MINI-REVIEW ARTICLE

Possibilities and Limitations of CNV Interpretation Software and Algorithms in *Homo Sapiens*

Maria A. Zelenova[1,2] and Ivan Y. Iourov[1,2,3,*]

[1]Mental Health Research Center, Moscow, 117152, Russia; [2]Veltischev Research and Clinical Institute for Pediatrics of the Pirogov Russian National Research Medical University, Ministry of Health of Russian Federation, Moscow, 125412, Russia; [3]Department of Medical Biological Disciplines, Belgorod State University, Belgorod, 308015, Russia

ARTICLE HISTORY

Received: May 11, 2022
Revised: July 15, 2022
Accepted: July 21, 2022

DOI:
10.2174/1574893617666220907121155

Abstract: ***Background***: Technical advances and cost reduction have allowed for the worldwide popularity of array platforms. Otherwise called “molecular karyotyping”, it yields a large amount of CNV data, which is useless without interpretation.

Objective: This study aims to review existing CNV interpretation software and algorithms to reveal their possibilities and limitations.

Results: Open and user-friendly CNV interpretation software is limited to several options, which mostly do not allow for cross-interpretation. Many algorithms are generally based on the Database of Genomic Variants, CNV size, inheritance data, and disease databases, which currently seem insufficient.

Conclusion: The analysis of CNV interpretation software and algorithms resulted in a conclusion that it is necessary to expand the existing algorithms of CNV interpretation and at least include pathway and expression data. A user-friendly freely available CNV interpretation software, based on the expanded algorithms, is yet to be created.

Keywords: SNP array, CNV interpretation, bioinformatics, genetics, brain diseases, molecular karyotyping.

1. INTRODUCTION

Copy number variations (CNVs) cause numerous neuropsychiatric disorders, especially the ones featured by intellectual disability and congenital malformations. Due to the size (up to 1 Mbp), CNVs may simultaneously disrupt dozens of genes and cause significant changes to the genome, undermining patients’ health. Data on CNVs may be obtained from different sources, mainly from molecular karyotyping (such as SNP array) and NGS (Next-Generation Sequencing). However, it carries little value without further interpretation. Numerous tools for CNV analysis are mostly aimed at professional bioinformaticians and offer CNV calling from NGS data [1]. Software allowing for the interpretation itself is limited. Thus, laboratories working with the microarray data have to either restrict their analysis to gross and mostly unambiguous rearrangements, missing on potentially significant smaller changes, or hire a specialist to develop or use specialized tools (which is often a problem for small laboratories with moderate load) [2]. In the current article, we considered the algorithms, software, tools and pipelines most similar to the task of CNV interpretation, whether CNV data were obtained from microarray or sequencing analysis. This article aims to highlight their advantages and disadvantages in terms of CNV interpretation and user-friendly design.

2. ANALYSIS OF THE EXISTING SOFTWARE FOR CNV INTERPRETATION

Different software selections give a broad overview of the purposes of commonly used instruments for CNV interpretation. However, most programs mentioning gene choice or algorithms for microarray analysis solely focus on gene expression analysis, which is not relevant to our review. Many tools are aimed at tumor analysis or CNV calling/detection from NGS data. A great majority of programs are meant for use by professionals with a background in programming and require training in R/Python/Perl [1]. However, little attention is paid to CNV interpretation or prioritization tools, which are aimed at processing array data and are suitable for use by researchers with no or little programming experience. Although programming is important for CNV calling, CNV interpretation relies on the profound knowledge of genomics, which makes genetic experience prevail over programming at this stage. Tools relevant to CNV interpretation that are not related to tumor cells [3-6] are presented in Table **1**.

*Address correspondence to this author at the Mental Health Research Center, 117152 Moscow, Russia; Tel: +7-495-109-03-93 + 3500; E-mail: ivan.iourov@gmail.com

1574-8936/22 $65.00+.00

Table 1. Tools most relevant to CNV interpretation. The inclusion criteria are a presence of a webserver/GUI, easy access (*e.g.*, no requests to the founders required) and no previous programming knowledge.

