Universidad Autónoma de Madrid

Escuela politécnica superior

Máster en Bioinformática Traslacional y Medicina Personalizada

TRABAJO FIN DE MÁSTER

CNVXPLORER: A WEB TOOL FOR THE INTERPRETATION OF COPY NUMBER VARIANTS

Autor: Francisco Requena Sánchez
Tutor: Antonio Rausell
Ponente: Enrique Carrillo de Santa Pau

SEPTIEMBRE 2019

CNVXPLORER: A WEB TOOL FOR THE INTERPRETATION OF COPY NUMBER VARIANTS

Autor: Francisco Requena Sánchez
Tutor: Antonio Rausell
Ponente: Enrique Carrillo de Santa Pau

Clinical Bioinformatics Lab, Imagine Institut, Paris (France)
Escuela Politécnica Superior
Universidad Autónoma de Madrid
SEPTIEMBRE 2019

Abstract

Abstract

The detection of Copy Number Variants (CNVs) has been gradually introduced into routine diagnostics over the last 15 years and has been described as an important source of pathogenic variants in rare diseases. Despite this, the clinical interpretation remains a challenge and our understanding of the functional impact of these alterations on biological processes is limited, which hinder the diagnosis and the discovery of new therapies. One main reason of these limitations is the lack of integrated data and resources that efficiently assess causative genes with the patient phenotype. To address this challenge, we present CNVxplorer, a user-friendly web application for the interpretation of CNVs. For any given genomic interval, CNVxplorer allows to assess genes mapped in the region, through the integration of human genetics databases of clinical interest, experiment-based information on tissue expression, biomedical ontologies, scientific literature, regulatory regions, functional annotation and the option to perform enrichment analysis. All the results are arranged into a concise output and an automated report can be downloadable. The app provides a highly dynamic and user-friendly interface and the website will be publicly available.

Keywords

copy number variant, rare diseases, R/Shiny, causal genes, regulatory regions

Contents

1 Introduction

Rare diseases affect more than 300 million people worldwide. There are about 6,000-8,000 rare diseases and the majority of these diseases affect children (50-75%) [1]. These disorders are responsible for 35% of deaths in the first year of life in United States and are a significant cause of paediatric hospital admissions [2], where two-third are serious and disabling and half are life-limiting [3]. In spite of the scientific advances during the last years, the clinical outcome has slightly improved for the majority of these patients since major challenges need to be yet solved:

1.1 Challenges for drug development

The development of new therapies is hindered by the reduced commercial viability due to the low prevalence of rare diseases in the population. As a way to revert this reality, regulatory and public entities have provided incentives and a growing public investment to enhance the development of drugs for rare diseases. Initiatives such as the Orphan Drug Act (ODA) in United States or the Regulation on Orphan Medicinal Products in the European Union (EU) in 2000 have been successful in the development of new orphan drugs. For instance, before this, few drugs had been launched to the market, but since its implementation, the FDA has approved more than 600 orphan drugs [4].

Nowadays, it has been estimated that only 6% of rare disease currently have available treatment, of which less that 1% are curative [5], leaving a large majority of patients still awaiting a therapy. Considering the actual pace of research and development, most clinically tractable rare disease will not have new treatments for a long time.

1.2 Challenges for clinical diagnostic

For children with a rare disease, an accurate diagnosis is crucial in order to provide advice, possible therapies and to assess the potential risk to existing family members of future generations. Public initiatives such as the International Rare Diseases Research Consortium (IRDiRC) set the goal for 2017-2027 of “enable all people living with a rare disease to receive an accurate diagnosis, care and available therapy within of coming to medical attention“ [6].

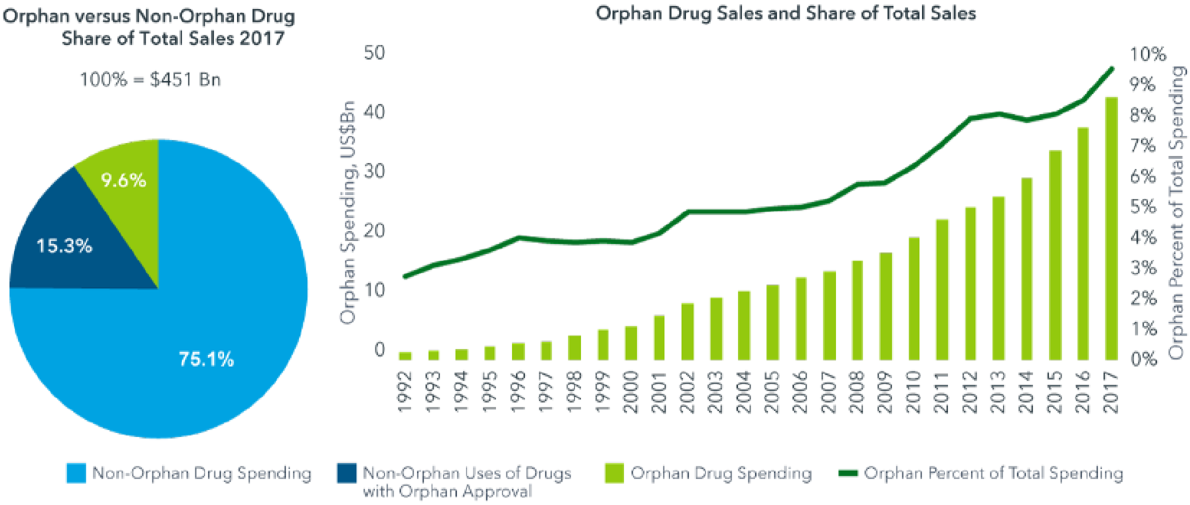

Figure 2.

