AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis

Veronique Geoffroy, Thomas Guignard, Arnaud Kress, Jean-Baptiste Gaillard, Tor Solli-Nowlan, Audrey Schalk, Vincent Gatinois, Helene Dollfus, Sophie Scheidecker, Jean Muller

Published online 22 May 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W21–W28***
https://doi.org/10.1093/nar/gkab402

AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis

Véronique Geoffroy[1,*,†], Thomas Guignard[2,†], Arnaud Kress[3], Jean-Baptiste Gaillard[2], Tor Solli-Nowlan[4], Audrey Schalk[5], Vincent Gatinois[2], Hélène Dollfus[1,6], Sophie Scheidecker[1,5] and Jean Muller[1,5,7,*]

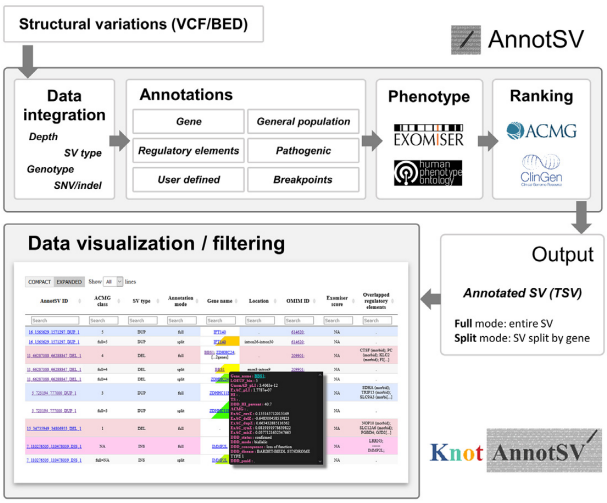
[1]Laboratoire de Génétique Médicale, U1112, INSERM, IGMA, FMTS, Université de Strasbourg, Strasbourg, France, [2]Unité de Génétique Chromosomique, CHU Montpellier, France, [3]Complex Systems and Translational Bioinformatics, ICube, UMR 7357, University of Strasbourg, CNRS, FMTS, Strasbourg, France., [4]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway., [5]Laboratoires de Diagnostic Génétique, IGMA, Hôpitaux Universitaires de Strasbourg, Strasbourg, France., [6]Centre de référence pour les Affections Rares en Génétique Ophtalmologique, Filière SENSGENE, Hôpitaux Universitaires de Strasbourg, Strasbourg, France. and [7]Unité Fonctionnelle de Bioinformatique Médicale appliquée au diagnostic (UF7363), Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Received March 11, 2021; Revised April 16, 2021; Editorial Decision April 27, 2021; Accepted April 29, 2021

ABSTRACT

With the dramatic increase of pangenomic analysis, Human geneticists have generated large amount of genomic data including millions of small variants (SNV/indel) but also thousands of structural variations (SV) mainly from next-generation sequencing and array-based techniques. While the identification of the complete SV repertoire of a patient is getting possible, the interpretation of each SV remains challenging. To help identifying human pathogenic SV, we have developed a web server dedicated to their annotation and ranking (AnnotSV) as well as their visualization and interpretation (knotAnnotSV) freely available at the following address: https://www.lbgi.fr/AnnotSV/. A large amount of annotations from >20 sources is integrated in our web server including among others genes, haploinsufficiency, triplosensitivity, regulatory elements, known pathogenic or benign genomic regions, phenotypic data. An ACMG/ClinGen compliant prioritization module allows the scoring and the ranking of SV into 5 SV classes from pathogenic to benign. Finally, the visualization interface displays the annotated SV in an interactive way including pop-ups, search fields, filtering options, advanced colouring to highlight pathogenic SV and hyperlinks to the UCSC genome browser or other public databases. This web server is designed for diagnostic and research analysis by providing important resources to the user.

GRAPHICAL ABSTRACT



INTRODUCTION

Clinical genetics applications are using more and more high throughput technologies in diagnostic and research settings. Current sequencing techniques encompassing targeted (panel and whole exome) and whole genome sequencing (WGS) but also array technologies are identifying tremendous amount of human variations (1,2).

*To whom correspondence should be addressed. Tel: +33 3 69 55 07 77; Fax: +33 3 69 55 18 94; Email: jeanmuller@unistra.fr
Correspondence may also be addressed to Véronique Geoffroy. Email: veronique.geoffroy@inserm.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