Name	Description and Usability Comments	Refs.	Link
CNVxplorer	Assessment of CNVs in multiple areas. The software yields data on clinically relevant information (*i.e.*, diseases), mouse model comparison, various statistics from literature (from titles and abstracts), expression, protein interactions, and regulatory regions overlap.	[3]	http://cnvxplorer.com/
ClinTAD	CNV interpretation in the context of topologically associated domains.	[4]	https://www.clintad.com/
The Ensembl Variant Effect Predictor	Predicts the effects of CNVs and SNPs on genes, transcripts, proteins, and regulatory sequences.	[5]	https://www.ensembl.org/info/docs/tools/vep/index.html
AnnotSV	Helps interpret the potential pathogenicity of genomic variants and provides a ranking based on ACMG classification.	[6]	https://lbgi.fr/AnnotSV/

According to Table **1**, all tools imply different output data, from TADs to ranked variants, and take input data in different formats, which eventually complicates cross-interpretation. ClinTAD allows for discovering topologically associated domains in CNVs from one or multiple patients. Researchers determine the CNV pathogenicity relying on data on gene and protein changes, which may not be sufficient for the complete picture. Several studies showed that CNVs might also change chromatin architecture, leading to different disorders [7, 8]. In case a CNVs is located inside a TAD, alterations of transcriptional regulation may occur. TADs might significantly assist in deciding on the potential pathogenicity of small CNVs, which is one of the major issues in clinical practice. Understanding a biological basis for CNV consequences is also of great importance. Therefore, ClinTAD may be considered a helpful tool for increasing the depth of CNV interpretation. The tool is, however, a separate instrument not integrated into a CNV evaluation system, and its use for the analysis of many variations at a time does not seem convenient. CNVxplorer, to our mind, is the most comprehensive online CNV evaluation tool. It aggregates a wide variety of data on one page, with a user-friendly design helping in a fast grasp of a CNV consequence. It includes regulatory regions overlap and pathway analysis (*via* KEGG or Reactome data), as well as functional annotations data (*via* Gene Ontology). However, the tool is more suitable for clinical use, with OMIM genes being restricted to monogenic Mendelian disease genes, leaving aside ‘somatic’ and ‘complex’ descriptions, and DECIPHER disease genes restricted to those labeled as ‘confirmed’ [3]. CNVxplorer also limits the number of variants that can be uploaded to 200. The Ensembl Variant Effect Predictor provides a thorough analysis of variant call format (VCF) files, and its web version is presented with a simple, user-friendly interface. The program is ideal for a small-scale analysis. The resulting VEP data does not include genomic pathways, expression or gene imprinting information. AnnotSV shows ACMG class, overlap with regulatory elements, and pathogenic and benign structural variations. The resource is easy to use but lacks at least pathway, gene expression, and exon-intron data, which may be of great importance for a CNV interpretation.

3. CNV INTERPRETATION ALGORITHMS: POSSIBILITIES AND LIMITATIONS

CNV interpretation software is based on specific interpretation algorithms. In order to consider the software advantages and disadvantages, and to provide reproducible analysis, understanding the underlying algorithm is required. The algorithms have changed dramatically throughout the last decade. For example, a paper from 2009 suggested considering a CNV common when it was reported at least twice in the DGV (Database of Genomic Variants), and the record had a complete overlap with a variation in question [9]. Currently, a CNV is thought to be rare if found in <1% of the population. The most recent recommendations of the American College of Medical Genetics and Genomics were released in 2020, covering enhancer, exon/intron, and literature analysis besides casual disease CNVs overlap [10] (Table **2**). When verifying the algorithm, authors and independent reviewers tested it on 114 CNVs using scoring metrics, with scoring based on statistical background. Consequently, the system has a CNV-centric approach, which may be a limitation of a CNV interpretation method. Considering individual CNVs separate from the rest of a genetic landscape, the patient’s clinical data may deprive clinicians of the information necessary for the interpretation. Other most recent interpretation algorithms and the parameters they consider are presented in Table **2**. Discussing the parameters, it is important to note that most algorithms abundantly use DGV to explore the frequency of an aberration and its pathogenicity in different cohorts. Although it is a strong tool for CNV analysis, there are details that one has to keep in mind. First of all, not every cohort may be completely relevant for CNV interpretation. For example, the database contains newborn studies or records with no sex information. Methods of research are also unlikely to be equally spread through the DGV data, meaning some CNVs (due to discrepancy in probe locations for different methods) will be more abundant than others. Choosing specific DGV tracks might solve this issue; however, the best solution to check for recurrent and, therefore, likely non-pathogenic CNVs is to compare them to a cohort obtained by the same method. Unfortunately, different platforms may have many probes in non-overlapping locations:

Table 2. CNV interpretation algorithms described in different articles. "Yes" means that exact or similar data is retrieved, "No" means that the step in question is not mentioned in the article.

Refs.	Recurrency Within Method	Use DGV	Enhancers Downstream/Upstream	Inherited/*De novo*	Genes/CNV Associated with Disease	Gene Expression	Pathways	Imprinting	Introns/Exons	Literature
[2]	No	Yes	No	No	Yes	No	No	No	No	No
[9]	No	Yes	No	Yes	Yes	No	No	No	No	No
[10]	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
[18]	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
[19]	No	Yes	No	Yes	Yes	No	No	No	Yes	Yes
[20]	No	Yes	No	Yes	Yes	No	No	No	No	Yes
[21]	No	Yes	No	No	Yes	No	No	No	No	Yes
[22]	Yes	No	No	Yes	Yes	No	No	No	No	No

for example, the genomic coordinates of our cohort obtained with Affymetrix Cytoscan HD [11] poorly overlap with the data from CytoScan™ XON tool. Since the comparison of the coordinates is usually direct, even a couple of nucleotide differences will yield discrepant results. However, only two algorithms purposedly iterate within in-house databases obtained using a certain platform. Inheritance data are considered in more than half of the algorithms mentioned in Table **2**. To retrieve inheritance data, a trio analysis must be completed. Despite being recommended by the latest American College of Medical Genetics and Genomics guidelines, genetic analyses for medical genomics purposes are not covered by insurance in many countries, neither for the patient nor for the parents. Since even one molecular karyotyping analysis is expensive, trio data are often unavailable. On the other side, the trio analyses may not be as necessary as they are described for the establishment of a CNV status. Often, one CNV is not enough for the disease to manifest (two/multiple hit model) or a smaller CNV in a parent is not causative, whereas an expanded variation in a child affects more genetic material and leads to a disorder [12]. Furthermore, inherited CNVs can have variable penetrance, and considering them benign due to their inherited nature would be incorrect. Gene expression may significantly help understanding if the gene in question is active in a tissue demonstrating pathologic changes. Pathway data indicate how variome is connected to different biological processes. Although almost no algorithms analyze if genes inside the CNV are imprinted, we find it necessary since known imprinting disorder-like phenotypes may arise in these cases [13-15]. In the case of CNVs that affect a small number of genes, analyzing the affected introns and exons is of high importance. One more appealing tendency used in many algorithms is checking if a certain gene is intolerant to a loss of function. Different resources offer probability scores. For example, ExAC Browser (https://gnomad.broadinstitute.org/) shows a pLI score and ClinGen (https: //www.ncbi.nlm.nih.gov/ projects/dbvar/clingen/index.shtml) shows dosage sensitivity scores.

CONCLUSION

Despite numerous changes in interpretation algorithms over the last decade, there is still place for growth. First, clear and user-friendly interpretation software is still missing. Ideally, we see it not only as a scoring system but as a tool for visually presenting and merging the existing data, able to help analyze individual and group variations with the possibility of including clinical features in the interpretation. Indeed, a multidisciplinary dialogue problem arises since essential patient descriptions may be ambiguous, limiting a referral only to a predicted diagnosis. Furthermore, previous genetic analyses, such as karyotype, are to be presented. Regardless of being costly and time-consuming, validation of CNVs is necessary, especially if the analysis is performed using a low-resolution platform. When interpreting the microarray results, it is important to remember that a presence of a large CNV (>0.5 Mbp) with condition-relevant gene contents should not be a reason to dismiss other variations. Coinciding diagnoses are not rare [16], and smaller CNVs may result in additional phenotypic features that are not characteristic of the main syndrome [17].