Figure 1. Spending on orphan drugs in the United States. US dollars. Source: IQVIA National Sales Perspectives, Jan 2018; FDA Orphan Drug Database. Report: Orphan Drugs in the United States Growth Trends in Rare Disease Treatments, IQVIA Institute for Human Data Science, Oct 2018.

In spite of this, achieving precision diagnosis for each individual remains a difficult challenge due to two main aspects:

- **Genetic variability**. The spectrum of genetic variants is broad and heterogeneous, from single nucleotide variants (SNVs) to structural variants (SV) and altered number of chromosomes (aneuploidy) or genome (diploid/triploid mosaicism). Another important source of variability is cis-acting variants that may alter regulation for a unique allele through a change to promoter/enhancer regions (transcription factor binding sites), or even through 3' UTR mutations that affect mRNA stability. Besides, these variants can affect only one allelic copy (heterocygotes) or both (homocygotes).
- **Phenotypic variability**. Due to the high number of rare diseases, clinical features of diseases tend to overlap. In addition, patients with the same disease can present a different phenotype as a result of incomplete penetrance or expressivity.

Two disorders that illustrate this heterogeneity are described below:

- **Cystic fibrosis**. Monogenic disorder caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Despite being a single-gene disorder, there are described around 150 disease causing mutations which are classified in 6 different classes according to the altered molecular function. The genetic variability has generated a broad range of disease severity. Therapeutic options have focused on preventing and treating complications of the disease, until recently, with a better understanding of the genetic abnormalities, new compounds are already being assessed in clinical trials [7].
- **Bardet–Biedl syndrome (BBS)**. Autosomal recessive disorder caused by variants in over 20 genes with indistinguishable clinical presentations [8].

All this heterogeneity adds up to some other aspects:

- Due to the low prevalence of these diseases, in most cases, the cohort studies comprise a small number of patients which reduce the statistical power to find causal variants.
- The influence of environmental factors that are difficult to identify and measure.
- Patients with monogenic diseases can harbour apart variants at loci (e.g. non-coding) that can alter the proper clinical features of the disease.

To sum up, the heterogeneity of rare diseases hinder the assessment of the association genotype-phenotype and therefore the development of new drugs and an accurate diagnostic of these patients.

1.3 Gene discovery, a key step to address these challenges

Genetic cause of at least one-third of rare diseases has yet to be discovered [3]. Development of new diagnostic tools and new drugs certainly depend on the identification of causal genetic drivers and our own understanding of the underlying biological mechanisms.

In this thesis, we focus on copy number variants (CNVs), a class of structural variants which increases or decreases the DNA content through deletions or duplications respectively. CNVs are known to contribute substantially to phenotypic diversity and disease. They have been associated with conditions such as autism [9], schizophrenia [10], Crohn's disease [11], rheumatoid arthritis [11], or type 1 diabetes [11]. In addition, it is a major cause of many developmental disorders including intellectual disability and congenital anomalies [12].

The interpretation of CNVs relies heavily on frequency information from healthy control cohorts and databases with previously reported clinically relevant CNVs. Other factors, such as length, location (interstitial/centromeric/repeat regions) and gene content of the CNV [13] are assess. Since the size of a CNV is directly related with the number of genes affected, the likeliness of pathogenicity increases accordingly with the size of the CNV.

Despite of its relevance, interpreting the pathogenicity of CNVs remain challenging and clinical geneticists need to discriminate arbitrarily pathogenic or high-risk from benign variants in patients. This process is laborious since the clinicians have to curate manually different databases. Even though the CNVs are correctly associated with the clinical features of the patient, the causal variant can remain unsolved since numerous genes can be mapped in the same CNV. Besides, there is not a systematic protocol to interpret these variants, which limit the reproducibility and traceability of the clinical decisions.

Recent efforts have been made to improve the interpretation of CNVs. In addition, some tools have been developed, such as Moon, a commercial software developed by the company Diploid. It consists of a cloud-based software which prioritize variants associated to the symptoms of the patient and displays a variant ranking. Despite the possible advantages, it works as a black-box, since the user can not explore the variables or the parameters used and therefore, it takes away the opportunity to interpret the results by a professional. Other tools, such as MARRVEL [14] or VarCards [15] focus on SNVs and solely permit to analyze genes individually without taking into account the possible aggregated effect of multiple genes and increase the analysis time. As a consequence, clinical and research laboratories which need to interpret CNVs do not have the chance to work efficiently due to the lack of tools oriented to this type of mutations.

Here, we present CNVxplorer, a web tool which allows the interpretation of CNVs in a rapid way through the integration of multiples databases with gene-level information, clinical relevance data, regulatory regions and a set of downstream analysis. The design of the tool is user-friendly and allow to download an automated report with results in order to improve the traceability required in a clinical environment.