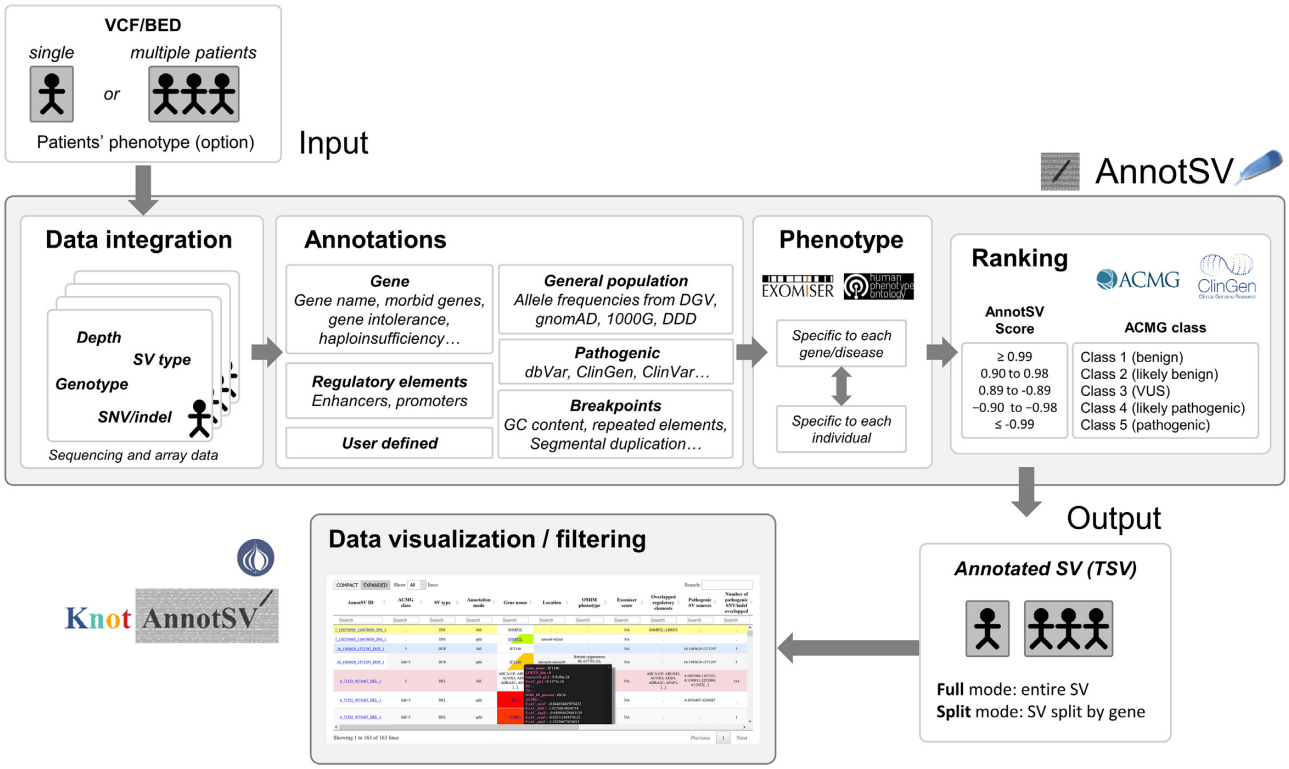


Figure 1. Schematic overview of the AnnotSV and knotAnnotSV workflow. The web server architecture comprises a two-tiered framework: first an annotation engine with AnnotSV (submitting a query, processing the annotations) and second a visualization and filtering interface with knotAnnotSV (showing results, generating the visualization and filtering system).

They include single nucleotide variants (SNV), small insertion/deletions (indel) and structural variations (SV). In particular, WGS generates more and more reliable SV calls in the context of human genetics studies. SV are of specific interest as they play an important role either in genome evolution or as a disease mechanism. Research and clinical communities still face major challenges in annotating and accurately classifying SV. Specific tools for annotation and prioritization are required for helping human geneticists to reliably and quickly interpret clinically relevant SV.

AnnotSV was developed to fulfil these requirements and was regularly updated since the first release (3) introducing a prioritization module from the more to the less pathogenic SV (version 2.0) and a phenotype matching module as well as an improved web based visualization service (version 3.0).

PROGRAM DESCRIPTION AND METHODS

General description of the web server

There are two major parts in the program workflow (Figure 1), first an annotation engine with AnnotSV and second a visualization and filtering interface with knotAnnotSV. AnnotSV is a fast and efficient tool to annotate and classify the SV identified from the human genome. This tool aims at providing annotations useful (i) to interpret SV potential pathogenicity and (ii) to filter out potential false positive variants from all the identified SV.

Inputs/Outputs

The web server accepts queries in three different ways: chromosomal coordinates and SV type (e.g. chr11:66 286 491–66 289 295 del) for a quick single SV analysis and a VCF (Variant Call Format) (4) or a BED file for multiple SV requests. In order to test the general use of the interface, an input example is available on the web page. Optionally, the user can also submit a VCF file with the SNV/indel calls from any sequencing experiment as input from the same sample. These annotations can be of substantial help to identify possible false positive SV calls (i.e. observing heterozygous SNV/indel at the same position of a large deletion) or report a second variant for a recessive disease. The output contains the overlaps of the SV with relevant genomic features where the genes refer to RefSeq (5) or ENSEMBL (6) genes (user defined) as well as phenotypic data or a pathogenic scoring of each SV (Figure 1). Output can be either visualized in a web browser directly, using a specific link, or downloaded as an html file or a tab separated file (tsv file).

