and revising the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by an Innovation Capability Support Plan of Shaanxi province (Grant number 2019KJXX-055). The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The example used in this paper (Fig. 3 B and D, Supplemental Table 1) is available in the free trial account as a demonstration case.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors (exception Ying Yang) are employees of Berry Genomics Co., Ltd. None of the authors hold stocks or bonds.

Author details

[1]Berry Genomics Company Limited, Building 5, Courtyard 4, Shengmingyuan Road, ZGC Life Science Park, Changping District, 102200 Beijing, China. [2]Xian Children's Hospital, 710003 Xian, China.

Received: 22 January 2021 Accepted: 20 May 2021
Published online: 03 June 2021

References

1. Kennedy MA. Mendelian Genetic Disorders. eLS. 2005. https://doi.org/10.1038/npg.els.0003934.
2. Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nat Rev Genet. 2006;7(4):277–82. https://doi.org/10.1038/nrg1826.
3. Chakravarti A. Genomic contributions to Mendelian disease. Genome Res. 2011;21(5):643–4. https://doi.org/10.1101/gr.123554.111.
4. Hartley T, Balci TB, Rojas SK, Eaton A, Canada CR, Dyment DA, et al. The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM(R). Am J Med Genet C Semin Med Genet. 2018;178(4):458–63. https://doi.org/10.1002/ajmg.c.31662.
5. Field MJ, Boat TF, editors. Rare Diseases and Orphan Products: Accelerating Research and Development. Washington (DC): National Academies Press (US); 2010. https://doi.org/10.17226/12953.
6. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19(5):253–68. https://doi.org/10.1038/nrg.2017.116.
7. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305–14. https://doi.org/10.1016/S0140-6736(14)61705-0.
8. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433–8. https://doi.org/10.1038/nature21062.

9. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42(1):30–5. https://doi.org/10.1038/ng.499.
10. Oti M, Brunner HG. The modular nature of genetic diseases. Clin Genet. 2007;71(1):1–11. https://doi.org/10.1111/j.1399-0004.2006.00708.x.
11. Kaname T, Yanagi K, Naritomi K. A commentary on the promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59(3):117–8. https://doi.org/10.1038/jhg.2014.7.
12. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59(1):5–15. https://doi.org/10.1038/jhg.2013.114.
13. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008; 26(10):1135–45. https://doi.org/10.1038/nbt1486.
14. Dragojlovic N, Elliott AM, Adam S, van Karnebeek C, Lehman A, Mwenifumbo JC, et al. The cost and diagnostic yield of exome sequencing for children with suspected genetic disorders: a benchmarking study. Genet Med. 2018;20(9):1013–21. https://doi.org/10.1038/gim.2017.226.
15. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Koster J, Marais A, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. Eur J Hum Genet. 2017;25(2):176–82. https://doi.org/10.1038/ejhg.2016.146.
16. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875. https://doi.org/10.1038/srep17875.
17. Liu M, Zhong Y, Liu H, Liang D, Liu E, Zhang Y, et al. REDBot: Natural language process methods for clinical copy number variation reporting in prenatal and products of conception diagnosis. Mol Genet Genomic Med. 2020;8(11):e1488. https://doi.org/10.1002/mgg3.1488.
18. Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep. 2019;9(1):9345. https://doi.org/10.1038/s41598-019-45835-3.
19. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823–8. https://doi.org/10.1126/science.1215040.
20. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535–48. https://doi.org/10.1016/j.cell.2018.12.015. e24.
21. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. https://doi.org/10.1093/nar/gkq603.
22. Smedley D, Jacobsen JO, Jager M, Kohler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10(12):2004–15. https://doi.org/10.1038/nprot.2015.124.
23. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92. https://doi.org/10.4161/fly.19695.
24. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;76(1):7.20.21-27.20.41. https://doi.org/10.1002/0471142905.hg0720s76.
25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94. https://doi.org/10.1093/nar/gky1016.
26. Li Q, Wang K, InterVar. Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017;100(2):267–80. https://doi.org/10.1016/j.ajhg.2017.01.004.
27. Zhang F, Drabier R. IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. BMC Bioinformatics. 2012;13(15):S7. https://doi.org/10.1186/1471-2105-13-S15-S7.
28. Cheng L, Wang G, Li J, Zhang T, Xu P, Wang Y. SIDD: a semantically integrated database towards a global view of human disease. PLoS One. 2013;8(10):e75504. https://doi.org/10.1371/journal.pone.0075504.
29. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018-D27. https://doi.org/10.1093/nar/gky1105.
30. Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845-D55. https://doi.org/10.1093/nar/gkz1021.
31. Tasleem M, Ishrat R, Islam A, Ahmad F, Hassan MI. Human Disease Insight: An integrated knowledge-based platform for disease-gene-drug information. J Infect Public Health. 2016;9(3):331–8. https://doi.org/10.1016/j.jiph.2015.10.018.
32. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83(5):610–5. https://doi.org/10.1016/j.ajhg.2008.09.017.
33. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):D966-74. https://doi.org/10.1093/nar/gkt1026.
34. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc. 2014; 21(5):808–14. https://doi.org/10.1136/amiajnl-2013-002381.
35. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. AMIA Annu Symp Proc. 2015; 2015:1326-33.
36. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017;33(14):i37–48. https://doi.org/10.1093/bioinformatics/btx228.
37. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393.
38. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(5):405–24. https://doi.org/10.1038/gim.2015.30.
39. Jiang Y, Wangler MF, McGuire AL, Lupski JR, Posey JE, Khayat MM, et al. The phenotypic spectrum of Xia-Gibbs syndrome. Am J Med Genet A. 2018; 176(6):1315–26. https://doi.org/10.1002/ajmg.a.38699.
40. Liu MY, Yang YL, Chang YC, Chiang SH, Lin SP, Han LS, et al. Mutation spectrum of MMACHC in Chinese patients with combined methylmalonic aciduria and homocystinuria. J Hum Genet. 2010;55(9):621–6. https://doi.org/10.1038/jhg.2010.81.
41. Dahary D, Golan Y, Mazor Y, Zelig O, Barshir R, Twik M, et al. Genome analysis and knowledge-driven variant interpretation with TGex. BMC Med Genomics. 2019;12(1):200. https://doi.org/10.1186/s12920-019-0647-8.
42. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: From raw data to diagnosis with lifetime value. Clin Genet. 2018;93(3):508–19. https://doi.org/10.1111/cge.13190.
43. Hintzsche JD, Robinson WA, Tan AC. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. Int J Genomics. 2016; 2016:7983236. https://doi.org/10.1155/2016/7983236.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.