It is necessary to remember that the final goal of the interpretation is to help clinical specialists not only with the prognosis for the patient but also with the treatment, even in cases when it is considered impossible, such as aneuploidy syndromes. Therefore, the interpretation process should include genomic networks (pathways), such as the ones mentioned in Gene Ontology and KEGG. When multiple genes are affected, it is impossible to consider treatment that would restore the function of every single one, but plausible to find genomic networks (biological processes) undermined by genomic changes and target them with therapy [18, 23, 24, 25]. Thus, we consider it necessary to not only create a user-friendly open-source CNV interpretation software but also to expand the existing algorithms and include pathway information as their core part.

LIST OF ABBREVIATIONS

CNVs	=	Copy Number Variations
Mbp	=	Million base pairs
SNP	=	Single Nucleotide Polymorphism
NGS	=	Next-Generation Sequencing
ACMG	=	American College of Medical Genetics
DGV	=	Database of Genomic Variants

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

The study is partially supported by the Government Assignment of the Russian Ministry of Health, Assignment no. 121031000238–1 (Veltischev Research and Clinical Institute for Pediatrics of the Pirogov Russian National Research Medical University) and by the Government Assignment of the Russian Ministry of Science and Higher Education, Assignment no. AAAA-A19-119040490101–6 (Mental Health Research Center).

CONFLICT OF INTEREST

Dr. Ivan Y Iourov is the regional editor for the journal CBIO.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Pös O, Radvanszky J, Styk J, *et al.* Copy number variation: Methods and clinical applications. Appl Sci (Basel) 2021; 11(2): 819. http://dx.doi.org/10.3390/app11020819

[2] Magini P, Scarano E, Donati I, *et al.* Challenges in the clinical interpretation of small de novo copy number variants in neurodevelopmental disorders. Gene 2019; 706: 162-71. http://dx.doi.org/10.1016/j.gene.2019.05.007 PMID: 31085274

[3] Requena F, Abdallah HH, García A, *et al.* CNVxplorer: A web tool to assist clinical interpretation of CNVs in rare disease patients. Nucleic Acids Res 2021; 49(W1): W93-W103. http://dx.doi.org/10.1093/nar/gkab347 PMID: 34019647

[4] Spector JD, Wiita AP. A guide to using ClinTAD for interpretation of DNA copy number variants in the context of topologically associated domains. Curr Protoc Hum Genet 2020; 108(1): e106. http://dx.doi.org/10.1002/cphg.106 PMID: 33170544

[5] McLaren W, Gil L, Hunt SE, *et al.* The ensembl variant effect predictor. Genome Biol 2016; 17(1): 122. http://dx.doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

[6] Geoffroy V, Herenger Y, Kress A, *et al.* AnnotSV: An integrated tool for structural variations annotation. Bioinformatics 2018; 34(20): 3572-4. http://dx.doi.org/10.1093/bioinformatics/bty304 PMID: 29669011

[7] Lupiáñez DG, Kraft K, Heinrich V, *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 2015; 161(5): 1012-25. http://dx.doi.org/10.1016/j.cell.2015.04.004 PMID: 25959774

[8] Franke M, Ibrahim DM, Andrey G, *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature 2016; 538(7624): 265-9. http://dx.doi.org/10.1038/nature19800 PMID: 27706140

[9] Buysse K, Delle Chiaie B, Van Coster R, *et al.* Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. Eur J Med Genet 2009; 52(6): 398-403. http://dx.doi.org/10.1016/j.ejmg.2009.09.002 PMID: 19765681

[10] Riggs ER, Andersen EF, Cherry AM, *et al.* 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 2020; 22(2): 245-57. http://dx.doi.org/10.1038/s41436-019-0686-8 PMID: 31690835