Genomic annotations: categories and data sources

Genomic annotations can be performed using either the GRCh37 or GRCh38 build of the human genome (user defined). Some of the annotations are linked to the gene name and thus can be provided independently of the genome build. Numerous relevant annotations are provided encompassing the following categories: Gene, Reg-

Table 1. Summary of annotation sources and their versions available in the current version

Annotations source	Version
...Gene annotations	
Gene annotations (RefSeq)	8/17/2020
Gene annotations (ENSEMBL)	10/24/2020
...Regulatory Elements annotations	
Promoter data (RefSeq)	8/17/2020
Promoter data (ENSEMBL)	10/24/2020
EnhancerAtlas 2.0	6/11/2019
GeneHancer	Licence required
...Gene-based annotations	
DDD disease	12/3/2020
OMIM	11/7/2020
ACMG	ACMG SF v2.0
Gene intolerance (gnomAD)	V2.1.1
Gene intolerance (ExAC)	8/23/2016
Haploinsufficiency (DDD)	7/13/2020
Haploinsufficiency and triplosensitivity (ClinGen)	12/18/2020
Exomiser	20/08/2020 (v2007)
NCBI gene ID	12/18/2020
...Annotations with known pathogenic genes or genomic regions	
ClinVar	12/12/2020
ClinGen	12/18/2020
OMIM	11/7/2020
dbVar	12/2/2020
...Annotations with known pathogenic SNV/indel	
ClinVar	12/12/2020
...Annotations with known benign genes or genomic regions	
gnomAD (GRCh37)	06/03/2019 (v2.1)
ClinVar	12/12/2020
ClinGen	12/18/2020
DGV annotations	2/25/2020
DDD annotations (GRCh37)	3/18/2019
1000 genomes annotations (GRCh37)	5/19/2017
1000 genomes annotations (GRCh38)	11/5/2017
Ira M. Hall's lab annotations	12/31/2018
...Annotations with features overlapped with the SV	
COSMIC annotations	Licence required
TAD boundaries annotations	10/24/2017
...Breakpoints annotations	
GRCh37 FASTA genome	3/20/2009
GRCh38 FASTA genome	1/23/2014
Repeated sequences annotations	7/16/2020
Segmental Duplication annotations	10/8/2020
ENCODE blacklist annotations	2018 (v2)
GAP regions annotations	10/8/2020

ulatory elements, Pathogenic genomic regions, Benign genomic regions and Breakpoints. As an example, ∼30 000 benign SV (Gain, Loss, Insertion, Inversion) and 20 000 pLI (probability of being Loss-of-function Intolerant) and LOEUF (Loss-of-function Observed/Expected Upper bound Fraction) annotations were retrieved from the gnomAD database (7). Around 400 pathogenic regions (Gain, Insertion, Loss), 1100 haploinsufficiency/triplosensitivity genomic regions and 35 benign SV from the ClinGen (8). All the annotations sources and their corresponding version are given in Table 1 and the exhaustive output annotations are listed in Supplementary Table S1.

AnnotSV adjusts the overlapping method between each SV and annotation databases depending on the annotation features. Known pathogenic genes or genomic regions need to be fully overlapped with the SV to annotate, while benign genes or genomic regions need to be fully overlapping with the SV to annotate (Supplementary Figure S1). Genes are reported starting with a single base overlapped.

In order to make sure AnnotSV remains a prime resource for researchers and geneticists, data annotations are regularly updated (once or twice a year) and novel sources are integrated to meet new needs.

Genomic annotations: full and split visualization modes

Annotation is provided into 2 modes: one directly related to the full-length SV (Full mode) and one related to each gene within the SV (Split mode). The Full mode integrates the annotations and ranking of the highest scoring elements overlapped while, the Split mode provides detailed annotations for each overlapped gene (ID, OMIM (9), haploinsufficiency etc.). With the Split mode, users have access to the SV location within each overlapped gene (e.g. 'exon3-intron11', 'txStart-intron19' etc.) and each of these genes is evaluated with respect to individual phenotypic features observed in the patient.

Phenotype-driven prioritization of SV

Prior biological knowledge and phenotype information may help to identify disease genes from human sequencing and array data. Collecting a full and detailed phenotypic profile of the individuals being investigated will certainly improve the chance of prioritizing the correct causative SV (10–12). For this, the patient's phenotype needs to be coded using the Human Phenotype Ontology (HPO) (13). Using Exomiser (14), AnnotSV scores each SV taking into account the most relevant overlapped gene (i.e. the one having the most similar phenotypes with the phenotypic profile under investigation). Genes overlapped with an SV are scored from the lowest (0.0) to the highest similarity (1.0) so that:

- Genes previously associated with disease can be easily highlighted,
- Genes not previously associated with disease can be highlighted,
- Genes associated with diseases that have little or no similarity to the observed phenotypes can be easily removed.

