Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Notes

AnnotSV: An integrated tool for Structural Variations annotation

Véronique Geoffroy[1,*], Yvan Herenger[2], Arnaud Kress[3], Corinne Stoetzel[1], Amélie Piton[4,5], Hélène Dollfus[1,6] and Jean Muller[1,4]

[1]Laboratoire de Génétique médicale, UMR_S INSERM U1112, IGMA, Faculté de Médecine FMTS, Université de Strasbourg, Strasbourg, France, [2]Service de Génétique Médicale, CHU de Tours, Tours, France, [3]ICUBE UMR 7357, Complex Systems and Translational Bioinformatics (CSTB), Université de Strasbourg - CNRS - FMTS, Strasbourg, France, [4]Laboratoires de Diagnostic Génétique, Institut de Génétique Médicale d'Alsace (IGMA), Hôpitaux Universitaires de Strasbourg, Strasbourg Cedex, France, [5]Institut de Génétique et de Biologie Moleculaire et Cellulaire, INSERM U964, CNRS UMR7104, Université de Strasbourg, 67400 Illkirch, France, [6]Centre de Référence pour les affections rares en génétique ophtalmologique, CARGO, Filière SENSGENE, Hôpitaux Universitaires de Strasbourg, 67091 Strasbourg, France

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on December 22, 2017; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Structural Variations (SV) are a major source of variability in the human genome that shaped its actual structure during evolution. Moreover, many human diseases are caused by SV, highlighting the need to accurately detect those genomic events but also to annotate them and assist their biological interpretation. Therefore, we developed AnnotSV that compiles functionally, regulatory and clinically relevant information and aims at providing annotations useful to i) interpret SV potential pathogenicity and ii) filter out SV potential false positive. In particular, AnnotSV reports heterozygous and homozygous counts of single nucleotide variations and small insertions/deletions called within each SV for the analyzed patients, this genomic information being extremely useful to support or question the existence of an SV. We also report the computed allelic frequency relative to overlapping variants from DGV (MacDonald, et al., 2014), that is especially powerful to filter out common SV. To delineate the strength of AnnotSV, we annotated the 4,751 SV from one sample of the 1000 Genomes Project, integrating the sample information of 4 million of SNV/indel, in less than 60 seconds.

Availability and implementation: AnnotSV is implemented in Tcl and runs in command line on all platforms. The source code is available under the GNU GPL license. Source code, README and Supplementary data are available at http://lbgi.fr/AnnotSV/

Contact: veronique.geoffroy@inserm.fr

Supplementary information: In order to provide a ready to start installation of AnnotSV, each annotation source (that do not require a commercial license) is already provided with the AnnotSV sources. Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-Generation Sequencing (NGS) technologies have been widely used to identify the molecular causes of inherited human diseases. It reveals with a single assay the extreme variability of the human genome composed of millions of Single Nucleotide Variations (SNV), small insertions and deletions (indel) as well as thousands of structural variations (SV) (Sudmant, et al., 2015). SV are an important cause of genetic diseases that are routinely detected by array-based methods. Nevertheless, compared to whole genome sequencing, those methods are of low resolution and noisy signal can make detection of small SV more difficult (Engelhardt, et al., 2017). Since a decade, more than a hundred and fifty tools (Supplementary Table1) have been developed to detect SV from NGS data by using different algorithms, each with their own strengths and weaknesses. However, SV prediction programs often show a high false positive detection rate (Samarakoon, et al., 2016), that

© The Author(s) (2018). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty304/4970516
by University of Durham - Stockton Campus user
on 14 April 2018

requires careful inspection by the user. Thus, apart from the difficulty to properly detect those genomic events with NGS technologies, one of the current challenges is to enhance the automatic analysis of these anomalies.

Existing annotation tools provide functionally, regulatory and/or clinically relevant information using multiple datasets in order to highlight potentially significant findings and prioritize candidates for further analyses (Erikson, et al., 2015; Makarov, et al., 2012; Samarakoon, et al., 2016; Zhang, et al., 2015; Zhao and Zhao, 2013). Nevertheless, several crucial annotations sources are still lacking from the existing tools such as the compiled patient's own information (SNV/indel) or the computed allelic frequencies relative to overlapping variants from DGV (MacDonald, et al., 2014).

Here, we present AnnotSV a new simple tool that combines a complete panel of different datasets to provide high quality SV as well as breakpoints annotation to NGS users including repeated sequences, GC content, Topologically Associating Domains….