2

Results

2.1 CNVxplorer integrates data from human databases

CNVxplorer builds upon and complement existing tools by integrating genic intolerance scores, dosage effect, tissue-specific expression, disease-associated genes, downstream analysis and other information into a user-friendly web application. To analyze a CNV, CNVxplorer allows entry of a genomic interval or a chromosomal band. First, the app displays a general overview of the genes located in the region with broadly used genic intolerance scores (pLI, RVIS, ncRVIS, ncGERP) and identifies those genes located in human genetic databases (OMIM, FDA, GWAS, ClinGen, DECIPHER, ClinVar). Through this information, the user has the option to filter out those genes that not satisfy one or multiples criteria (e.g. genes with pLI >= 0.9 and found in OMIM). These filtered genes will be used along the rest of analysis, but the user can modify any filter and update this list during the same user session. Second, CNVxplorer displays if the region chosen by the user overlaps with any CNV from DECIPHER (CNVs from patients with developmental disorders) or gnomAD/DGV (CNVs from healthy population). Besides, CNVxplorer identifies if the input region is located in an inaccurate genomic annotation regions, such as segmental duplications or low-mappability regions.

2.2 CNVxplorer allows to explore the regulatory architecture

CNVxplorer allows to visualize regulatory regions located in the CNV, such as enhancers or long noncoding RNAs (lncRNAs). In addition, it displays TADs boundaries that are disrupted. Since the enhancers and lncRNAs are characterized by a high level of redundancy, many variants that affect these regulatory regions do not have a functional impact [16] [17]. As a consequence, CNVxplorer displays for each enhancer two scores broadly used in the identification of conserved regions in the genome:

- Conservation inter-species using PhastCon [18] with three different multiple alignments: vertebrates, mammals and primates.
- Conservation intra-species using the gnomAD database [19] of 15,708 genomes from human sequencing studies. To that end, we calculated the ratio of the number of observed variants

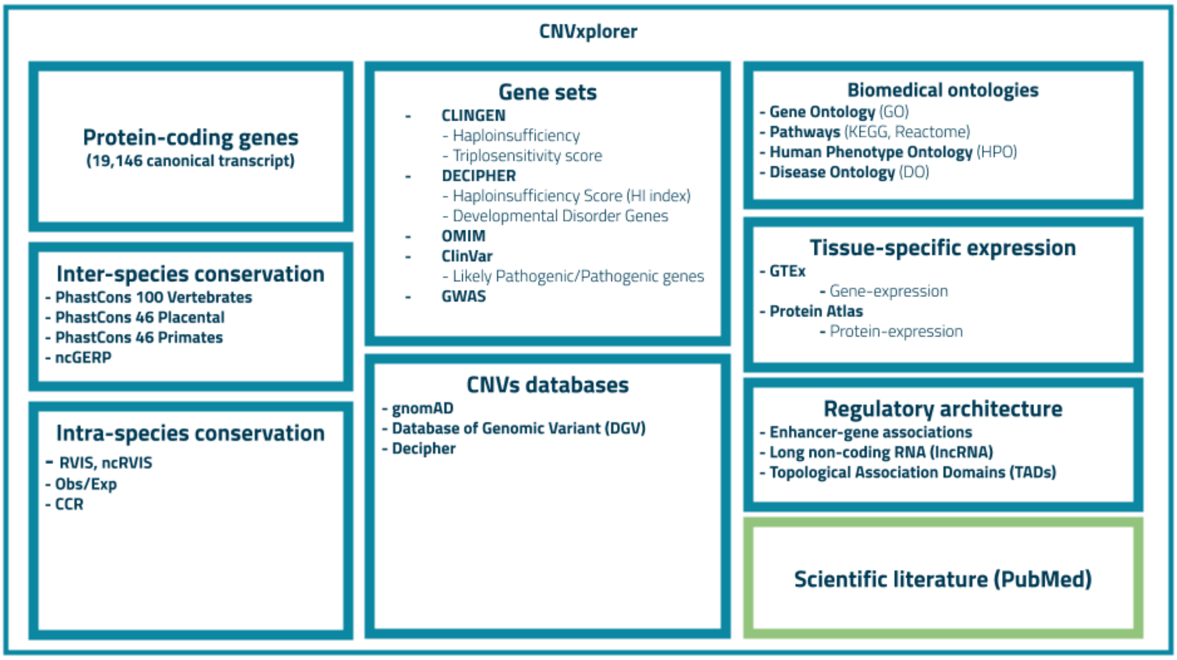
CNVxplorer
Protein-coding genes
(19,146 canonical transcript)
Gene sets
- CLINGEN
- Haploinsufficiency
- Triplosensitivity score
- DECIPHER
- Haploinsufficiency Score (HI index)
- Developmental Disorder Genes
- OMIM
- ClinVar
- Likely Pathogenic/Pathogenic genes
- GWAS
Biomedical ontologies
- Gene Ontology (GO)
- Pathways (KEGG, Reactome)
- Human Phenotype Ontology (HPO)
- Disease Ontology (DO)
Inter-species conservation
- PhastCons 100 Vertebrates
- PhastCons 46 Placental
- PhastCons 46 Primates
- ncGERP
Tissue-specific expression
- GTEx
- Gene-expression
- Protein Atlas
- Protein-expression
CNVs databases
- gnomAD
- Database of Genomic Variant (DGV)
- Decipher
Regulatory architecture
- Enhancer-gene associations
- Long non-coding RNA (lncRNA)
- Topological Association Domains (TADs)
Intra-species conservation
- RVIS, ncRVIS
- Obs/Exp
- CCR
Scientific literature (PubMed)

Figure 2. Data resources used by CNVxplorer. Blue color represents local data and green color represents information obtained through API queries.

and the sum of expected variants per each nucleotide in the region (see Material and Methods).