New ACMG/ClinGen based ranking

To assist clinical laboratories in the classification and reporting of SV, the ACMG has developed professional standards in collaboration with the National Institutes of Health (NIH)—funded Clinical Genome Resource (ClinGen) project (15). An automatic SV scoring based on these recommendations (15) has been implemented in AnnotSV which permits each SV to be categorized in one of the following classes: class 1 (benign), class 2 (likely benign), class 3 (variant of unknown significance), class 4 (likely pathogenic) and class 5 (pathogenic). The comprehensive and detailed scoring guidelines are available in the Supplementary Table S2.

Interface: annotation, visualization and filtering system

A user-friendly web server interface is freely available online (https://lbgi.fr/AnnotSV/runjob) to annotate, rank, visualize and filter SV. To help identifying human pathogenic SV,

knotAnnotSV displays the annotated SV in an interactive way including popups, search fields, filtering options, advanced colouring to highlight pathogenic SV and hyperlinks to the UCSC genome browser (16) or other public databases (Figure 2). The annotation is available for the SV as a whole (Full mode or Compact mode in knotAnnotSV) or divided for each overlapped gene (Split mode or Expanded mode in knotAnnotSV). The Full lines are sorted with the most likely pathogenic SV first. Split lines can be displayed below each corresponding Full line and genes are sorted with the most likely pathogenic first.

The SV lines are coloured depending on their SV type (e.g. duplication in blue, deletion in red). Each gene in the split lines is coloured according to their LOEUF score (gradient from the highest haploinsufficiency value in red to the lowest haploinsufficiency value in green). Fully overlapped gene harbour completely coloured cells while partially overlapped genes are partially coloured. This interface includes many functionalities (Table 2) among others hyperlinks to external databases (GeneCards (17), OMIM etc.) and direct link to the UCSC genome browser with specific highlight to the region of interest.

Implementation

AnnotSV is written in Tcl/Tk and runs on all Unix platforms with a standard Tcl/Tk 8.5 installation including four packages ('http', 'json', 'tar' and 'csv'). Bedtools (v2.25 and higher) is required. Optionally, bcftools (v1.10 and higher) can be used for VCF handling and a minimal Java 8 installation is required for the Exomiser dependencies. In order to provide a ready to start installation of AnnotSV, each annotation source (that do not require a commercial license) is automatically downloaded during the installation. knotAnnotSv is written in Perl and runs on all Unix platforms with a standard Perl installation (≥5.22) including 2 CPAN libraries ('YAML::XS' and 'Sort::Key::Natural'). knotAnnotSV implements DataTables (https://datatables.net/) which is a plug-in for the jQuery Javascript library that brings advanced features to HTML tables.

The web application is based on the PHP Symfony framework and the jQuery library for the frontend. Commands are sent to the compute nodes (Linux servers under Ubuntu 18.04) using the SLURM scheduler (v15.08).

RESULTS AND USE CASES

In order to assess the running performance of the web server, datasets from different sources (CGH/SNP arrays, WES, WGS) were annotated (Table 3). In total, from 46 to 3,400 SV ranging in size from 50 bp to 205 Mb were annotated using the GRCh37 build of the human genome taking into account 2 different situations: with or without HPO and with or without sample SNV/indel. As a result, the web server completed the annotation for these SV within a reasonable time frame (∼1–27 min) and with minimal memory (∼100 Mo–1.5 Go) (Table 3). The running time depends mainly on the number of overlapped genes (indirectly both on the SV number and the SV size) resulting in as many annotations split lines. Taking into account additional information especially the HPO scoring impacts the running time of the web server.

Then, we tested the ability of our web server to highlight pathogenic SV in different situations. The challenges in prioritizing functionally and clinically relevant SV has been addressed from array-based techniques and next-generation sequencing datasets (see Supplementary methods).

Case study 1: SNP array (Causal regulatory element overlapped)

SNP array (Affymetrix 2.7M) in a patient with Rieger anomaly (MIM# 137600) identified 3 losses and 3 gains of one copy. Among those SV, a 950 kb deletion at heterozygous state on 4q25 localized 194 kb before the *PITX2* gene. This deletion without any known gene encompasses *PITX2* regulatory elements (Supplementary Figure S2). Deletions located in *PITX2* are already known to cause the Axenfeld-Rieger syndrome (18). This illustrates how the annotation of regulatory elements can identify causative SV.

Case study 2: CGH array (Known pathogenic SV overlapped)

Prenatal diagnosis using a CGH array (Agilent® 180K) in a foetus with retrognathism and cleft palate revealed among 11 SV (4 gains and 7 losses of one copy ranging from ∼62 kb to 3.2 Mb) a single de novo 3.2 Mb heterozygous deletion close to the *SOX9* gene. Direct submission of the deletion coordinates (chr17:66,426,540–69,629,770) to AnnotSV revealed the existence of known pathogenic Copy Number Variant (CNV) in ClinGen and OMIM (Supplementary Figure S3). Deletions of regulatory elements from the *SOX9* gene have been already described in a Pierre-Robin sequence (19).