2 Overview of AnnotSV

AnnotSV can be accessed from the well documented http://lbgi.fr/AnnotSV/ website. It is a command-line tool written in the Tcl programming language, which can be executed on a variety of operating systems and integrated in any automatic NGS analysis pipeline.

Starting with SV's genomic coordinates called from NGS data and available in a standard VCF or BED file, AnnotSV performs the annotation process first by identifying the genomic overlaps between the input and the annotation features. The overlapping criteria can be either a reciprocal or non-reciprocal overlap (user defined) between the SV and the annotation. These annotations can be performed using either the GRCh37 or GRCh38 build of the human genome. Second, annotations linked to the gene name are also reported. Finally, to offer the most comprehensive and accurate annotation, AnnotSV generates for each SV (i) one annotation based on the full length SV and (ii) one annotation for each gene within the SV. Indeed, this latter annotation is extremely powerful to shorten the identification of mutation in a specific gene. The output is produced as a tab-separated file that can be directly opened in a spreadsheet program. Three different types of annotations are provided, which are summarized here and detailed in Table 1:

-**A genomic based annotations**, providing annotations for the overlapping features with the annotated SV. These annotations include the definition of the genes/transcripts from RefSeq (i.e. ID, Coding DNA Sequence (CDS), transcript length, SV coordinates within the gene), DGV, DECIPHER (Firth, et al., 2009), 1000 Genomes project (phase 3) (Sudmant, et al., 2015), OMIM (Hamosh, et al., 2000), gene intolerance from the ExAC dataset (Lek, et al., 2016), haploinsufficiency (Huang, et al., 2010), promoters, Topologically Associating Domain, GC content, repeated sequences....

In particular, we used the DGV Gold Standard dataset by reporting SV identifiers and by calculating the counts of unique samples with gains and losses, the number of non-redundant samples tested in the related studies and the subsequent relative frequency (Supplementary Figure 1). These annotations are especially powerful for filtering common SV.

-**A patient based annotation**, using the SNV and indel variations identified from the patient's NGS data using a VCF file as an input. The numbers of homozygous and heterozygous variants covered by a SV are reported and can be used to filter out false positives (Supplementary Figure 2). Indeed, first, false positive homozygous deletions can be highlighted by identifying SNV/indel called within the SV. Second, false positive heterozygous deletions can be identified thanks to the presence of heterozygous SNV/indel in the overlapping region. Finally, the presence of only homozygous SNV/indel can be a way to confirm the heterozygous deletions.

-**A custom based annotation** dedicated to the users' specific practice via a custom tab-separated files. For example, one could add each gene transmission mode, known genes in a given group of pathologies…

Table 1. AnnotSV annotations descriptions

Source	Extracted information
SV calling information	
	Name of the chromosome
	Starting position of the SV in the chromosome
	Ending position of the SV in the chromosome
	"Full" or "split by gene" SV annotation
Genomic based annotation	
Refseq genes	Gene symbol
	Transcript symbol
	Length of CDS (bp) overlapping with the SV
	Length of transcript (bp) overlapping with the SV
	SV location in the gene (e.g. « txStart-exon3 »)
	Start and End positions of the intersection between the SV and the transcript
Promoter	List of the genes whose promoters are overlapped by the SV
DGV gold standard	DGV Gold Standard GAIN IDs overlapped with the annotated SV
	Number of individuals with a shared DGV_GAIN_ID
	Number of individuals tested
	Relative GAIN frequency
	DGV Gold Standard LOSS IDs overlapped with the annotated SV
	Number of individuals with a shared DGV_LOSS_ID
	Number of individuals tested
	Relative LOSS frequency
Deciphering Developmental Disorders (DDD)	SV coordinates from the DDD study overlapped with the annotated SV

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty304/4970516
by University of Durham - Stockton Campus user
on 14 April 2018