In addition, the user can filter the list of enhancers based on these scores and has the option to include the target genes, that are not located in the CNV, in the analysis.

2.3 CNVxplorer eases functional analysis of gene groups

CNVxplorer allows the annotation of genes based on Gene Ontology (GO), where the user can specify the GO level and the ontology source. In addition, it can perform functional enrichment analysis from multiples ontologies: biological process, molecular function or cellular component (Gene Ontology), pathways (KEGG, REACTOME) and diseases (Disease Ontology). Besides, the user can specify the p-value cutoff of each analysis. Lastly, these results are displayed through plots and table format.

2.4 CNVxplorer facilitates the clinical interpretation of CNVs

In order to enhance the clinical interpretation between the genes of the region affected and the patient's phenotype, CNVxplorer offers three different approaches:

First, CNVxplorer summarizes human expression data per gene from two sources. For gene expression data, the source is GTEx, which provides quantitative expression data of each gene in 53 human tissues. For protein levels, the source is the Protein Atlas database.

The information about gene-level expression across tissues brings the possibility to assess genes expressed in tissues that are associated with the affected organ(s) in the patient.

Second, the user can select the phenotype terms observed in the patient. CNVxplorer will display only the list of genes located in the CNV that have been associated previously with these clinical features.

Third, CNVxplorer displays the articles from PubMed that has been associated before with that genomic region. Every displayed article includes the title, date of publication, number of cites and a link to the original manuscript.

Here, we provide an example of how CNVxplorer displays information useful for gene prioritization. We describe a specific case for which CNVxplorer can be used to facilitate manual analysis of genes located in a CNV.

Zhu et al. [20] recently successfully performed the identification of six CNV regions associated with congenital diaphragmatic hernia (CDH). This disorder is characterized by a developmental discontinuity of the diaphragm and hypoplasia of the lungs.

As an example of how CNVxplorer can help to interpret CNVs, we investigate the gene content of one of the CNV region described by the authors (chr17:34813719–36278623). To emphasize the flexibility of CNVxplorer, we use two different approaches (Figure 3) to interpret 15 genes mapped in this region:

Approach 1. The first step is to filter genes based on pLI score (pLI $>=$ 0.9). Of the 15 genes, 4 (ACACA, GGNBP2, HNF1B and SYNRG) have a pLI score above or equal to 0.9, indicating a high intolerance of loss-of-function variants. Based on functional analysis (Gene Ontology - Biological Process), we detect 2 genes (GGNBP2, HNF1B) enriched significantly (p-value $<$ 0.05) for the terms: utero embryonic development (GO:0001701) and embryonic organ development (GO:0048568). To corroborate the relevance of these two genes, we compare it with clinical features of patients with CDH. We found one gene (HNF1B) associated with Pancreatic hypoplasia (HP:0002594) but not with Lungs hypoplasia (HP:0002089). Since CDH is characterized by extradiaphragmatic associated anomalies (46% of the cases [21], the disruption of this gene could explain partially the phenotypic heterogeneity of this disease.

Approach 2. In this case, we decided to use HI score (HI $<=$ 10). Of the 15 genes, 5 (AATF, ACACA, HNF1B, LHX1, TADA2A) have a HI score below or equal to 10, indicating a high probability of being a haploinsufficient gene. Based on functional analysis (Gene Ontology - Biological Process), we detect 2 genes (HNF1B, LHX1) enriched significantly (p-value $<$ 0.05) for the terms: regulation of gastrulation (GO:0010470) and embryonic organ development (GO:0048568). We assess these two genes with the current biomedical literature and both genes have been described before playing a role in congenital disorders. In the case of HNF1B in Duodenal atresia (PMID: 25256560) and developmental kidney disease (PMID:27234567) and LHX1 in Müllerian aplasia (PMID:23954021, PMID:22740494).

These two approaches reflect the flexibility of the application to analyse a genomic interval, since the user has the opportunity to explore different parameters to filter genes. For instance, in the first approach, the gene LHX1 is filtered out due to a low pLI score (0.03). Thanks to a second approach, using HI score as the parameter to filter out genes, this gene is included and identified as a potential gene due to its association with Mullerian aplasia.

CNVxplorer allows to identify novel candidate genes such as HNF1B, LHX1 and GGNBP2 (Figure 3), which were described as relevant by Zhu et al. [20] aswell.

2.5 CNVxplorer permits the reproducibility of the results

CNVxplorer offers the option to download an automated generated report with the results obtained and the information required to reproduce the analysis. In addition, the app offers an optional feature to include notes, patient and clinician information.