Case study 3: WGS (Repeated elements highlighted around the breakpoints)

WGS in two affected siblings with a Mainzer-Saldino syndrome (MIM# 26692) identified ∼1400 deletions and ∼2000 duplications per patient using SoftSV (20) and CANOES (21). Among those SV, a 6.7 kb tandem duplication in the *IFT140* gene (22) was highlighted. The breakpoints of the duplication of exon 26 to exon 30 (chr16:1 565 629–1 571 297) overlapped two distinct repeated elements from the Alu family (AluJb and AluJr) potentially suggesting a recurrent mechanism (Supplementary Figure S4).

Case study 4: Multiple cases with custom annotations

In the contexts of cohort analysis or periodic negative cases re-analysis, users can provide SV from multiple samples at once (Supplementary Figure S5). Complementary annotations added in the input file appear as new columns in the output (e.g. Sample Names, associated phenotype etc.).

Our workflow addresses the need for a user-friendly, integrated annotation and visualization SV tool to help the analysis and biological interpretation of SV derived from small SNP/CGH array to large NGS datasets, and to guide precision medicine.

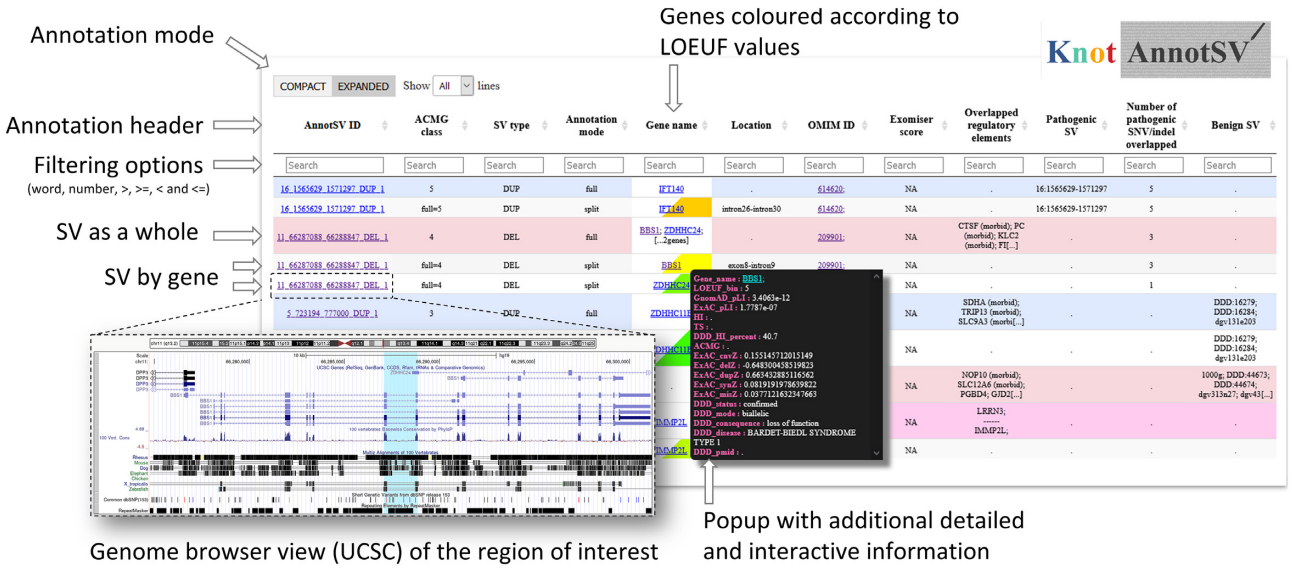


Figure 2. knotAnnotSV interface overview. The interface includes multiple annotations columns with detailed popup. Annotation mode includes presentation of annotation for the Full SV (Compact) or the detailed Split by gene view (Expanded). Each annotation column can be easily filtered. External links allows the user see the SV in the UCSC genome browser.

Table 2. Summary of knotAnnotSV functions available in the current version

Features		Description
Display modes	*COMPACT*	Display only the "Full" AnnotSV lines, giving a SV view
	EXPANDED	Display both "Full" and "Split" AnnotSV lines, giving a SV + gene view
	SINGLE SV FOCUS	Display only the "Full" and "Split" AnnotSV lines of a single SV (by double-clicking on a full line in compact mode)
Data knotting	*Lines sorting*	Lines are sorted according to these priorization rules: ACMG class > Exomiser Score > OMIM morbid > LOEUF bin (this last one is applied on split lines only)
	Tooltips	Hover annotation with mouse to display related informations
	SV type Color coding	The "Full" lines are highlighted depending on their SV type (e.g. duplication in blue, deletion in red). Split lines have a white background
	LOEUF Color coding	Gene annotations are red-to-green color-coded depending on the gene LOEUF values. The color is full or truncated depending on the overlapping feature of the SV
	Regulatory elements	Regulatory elements, whose targets are not overlapped by SV (hence absents from "Gene name" list), are top-listed and ranked according to their Exomiser and morbid features
External links	*UCSC genome browser*	Click on the "AnnotSV ID" to open the SV in the UCSC Genome browser. The SV region is automatically highlighted in light blue and zoomed out by 1.5x for a better genomic contextualisation
	OMIM, GeneCards	Click on the blue hyperlinks to open the corresponding entry web page
Data handling	*Values Sorting*	Click on the column headers to sort values. Come back to the original sorting by clicking on the header of the first column (AnnotSV ID)
	Word Searching	Search words to extract matching records in the displayed annotations as well as in the OMIM, Pathogenic and Gene Name tooltips
	Numerical Filtering	Filter a range of numerical values by preceding a value with the ">", ">=", "<" and "<=" symbols (e.g. to select frequencies less than 1%, type "<0.01")
	Filtering Out	Filter out the numerical values or words respectively preceded with "!=" or "!" (e.g. "! word" or "!= value")
	Filter Saving	At each handling step, all the set-up filters are locally stored