[11] Iourov IY, Vorsanova SG, Yurov YB, *et al.* The cytogenomic "theory of everything": Chromohelkosis may underlie chromosomal instability and mosaicism in disease and aging. Int J Mol Sci 2020; 21(21): 8328. http://dx.doi.org/10.3390/ijms21218328 PMID: 33171981

[12] Iourov IY, Vorsanova SG, Yurov YB. *In silico* molecular cytogenetics: A bioinformatic approach to prioritization of candidate genes and copy number variations for basic and clinical genome research. Mol Cytogenet 2014; 7(1): 98. http://dx.doi.org/10.1186/s13039-014-0098-z PMID: 25525469

[13] Iourov IY, Vorsanova SG, Korostelev SA, Zelenova MA, Yurov YB. Long contiguous stretches of homozygosity spanning shortly the imprinted loci are associated with intellectual disability, autism and/or epilepsy. Mol Cytogenet 2015; 8(1): 77. http://dx.doi.org/10.1186/s13039-015-0182-z PMID: 26478745

[14] Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: Windows into population history and trait architecture. Nat Rev Genet 2018; 19(4): 220-34. http://dx.doi.org/10.1038/nrg.2017.109 PMID: 29335644

[15] Szpiech ZA, Mak ACY, White MJ, *et al.* Ancestry-dependent enrichment of deleterious homozygotes in runs of homozygosity. Am J Hum Genet 2019; 105(4): 747-62. http://dx.doi.org/10.1016/j.ajhg.2019.08.011 PMID: 31543216

[16] Iourov IY, Vorsanova SG, Yurov YB. The variome concept: Focus on CNVariome. Mol Cytogenet 2019; 12(1): 52. http://dx.doi.org/10.1186/s13039-019-0467-8 PMID: 31890032

[17] Wyandt HE, Wilson GN, Tonk VS. Human chromosome variation: Heteromorphism, polymorphism and pathogenesis. Springer Singapore 2017; pp. 235-417. http://dx.doi.org/10.1007/978-981-10-3035-2_10

[18] Zelenova MA, Yurov YB, Vorsanova SG, Iourov IY. Laundering CNV data for candidate process prioritization in brain disorders. Mol Cytogenet 2019; 12(1): 54. http://dx.doi.org/10.1186/s13039-019-0468-7 PMID: 31890034

[19] Nowakowska B. Clinical interpretation of copy number variants in the human genome. J Appl Genet 2017; 58(4): 449-57. http://dx.doi.org/10.1007/s13353-017-0407-4 PMID: 28963714

[20] Khelifa HB, Soyah N, Labalme A, *et al.* Genomic microarray in intellectual disability: The usefulness of existing systems in the interpretation of copy number variation. J Pediatr Genet 2017; 6(02): 084-91.

[21] Hollenbeck D, Williams CL, Drazba K, *et al.* Clinical relevance of small copy-number variants in chromosomal microarray clinical testing. Genet Med 2017; 19(4): 377-85. http://dx.doi.org/10.1038/gim.2016.132 PMID: 27632688

[22] Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ. Genome-wide arrays: Quality criteria and platforms to be used in routine diagnostics. Hum Mutat 2012; 33(6): 906-15. http://dx.doi.org/10.1002/humu.22076 PMID: 22415865

[23] Ghulam A, Lei X, Guo M, Bian C. Comprehensive analysis of features and annotations of pathway databases. Curr Bioinform 2021; 15(8): 803-20.

http://dx.doi.org/10.2174/1574893615999200413123352

[24] Iourov IY, Vorsanova SG, Yurov YB. Pathway-based classification of genetic diseases. Mol Cytogenet 2019; 12(1): 4. http://dx.doi.org/10.1186/s13039-019-0418-4 PMID: 30766616

[25] Iourov IY, Vorsanova SG, Voinova VY, Yurov YB. 3p22.1p21.31 microdeletion identifies CCK as Asperger syndrome candidate gene and shows the way for therapeutic strategies in chromosome imbalances. Mol Cytogenet 2015; 8(1): 82. http://dx.doi.org/10.1186/s13039-015-0185-9 PMID: 26523151

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.