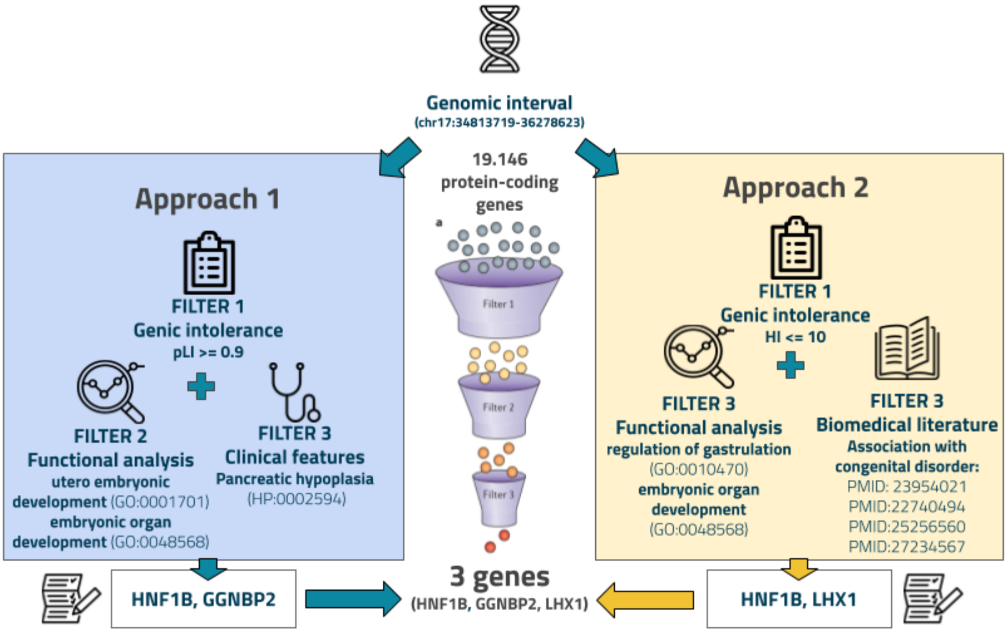

Figure 3. Example of how CNVxplorer can be used to analyze CNVs. Once the user selects the genomic interval (chr17:34813719–36278623), two approaches are performed: approach 1 (HNF1B, GGNBP2) and approach 2 (HNF1B, LHX1).

3
Material and Methods

3.1 CNVxplorer overview

CNVxplorer has been developed in R. The program follows a modular design that generates an automated pipeline workflow with optional steps, such as the enrichment analysis. The code was implemented in an interactive web application using Shiny (http//shiny.rstudio.com), an R framework that couples the R-based statistics computation and graphics generation to the rendering of a Web-based user interface. This technology allows the fast implementation of R code in an easy-to-use frontend in a "reactive" environment, where the user can interact with the data (e.g. filtering out genes through different approaches) and see the changes during the same user session. Since the application is hosted in a remote server, the user does not need to consume local resources and just requires a web browser to use the tool. During every session, the parameters and data input are rested for every user. Additional materials, documentation and FAQs will be soon available on the CNVxplorer website.

3.2 Gene-level data source

Based on HUGO Gene Nomenclature Committee (HGNC), we retrieve a list of 19,146 protein-coding genes symbols. We set the coding boundaries of HGNC genes to Ensembl Gene 97 with the reference human genome GRCh37.p13.

We collected genic intolerance scores of each gene from three different studies: (i) the residual variation intolerance score (RVIS) [22], (ii) loss-of-function (LoF) intolerance (gene intolerance score based on loss-of-function variants in 141,456 individuals), defined as pLI score, observed/expected ratio [19] and (iii) the number of constrained coding regions (CCRs) in the 99th percentile or higher [23]. An alternative version of RVIS (ncRVIS) and GERP (ncGERP) using proximal regulatory regions (5'UTR, 3'UTR, 250bp upstream of TSS) was used [22].

Besides, we included curated databases that classify genes according to its dosage effect: haploinsufficieny, triplosensitivity genes [24] and a quantitative score of the haploinsufficiency (HI index) using a predictive model based on the differences between those genes and the rest of them [25]. We curated genes associated with developmental disorders [24] and genes whose protein products are known to be the mechanistic targets of FDA-approved drugs [26]. In addition, genes associated with variants classified as likely-pathogenic or pathogenic were curated from ClinVar [27]

Finally, we collected gene expression data of 42 tissue from the genotype-tissue Expression Project (GTEx) [28] and protein expression of 48 tissues from Human Protein Atlas (HPA) [29]

3.3 Clinical data source

HPO terms and diseases associated with genes were obtained from Human Phenotype Ontology (HPO) [30], Online Mendelian Inheritance in Man (OMIM) [31] and ORPHANET [32]. Description of each human phenotype term is retrieved with the R package rols (v. 2.12.0). Access to PubMed articles is performed with the R package rentrez (v.1.2.2). The query is performed with the chromosomal bands that overlap with the genomic interval defined by the user.

3.4 Functional genomics data source

We curated associations of enhancers with target genes [33]. We filtered out predicted enhancers and enhancer-gene associations below score 1. In total, we collected 107,168 enhancer-target pairs. The TAD boundaries file used [34] was generated using H1 human embryonic stem cells (hESCs), a bin size of 40kbs and a window size of 2 Mb. We curated lncRNAs data from human with their genomic coordinates and expression profiles across 7 organs, from early organogenesis to adulthood [35].

3.5 Sequence conservation of regulatory regions

PhastCons scores for three multi-species alignment (verterbrates, mammals and primates) were obtained from CADD (version 1.3) [36]. Intra-species conservation score is calculated as the ratio between the number of observed variants and the expected number of variants in a specific genomic region. A value closer to 0 means that the region is highly conserved since the number of variants found is much lower in comparison with the number of expected variants. Observed variants are obtained from VCF files of gnomAD and the expected number of variants are calculated as the sum of the expected probabilities of mutability for each nucleotide based on its local sequence context. Expected probabilities for each triplet of nucleotides were obtained from Supplementary Table 1 [37].