DISCUSSION AND FUTURE UPDATES

Thanks to the many genomic initiatives around the world, CGH and SNP arrays or sequencing experiments generate a tremendous amount of human structural variations leading to the need of specific annotations tools. There are several online tools available like VEP (23) or DeAnnCNV (24) (for review see (25)). Based on these annotations, SV prioritization is gaining interest with a few programs like SVScore (26) and recently machine learning methods paving the future of SV interpretation (27,28). Among those, AnnotSV has been widely used owing to its large amount of available annotation sources and multiple functionalities including false positive detection and prioritization. Here we propose a web server allowing quick and robust annotation and visualization of SV including phenotype driven and ACMG/ClinGen compliant ranking.

Table 3. Comparison of computational speeds for the annotation of different datasets

Experiment source	SV count	Min SV size (bp)	Max SV size (bp)	Mean SV size (bp)	Median SV size (bp)	Running time (without HPO* or SNV/indel**)	Running time with HPO* (without SNV/indel**)	Running time with SNV/indel** (without HPO*)	Running time with HPO* and SNV/indel**	Max memory	Split annotations count
SNP array	2 DEL 4 DUP	53 476	622 597	202 741	119 731	~10 s	~40 s	NA	NA	~100 Mo	21
CGH array	7 DEL 4 DUP	100 000	3 203 230	562.208	144,753	~15 s	~1 min	NA	NA	~100 Mo	301
WES	11 DEL 26 DUP	214	614.757	56.506	12.178	~10 s	~1 min	~40 s	~1 min 30	~150 Mo	110
WGS	~1400 DEL ~2000 DUP	50	205 311 845	2 202 342	101	~9 min	~21 min	~ 14 min	~27 min	~1.5 Go	60.003
The smallest SV from the WGS example	1 DEL	50				~10 s	~30 s	~ 1 min 10	~1 min 30	~380 Mo	1
The largest SV from the WGS example	1 DEL	205 311 845				~30 s	~1 min 50	~1 min 40	~3 min	~100 Mo	1.472

SV datasets from different sources (CGH/SNP arrays, WES, WGS) were annotated to evaluate the running time depending on the number of split annotations, the integration of human phenotype and SNV/Indel information from patients next generation sequencing data (vcf).
*HPO terms used: HP:0004322; HP:0011314; HP:0011297; HP:0001156; HP:0000556; HP:0012047; HP:0000083.
**SNV/indel data used: 314 349 variants for the WES study; 5 134 967 variants for the WGS study

Since the initial release of the program (v1.0) (3), we have tripled the available data sources in our annotation engine, including current versions of the largest available datasets like gnomAD (7), DGV (29), DDD (30) but also gene relevant information such as haploinsufficiency and triplosensitivity from the ClinGen (8), the pLI and LOEUF from ExAC/gnomAD, the OMIM morbid gene list (9). We provide also annotations of regulatory elements from GeneHancer (31) (upon Licence agreement) and Enhancer Atlas (32) as well as specific data for breakpoint analysis such GC content and the presence of repeated elements. Users own annotations (e.g. patient's SNV/indel…) can also be added for annotation. The addition of a phenotype driven module powered by Exomiser (14) will help to automatically integrate the phenotypic data using HPO (13) in large scale studies and score the relation between a gene and the patient's phenotype. As described below, the prioritization module has been added in the second version of AnnotSV and a major update is presented in this version (v3.0). Consequently, less annotations columns are displayed and are directly integrated in the ranking. To our knowledge, this makes our web server the most comprehensive online SV annotation, ranking, filtering and interpretation tool (Supplementary Table S3).

However, we plan to update on a regular basis the annotations sources already available and provide new ones. Mobile element insertions (MEI) are rare but sometimes recurrent genomic events that can have dramatic impact in pathology (33,34). Identification of known polymorphic MEI can be of high importance, thus implementation of specific databases such as dbRIP (35) can be very interesting. False positive calls are still a major issue for SV callers (36). Although, we have already implemented the possibility to report SNV/indel genotype of the same sample together with the SV annotations, more can be done. For example, reporting internal occurrence of SV calls may help to remove recurrent technical false positives. At the interface level, knotAnnotSV is optimized for handling CGH to WES scale data. In fact, DataTables processing of the html output is performed on the user browser side, reducing interface fluidity after 1000 lines (Full/Split). The next developments of knotAnnotSV will be focused on performance enhancement through a server-side processing. Along those lines on a more technical point of view, a database backend will be implemented. This will further improve the annotation running time.