	Number of individuals with a shared DDD_DUP
	DUP frequency
	Number of individuals with a shared DDD_DEL
	DEL frequency
	DDD category (e.g. confirmed, probable, possible…)
	DDD allelic requirement (e.g. biallelic, hemizygous…)
	DDD mutation consequence (e.g. "loss of function", uncertain…)
	DDD disease name (e.g. "OCULOAURICULAR SYNDROME")
	DDD pubmed identifiers
1000 Genomes project (phase 3) (1000g)	Event types, global allele frequency and maximum allele frequency across the 1000g populations
Gene intolerance	Positive synZ (Z score) indicates gene intolerance to synonymous variation
	Positive misZ (Z score) indicates gene intolerance to missense variation
	pLI score (probability that a gene is intolerant to a loss of function mutation
Haploinsufficiency	Haploinsufficiency ranks
OMIM	OMIM unique six-digit identifier
	OMIM phenotype (e.g. Charcot-Marie-Tooth disease)
	OMIM inheritance (e.g. AD for "autosomal dominant")
GC content	GC content around each SV breakpoint (+/- 100bp)
Repeats	Repeats coordinates around each SV breakpoint (+/- 100bp)
	Repeats type around the two SV breakpoint (+/- 100bp). (e.g. AluSp, SVA_D…)
Topologically Associating Domain (TAD)	TAD coordinates whose boundaries overlapped with the annotated SV
	ENCODE experiments from where the TAD have been defined
Patients based annotation	
VCF file(s) with SNV/Indel	For each patient, number of homozygous SNV/Indel presents in the SV
	For each patient, number of heterozygous SNV/Indel presents in the SV
Filtered VCF file(s) with SNV/Indel	List of heterozygous SNV/Indel presents in the gene overlapped by the SV
Custom based annotation	
Example	Interacting genes…

The different annotations supported by AnnotSV are organized in 3 types: Genomic based, Patient based and Custom based annotations

3 Performance testing

In order to assess the performance of AnnotSV, the genome data from one sample (HG00096) of the 1000 Genome Project was downloaded and annotated. In total, 4,751 SV ranging in size from 50 bp to 1.2 Mb were annotated based on the GRCh37 build of the human genome, integrating also the information from the VCF files of 4 million of SNV/Indel (Supplementary method). As a result, 38% of the SV overlapped a least one gene and 52% overlapped at least one SV from DGV, among which 77% are frequents (present in more than 1% of the samples tested). AnnotSV completed the annotation for these SV in less than 60 seconds on a Linux x86_64 server (Xeon E5-2670). High annotation speed makes AnnotSV suitable for high-throughput sequencing facilities, making it practical to handle hundreds of human genomes in a day.

4 Conclusion

In summary, we developed a new tool for annotating human SV identified from NGS dataset. This tool reduces the time and efforts required to highlight disease-causing SV and so improves the clinical utility of SV detection in NGS data. These analyses are a crucial step to get a better understanding of the human genome and a prerequisite towards personalized medicine.

Funding

Conflict of Interest: none declared.

Acknowledgements

We would like to thank all the colleagues for useful discussions during the study and testing of the program, in particular Sophie Scheidecker, Sinthuja Pachchek, Vincent Zilliox and Luc Moulinier. We also would like to thank Jeffrey MacDonald for his help with the DGV dataset. This work use the computing resources available at the Complex Systems and Translational Bioinformatics laboratory in Strasbourg.

References

Engelhardt, K.R., *et al.* (2017) Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by Whole Exome Sequencing, *Journal of clinical immunology*, **37**, 42-50.

Erikson, G.A., *et al.* (2015) SG-ADVISER CNV: copy-number variant annotation and interpretation, *Genetics in medicine : official journal of the American College of Medical Genetics*, **17**, 714-718.

Firth, H.V., *et al.* (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources, *Am J Hum Genet*, **84**, 524-533.

Hamosh, A., *et al.* (2000) Online Mendelian Inheritance in Man (OMIM), *Human mutation*, **15**, 57-61.

Huang, N., *et al.* (2010) Characterising and predicting haploinsufficiency in the human genome, *PLoS genetics*, **6**, e1001154.

Lek, M., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans, *Nature*, **536**, 285-291.

MacDonald, J.R., *et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome, *Nucleic acids research*, **42**, D986-992.

Makarov, V., *et al.* (2012) AnnTools: a comprehensive and versatile annotation toolkit for genomic variants, *Bioinformatics*, **28**, 724-725.

Samarakoon, P.S., *et al.* (2016) cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data, *BMC genomics*, **17**, 51.

Sudmant, P.H., *et al.* (2015) An integrated map of structural variation in 2,504 human genomes, *Nature*, **526**, 75-81.

Zhang, Y., *et al.* (2015) DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data, *Nucleic acids research*, **43**, W289-294.

Zhao, M. and Zhao, Z. (2013) CNVannotator: a comprehensive annotation server for copy number variation in the human genome, *PloS one*, **8**, e80170.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty304/4970516
by University of Durham - Stockton Campus user
on 14 April 2018