3.6 Functional analysis

Functional enrichment of annotations from the GO Consortium are performed using R package ClusterProfiler (v. 3.12.0) [38]. For the pathway analysis, we considered two sources: KEGG [39] and REACTOME [40]. The enrichment analysis is performed with the R packages ClusterProfiler [38] and reactomePA [41] respectively. Disease Ontology enrichment analysis is computed and visualized with the R package DOSE (v. 3.10.2) [42]

3.7 Data processing

All input features were mapped to the human reference build hg19. Since enhancer data was originally mapped in hg38, we used Liftover chains [43] to convert it to hg19. The list of inaccurate genomic regions, such as segmental duplications or low mappability were obtained [44].

3.8 Clinical report

Clinical reports are generated with Rmarkdown (http://rmarkdown.rstudio.com/), a R package to elaborate dynamic documents with R. The reports are generated automatically depending on the data input and configurable parameters introduced by the user. The design of the document is generated with the R package prettydoc (v 0.1.0).

3.9 Software availability

CNVxplorer will be freely accessible from a website. For a local installation, a github page with the source code will be soon available. In order to provide easy installation, a docker-image container with required data and code will be available for the user too.

4
Dicussion

In summary, CNVxplorer comprises an efficient aggregation of information from human genomics databases and the possibility of multiple functional analysis, which allow for quick overview and assessment of candidate genes and their underlying mechanism in a genomic interval defined by the user. OMIM provides essential information about the association of mendelian diseases and genes. DGV and gnomAD provide a complete database of CNVs from healthy population. Besides, gnomAD provides genic-intolerance scores which assess the potential pathogenicity. ClinVar provides variants which are classified by their pathogenicity. GWAS Catalog provides SNP-trait associations and their mapped genes. Clingen offers a curated list of haploinsufficiency and triplosensitivity genes. DECIPHER provides CNVs from patients with developmental disorders.

Since the identification of causal genes is a challenge due to the heterogeneity of both the genotype and the phenotype, the user can filter genes mapped in the region based on the gene-level information displayed. In addition, the app discards those genes identified as less relevant from a clinical point of view. Once the most relevant genes have been selected, the user can assess clinical features of the patient (e.g. organ(s) affected) with tissue-specific expression by directly selecting the phenotype terms observed in the interface (e.g. Blindness, Heart defect...). CNVxplorer will search for genes, which have been associated previously with these phenotype terms.

Despite the actual version of the application uses diverse sources of data, a next step would be the inclusion of more resources of clinical interest, such as: (i) Databases like NephQTL [45] or CKDdb [46] , both based on kidney-related pathologies, that summary expression quantitative trait loci (eQTL) studies and provide a reliable way to assess genes and phenotypes tissue-specific. (ii) Another source of interesting databases that can be useful are those that describe drug-gene interactions such as DGIdb or Drugbank which would allow to infer possible alternative treatments. (iii) Model organism databases that have demonstrated to be practical in the interpretation of variants [14] [47], such as mouse (MGD) [48], zebrafish (ZFIN) [49] and drosophila (FlyBase) [50]

CNVxplorer provides access to regulatory elements such as enhancers, lncRNA and TADs. Due to the redundancy of the regulatory circuit of the genome, the assessment of causal variants mapped in enhancers or lncRNAs with a phenotype remains a challenge. CNVxplorer displays inter-species conservation scores in order to identify accurately those susceptible regions. Even though the conservation scores have been widely used in the interpretation of variants located in mapped-regions, it has been observed that evolutionary conservation can not help much in the prediction of disease-related enhancers, since human-specific enhancers can harbour important functions [51]. CNVxplorer offers a complementary score based on intra-species conservation that should facilitates the interpretation of CNVs that disrupt these regions. CNVxplorer also gives the option to include , the target genes of enhancers that have been disrupted and are not located in the CNV in the downstream analysis. Through these resources, our software allows to assess the effect of regulatory regions and their target genes that are likely not included in manual analysis or using other software tools.

CNVxplorer was conceived as a tool used in a clinical and research environment. To that end, CNVxplorer has been developed as a quick and user-friendly tool which allows to download and export the results obtained for further analysis in other software (e.g. Microsoft Excel), as well as to download an automated report with the required data to replicate the analysis. These features suppose an improvement in comparison with manual analysis based on inconsistency and subjectivity.

Despite these advantages, our application have limitations. Since the purpose of the application is to be implemented in a clinical and research environment, a local installation is recommended. In this case, the user would have to download the source code, install the dependencies and download the data needed to run the app. This process can hinder its implementation. This issue can improve using software containers such as Docker which encapsulate all the requirements to run the app and allow an easy installation. Another issue is the GeneHancer data which identifies enhancers without taking into account their tissue-specificity. Some projects [52] [53] provide tissue-specific regulatory regions, which can improve the assessment of causal variants mapped in those regions and the clinical features of the patient.

A next step, will be to adapt our program to analyze Next Generation Sequencing (NGS) data accepting multiple CNVs at the same time. Another possible feature, it is the development of new techniques, such as machine learning models, that would allow a ranking of each variant based on the gene content.