The ACMG/ClinGen recommendations provide a framework to score each SV. AnnotSV is compliant with those recommendations but some of the proposed scoring criteria are not easily accessible and/or requires close evaluation of the clinical context for a given patient. For example, inheritance patterns are not yet included as well as cases from published literature, public databases, and/or internal lab data. However, many ranking improvements are planned. Among the known SV types (i.e. deletions, duplications, insertions, inversions, translocations or even more complex rearrangements), the ACMG-based scoring is only supported for CNV (i.e. gain and loss). As each SV can arbitrary be summarized as a set of novel adjacencies, new SV types will be scored soon using their breakpoints analysis. Indeed, complex forms of SV such as triplica-

tions, inverted duplications, insertional translocations, and chromothripsis could be analysed at the breakpoint-level to determine how genes are disrupted, fused, and/or misregulated by breakpoints or to reveal signatures of diverse DNA repair mechanisms (37). Currently, even if not scored, each SV is ranked on the interface respectively according to its Exomiser score, number of overlapped genes, OMIM and LOEUF.

We believe that AnnotSV and knotAnnotSV will be of great interest and use for the audience interested in human genomics including medical geneticists, cytogeneticists, bioinformaticians and genomics scientists.

DATA AVAILABILITY

The web server is available for free public usage at https://lbgi.fr/AnnotSV. Examples are available on the running page. The source code is available from https://github.com/lgmgeo/AnnotSV and https://github.com/mobidic/knotAnnotSV.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Julius Jacobsen for his assistance during the phenotype-driven analysis module implementation and Jonathan Mercier for his guidance using the Filesystem Hierarchy Standard. We also wish to thank Vincent Zilliox for testing and users for their constant support suggesting new annotations and features. The work in the Chromosomal Genetics unit is supported by the CHROMOSTEM research platform. The authors would like to thank the BiGEst-ICube Platform for assistance.

FUNDING

Inserm; University of Strasbourg; Strasbourg University Hospital. Funding for open access charge: Inserm; University of Strasbourg; Strasbourg University Hospital.
Conflict of interest statement. None declared.

REFERENCES

1. Clark,M.M., Stark,Z., Farnaes,L., Tan,T.Y., White,S.M., Dimmock,D. and Kingsmore,S.F. (2018) Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Med.*, **3**, 16.
2. Lionel,A.C., Costain,G., Monfared,N., Walker,S., Reuter,M.S., Hosseini,S.M., Thiruvahindrapuram,B., Merico,D., Jobling,R., Nalpathamkalam,T. *et al.* (2018) Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.*, **20**, 435–443.
3. Geoffroy,V., Herenger,Y., Kress,A., Stoetzel,C., Piton,A., Dollfus,H. and Muller,J. (2018) AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, **3**, 3572–3574.
4. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
5. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
6. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2017) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
7. Collins,R.L., Brand,H., Karczewski,K.J., Zhao,X., Alföldi,J., Francioli,L.C., Khera,A.V., Lowther,C., Gauthier,L.D., Wang,H. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.
8. Rehm,H.L., Berg,J.S., Brooks,L.D., Bustamante,C.D., Evans,J.P., Landrum,M.J., Ledbetter,D.H., Maglott,D.R., Martin,C.L., Nussbaum,R.L. *et al.* (2015) ClinGen — the clinical genome resource. *N. Engl. J. Med.*, **372**, 2235–2242.
9. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
10. Zemojtel,T., Köhler,S., Mackenroth,L., Jäger,M., Hecht,J., Krawitz,P., Graul-Neumann,L., Doelken,S., Ehmke,N., Spielmann,M. *et al.* (2014) Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.*, **6**, 252ra123.
11. Javed,A., Agrawal,S. and Ng,P.C. (2014) Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods*, **11**, 935–937.
12. Yang,H., Robinson,P.N. and Wang,K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods*, **12**, 841–843.
13. Köhler,S., Carmody,L., Vasilevsky,N., Jacobsen,J.O.B., Danis,D., Gourdine,J.-P., Gargano,M., Harris,N.L., Matentzoglu,N., McMurry,J.A. *et al.* (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic. Acids. Res.*, **47**, D1018–D1027.
14. Smedley,D., Jacobsen,J.O.B., Jager,M., Köhler,S., Holtgrewe,M., Schubach,M., Siragusa,E., Zemojtel,T., Buske,O.J., Washington,N.L. *et al.* (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.*, **10**, 2004–2015.
15. Riggs,E.R., Andersen,E.F., Cherry,A.M., Kantarci,S., Kearney,H., Patel,A., Raca,G., Ritter,D.I., South,S.T., Thorland,E.C. *et al.* (2020) Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.*, **22**, 245–257.
16. Lee,C.M., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Gonzalez,J.N., Hinrichs,A.S., Lee,B.T., Nassar,L.R., Powell,C.C. *et al.* (2019) UCSC Genome Browser enters 20th year. *Nucleic Acids Res.*, **48**, D756–D761.
17. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
18. Volkmann,B.A., Zinkevich,N.S., Mustonen,A., Schilter,K.F., Bosenko,D.V., Reis,L.M., Broeckel,U., Link,B.A. and Semina,E.V. (2011) Potential novel mechanism for Axenfeld-Rieger Syndrome: deletion of a distant region containing regulatory elements of PITX2. *Invest. Ophthalmol. Vis. Sci.*, **52**, 1450–1459.
19. Smyk,M., Roeder,E., Cheung,S.W., Szafranski,P. and Stankiewicz,P. (2015) A de novo 1.58 Mb deletion, including MAP2K6 and mapping 1.28 Mb upstream to SOX9, identified in a patient with Pierre Robin sequence and osteopenia with multiple fractures. *Am. J. Med. Genet. A*, **167**, 1842–1850.
20. Bartenhagen,C. and Dugas,M. (2016) Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. *Brief. Bioinform.*, **17**, 51–62.
21. Backenroth,D., Homsy,J., Murillo,L.R., Glessner,J., Lin,E., Brueckner,M., Lifton,R., Goldmuntz,E., Chung,W.K. and Shen,Y.F. (2014) CANOES: detecting rare copy number variants from whole exome sequencing data. *NucleicAcids Res.*, **42**, e97.
22. Geoffroy,V., Stoetzel,C., Scheidecker,S., Schaefer,E., Perrault,I., Bär,S., Kröll,A., Delbarre,M., Antin,M., Leuvrey,A.-S. *et al.* (2018) Whole-genome sequencing in patients with ciliopathies uncovers a

novel recurrent tandem duplication in IFT140. *Hum. Mutat.*, **39**, 983–992.
23. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
24. Zhang,Y., Yu,Z., Ban,R., Zhang,H., Iqbal,F., Zhao,A., Li,A. and Shi,Q. (2015) DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data. *Nucleic Acids Res.*, **43**, W289–W294.
25. Pös,O., Radvanszky,J., Styk,J., Pös,Z., Buglyó,G., Kajsik,M., Budis,J., Nagy,B. and Szemes,T. (2021) Copy number variation: methods and clinical applications. *Appl. Sci.*, **11**, 819.
26. Ganel,L., Abel,H.J. and FinMetSeq Consortium and Hall, I.M. (2017) SVScore: an impact prediction tool for structural variation. *Bioinformatics*, **33**, 1083–1085.
27. Sharo,A.G., Hu,Z. and Brenner,S.E. (2020) StrVCTVRE: a supervised learning method to predict the pathogenicity of human structural variants. bioRxiv doi: https://doi.org/10.1101/2020.05.15.097048, 13 July 2020, preprint: not peer reviewed.
28. Althagafi,A., Alsubaie,L., Kathiresan,N., Mineta,K., Aloraini,T., Almutairi,F., Alfadhel,M., Gojobori,T., Alfares,A. and Hoehndorf,R. (2021) DeepSVP: integration of genotype and phenotype for structural variant prioritization using deep learning. bioRxiv doi: https://doi.org/10.1101/2021.01.28.428557, 09 April 2021, preprint: not peer reviewed.
29. MacDonald,J.R., Ziman,R., Yuen,R.K.C., Feuk,L. and Scherer,S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
30. Firth,H.V., Wright,C.F. and Study,D.D.D. (2011) The Deciphering Developmental Disorders (DDD) study. *Dev. Med. Child Neurol.*, **53**, 702–703.
31. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, **2017**, doi:10.1093/database/bax028.
32. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
33. Torene,R.I., Galens,K., Liu,S., Arvai,K., Borroto,C., Scuffins,J., Zhang,Z., Friedman,B., Sroka,H., Heeley,J. *et al.* (2020) Mobile element insertion detection in 89, 874 clinical exomes. *Genet. Med.*, **22**, 974–978.
34. Delvallée,C., Nicaise,S., Antin,M., Leuvrey,A.-S., Nourisson,E., Leitch,C.C., Kellaris,G., Stoetzel,C., Geoffroy,V., Scheidecker,S. *et al.* (2021) A BBS1 SVA F retrotransposon insertion is a frequent cause of Bardet-Biedl syndrome. *Clin. Genet.*, **99**, 318–324.
35. Wang,J., Song,L., Grover,D., Azrak,S., Batzer,M.A. and Liang,P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.*, **27**, 323–329.
36. Cameron,D.L., Di Stefano,L. and Papenfuss,A.T. (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.*, **10**, 3240.
37. Weckselblatt,B. and Rudd,M.K. (2015) Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet.*, **31**, 587–599.