In conclusion, CNVxplorer offers a new approach to the interpretation of CNVs that provides the opportunity to prioritize genes located in CNVs that may be causal for a specific phenotype thus improving our knowledge of the underlying biological mechanism responsible for the disease. The use of this application together with other tools will provide new insight into the effect of CNVs in a clinical context.

5
Supplementary information

5.1 Table S1: List of Human Genetics Databases

Source	Description	Reference
GWAS Catalog	Genes associated with GWAS studies	[54]
DrugBank	Protein targets of FDA-approved drugs	[26]
DECIPHER	Unhealthy population (CNVs)	[24]
OMIM	Genes associated with mendelian disorders	[24]
gnomAD	Gene intolerance score (pLI)	[19]
gnomAD	Gene intolerance score (obs/exp)	[19]
gnomAD	Healthy population (CNVs)	[19]
DGV	Healthy population (CNVs)	[55]
ClinVar	Genes associated with human variants	[27]
Clingen	Haploinsufficiency genes	[54]
Clingen	Triplosensitivity genes	[54]
GTEx	Gene-expression data	[28]
HPA	Protein-expression data	[29]

Acronyms

- **CNV**: Copy Number Variant
- **SNV**: Single Nucleotide Variant
- **TAD**: Topological Associating Domain
- **lncRNAs**: long noncoding RNA
- **GO**: Gene Ontology
- **HPO**: Human Phenotype Ontology
- **DO**: Disease Ontology
- **OMIM**: Online Mendelian Inheritance in Man

Bibliography

[1] Joachim Rode. Rare diseases: understanding this public health priority. *Paris: EURORDIS*, 2005.

[2] Paula W Yoon, Richard S Olney, Muin J Khoury, William M Sappenfield, Gilberto F Chavez, and Don Taylor. Contribution of birth defects and genetic diseases to pediatric hospitalizations: a population-based study. *Archives of pediatrics & adolescent medicine*, 151(11):1096–1103, 1997.

[3] Kym M Boycott and Diego Ardigo. Addressing challenges in the diagnosis and treatment of rare genetic diseases. *Nat Rev Drug Discov*, 17:151–2, 2018.

[4] Nina L Hunter, Gayatri R Rao, and Rachel E Sherman. Flexibility in the fda approach to orphan drug development. *Nature Reviews Drug Discovery*, 16(11):737, 2017.

[5] Hugh JS Dawkins, Ruxandra Draghia-Akli, Paul Lasko, Lilian PL Lau, Anneliene H Jonker, Christine M Cutillo, Ana Rath, Kym M Boycott, Gareth Baynam, Hanns Lochmüller, et al. Progress in rare diseases research 2010–2016: an irdirc perspective. *Clinical and translational science*, 11(1):11, 2018.

[6] Christopher P Austin, Christine M Cutillo, Lilian PL Lau, Anneliene H Jonker, Ana Rath, Daria Julkowska, David Thomson, Sharon F Terry, Béatrice de Montleau, Diego Ardigò, et al. Future of rare diseases research 2017–2027: an irdirc perspective. *Clinical and translational science*, 11(1):21–27, 2018.

[7] F Ratjen, SC Bell, SM Rowe, CH Goss, and AL Quittner. Bush a. cystic fibrosis. *Nat Rev Dis Primers*, 1:15010, 2015.

[8] Asli Ece Solmaz, Huseyin Onay, Tahir Atik, Ayca Aykut, Meltem Cerrah Gunes, Ozge Ozalp Yuregir, Veysel Nijat Bas, Filiz Hazan, Ozgur Kirbiyik, and Ferda Ozkinay. Targeted multigene panel testing for the diagnosis of bardet biedl syndrome: identification of nine novel mutations across bbs1, bbs2, bbs4, bbs7, bbs9, bbs10 genes. *European journal of medical genetics*, 58(12):689–694, 2015.

[9] Dalila Pinto, Elsa Delaby, Daniele Merico, Mafalda Barbosa, Alison Merikangas, Lambertus Klei, Bhooma Thiruvahindrapuram, Xiao Xu, Robert Ziman, Zhuozhi Wang, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics*, 94(5):677–694, 2014.

[10] Dheeraj Malhotra and Jonathan Sebat. Cnvs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, 148(6):1223–1241, 2012.

[11] Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F Conrad, Eleni Giannoulatou, et al. Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713, 2010.

[12] Charles Lee and Stephen W Scherer. The clinical context of copy number variation in the human genome. *Expert reviews in molecular medicine*, 12, 2010.

[13] Karen Buysse, Barbara Delle Chiaie, Rudy Van Coster, Bart Loeys, Anne De Paepe, Geert Mortier, Frank Speleman, and Björn Menten. Challenges for cnv interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European journal of medical genetics*, 52(6):398–403, 2009.

[14] Julia Wang, Rami Al-Ouran, Yanhui Hu, Seon-Young Kim, Ying-Wooi Wan, Michael F Wangler, Shinya Yamamoto, Hsiao-Tuan Chao, Aram Comjean, Stephanie E Mohr, et al. Marrvel: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *The American Journal of Human Genetics*, 100(6):843–853, 2017.

[15] Jinchen Li, Leisheng Shi, Kun Zhang, Yi Zhang, Shanshan Hu, Tingting Zhao, Huajing Teng, Xianfeng Li, Yi Jiang, Liying Ji, et al. Varcards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic acids research*, 46(D1):D1039–D1048, 2017.

[16] Marco Osterwalder, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J Mannion, Sarah Y Afzal, Elizabeth A Lee, Yiwen Zhu, Ingrid Plajzer-Frick, Catherine S Pickle, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239, 2018.

[17] Aurélie Kapusta and Cédric Feschotte. Volatile evolution of long noncoding rna repertoires: mechanisms and biological implications. *Trends in Genetics*, 30(10):439–452, 2014.

[18] Melissa J Hubisz, Katherine S Pollard, and Adam Siepel. Phast and rphast: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, 12(1):41–51, 2010.

[19] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

[20] Qihui Zhu, Frances A High, Chengsheng Zhang, Eliza Cerveira, Meaghan K Russell, Mauro Longoni, Maliackal P Joy, Mallory Ryan, Adam Mil-Homens, Lauren Bellfy, et al. Systematic analysis of copy number variation associated with congenital diaphragmatic hernia. *Proceedings of the National Academy of Sciences*, 115(20):5247–5252, 2018.

[21] Joanna CE Wright, Judith LS Budd, David J Field, and Elizabeth S Draper. Epidemiology and outcome of congenital diaphragmatic hernia: a 9-year experience. *Paediatric and perinatal epidemiology*, 25(2):144–149, 2011.

[22] Slavé Petrovski, Ayal B Gussow, Quanli Wang, Matt Halvorsen, Yujun Han, William H Weir, Andrew S Allen, and David B Goldstein. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS genetics*, 11(9):e1005492, 2015.

[23] James M Havrilla, Brent S Pedersen, Ryan M Layer, and Aaron R Quinlan. A map of constrained coding regions in the human genome.

[24] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533, 2009.

[25] Ni Huang, Insuk Lee, Edward M Marcotte, and Matthew E Hurles. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics*, 6(10):e1001154, 2010.

[26] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2017.

[27] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2013.

[28] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[29] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[30] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic acids research*, 47(D1):D1018–D1027, 2018.

[31] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

[32] Steffanie S Weinreich, R Mangon, JJ Sikkens, ME Teeuw, and MC Cornel. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

[33] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, 2017, 2017.

[34] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331, 2015.

[35] Ioannis Sarropoulos, Ray Marin, Margarida Cardoso-Moreira, and Henrik Kaessmann. Developmental dynamics of lncrnas across mammalian organs and species. *Nature*, page 1, 2019.

[36] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2018.

[37] Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, Christine Stevens, Aniko Sabo, Lauren M McGrath, Jack A Kosmicki, Karola Rehnström, Swapan Mallick, Andrew Kirby, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944, 2014.

[38] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[39] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[40] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[41] Guangchuang Yu and Qing-Yu He. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12(2):477–479, 2016.

[42] Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2014.

[43] Angela S Hinrichs, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598, 2006.

[44] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. The encode blacklist: Identification of problematic regions of the genome. *Scientific Reports*, 9(1):9354, 2019.

[45] Christopher E Gillies, Rosemary Putler, Rajasree Menon, Edgar Otto, Kalyn Yasutake, Viji Nair, Paul Hoover, David Lieb, Shuqiang Li, Sean Eddy, et al. An eqtl landscape of kidney tissue in human nephrotic syndrome. *The American Journal of Human Genetics*, 103(2):232–244, 2018.

[46] Marco Fernandes and Holger Husi. Establishment of a integrative multi-omics expression database ckddb in the context of chronic kidney disease (ckd). *Scientific reports*, 7:40367, 2017.

[47] Matthew Jensen and Santhosh Girirajan. An interaction-based model for neuropsychiatric features of copy-number variants. *PLoS genetics*, 15(1):e1007879, 2019.

[48] Carol J Bult, Judith A Blake, Cynthia L Smith, James A Kadin, and Joel E Richardson. Mouse genome database (mgd) 2019. *Nucleic acids research*, 47(D1):D801–D806, 2018.

[49] Douglas G Howe, Yvonne M Bradford, Tom Conlin, Anne E Eagle, David Fashena, Ken Frazer, Jonathan Knight, Prita Mani, Ryan Martin, Sierra A Taylor Moxon, et al. Zfin, the zebrafish model organism database: increased support for mutants and transgenics. *Nucleic acids research*, 41(D1):D854–D860, 2012.

[50] Jim Thurmond, Joshua L Goodman, Victor B Strelets, Helen Attrill, L Sian Gramates, Steven J Marygold, Beverley B Matthews, Gillian Millburn, Giulia Antonazzo, Vitor Trovisco, et al. Flybase 2.0: the next generation. *Nucleic acids research*, 47(D1):D759–D765, 2018.

[51] Antonio CA Meireles-Filho and Alexander Stark. Comparative genomics of gene regulation—conservation and divergence of cis-regulatory information. *Current opinion in genetics & development*, 19(6):565–570, 2009.

[52] Daniel Marbach, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods*, 13(4):366, 2016.

[53] Aziz Khan and Xuegong Zhang. dbsuper: a database of super-enhancers in mouse and human genome. *Nucleic acids research*, 44(D1):D164–D171, 2015.

[54] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2018.

[55] Jeffrey R MacDonald, Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992, 2013.