COMMENT

ParseCNV2: a versatile and integrated tool for copy number variation association studies

Tze Y. Lim[1], Miguel Verbitsky[1] and Simone Sanna-Cherchi[1]✉

European Journal of Human Genetics (2023) 31:275–277; https://doi.org/10.1038/s41431-022-01280-x

Genomic structural variants (SVs), including copy number variations (CNVs) and copy neutral variation, are understudied classes of genomic variations [1, 2]. They can comprise large DNA segments and include multiple genes and regulatory element, and they present in variable structure or copies in comparison to the reference genome. In contrast, single nucleotide variants (SNV), common or rare, identified in family-based or case-control association studies, have been the object of the vast majority of human genetic studies, but alone cannot fully explain the genetic basis of complex diseases [3]. The role of CNVs in human disease predisposition is well established, but their full contribution to this "missing heritability" still remains largely unknown.

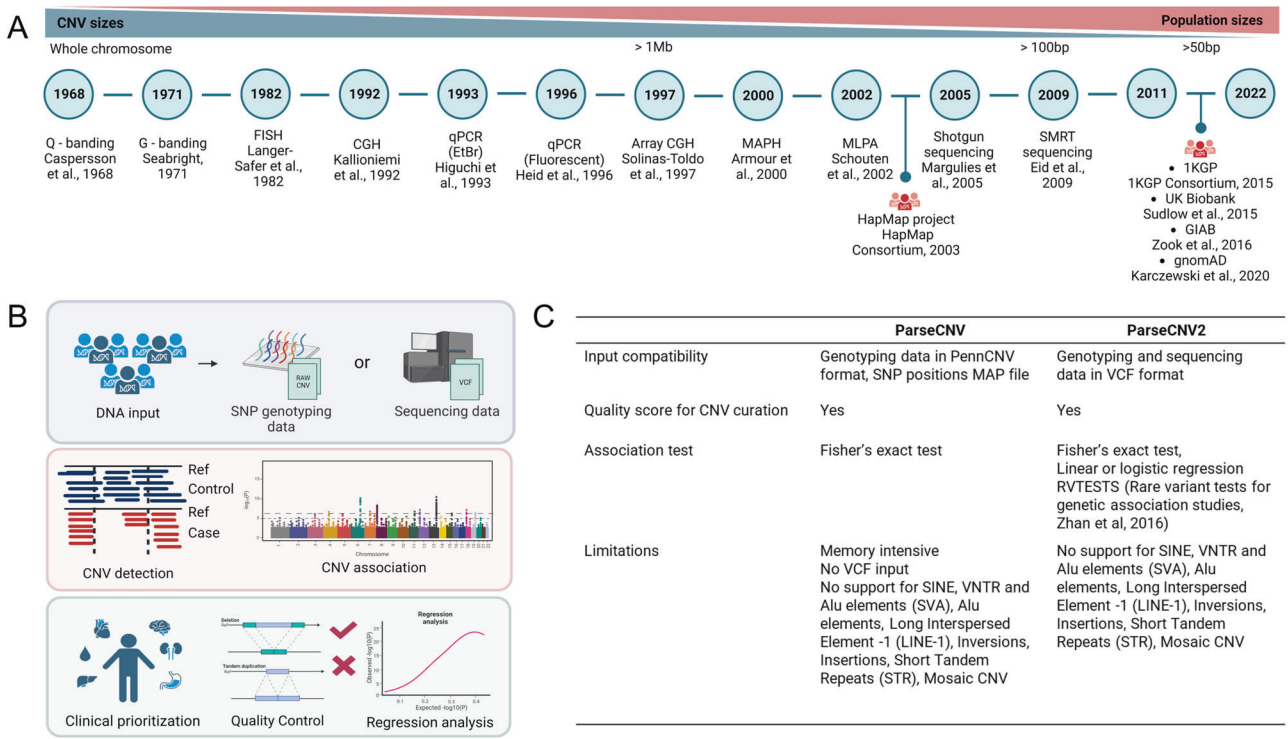
Two of the main reasons for the paucity of studies on SV/CNVs as compared to SNVs are in the difficulty of accurately genotyping CNV regions (CNVr) and in the challenges to effectively conduct an adequately powered, hypothesis-free case-control association study, thus limiting new discoveries. In the past century the field of structural variation identification and analysis has evolved from very simple and targeted approaches with limited resolution to more recent array-based or sequencing-based genome-wide approaches (Fig. 1A). Hence, using clone-based comparative genomic hydbridization (aCGH), high density SNP genotyping array and more recent deep sequencing of the human genome to conduct SV/CNV analysis, we have achieved a far higher resolution but also added significant complexity in data analysis. Modern large-scale case-control studies often require pooling data from different cohorts which may be generated using different technologies (i.e. DNA microarrays, exome or genome sequencing), different capture versions of the same technology, different processing procedures, as well as diverse CNV callers. Therefore, meta-analyses that integrate such diverse datasets accounting for cohort heterogeneity, batch effects, and data compatibility issues present significant challenges. Continuous evolution of methods for association analysis and interpretation is therefore key to fully understand the contribution of structural variants to health and disease.

In this issue, Glessner et al, report on ParseCNV2 [4], a new update of a software that performs CNV association with extra functionalities to natively support genotyping array and sequencing data. The critical additions in ParseCNV2 include: 1) a code revamp to develop a unified CNV variant call format (VCF) parser, as current VCF file format lacks a standard convention for reporting CNVs; 2) new options to perform either linear or logistic regression tests for quantitative and binary traits while adjusting for covariates and an extra option to perform rare variant association tests using the statistical models implemented in RVTESTS in addition to the existing Fisher's exact test; and 3) interactive quality control support for the interpretation of a wider range of CNV calling tools output from sequencing data. One of the key highlights of ParseCNV2 is in its strategic way of merging probe markers or exon boundaries into larger population-level CNVrs. The tool functions by dynamically combining individual-level CNV segments based on similarities in P-values, direction of effect, and distance from adjacent markers or exon regions, rather than relying on a static list of CNVrs. This feature makes ParseCNV2 suitable for analyzing CNVrs derived from different genotyping arrays and, potentially, CNVrs derived from both arrays and sequencing data. While this has not been formally tested in the current manuscript, the potential for seamlessly combining data from multiple sources and technologies without the need for an added step of reformatting represents a significant improvement for the tool. Another important new feature of this suite is that it has a plugin implementation of RVTESTS, which is a comprehensive statistical tests toolkit that allows rare CNV association analysis including logistic Firth regression, variable threshold test, SKAT, burden test, and meta-analysis test, while correcting for issues like cohort heterogeneity in large-cohort studies (Fig. 1B). The new code efficiency and scalability also allow association analyses to be performed for CNV studies with large sample sizes ($N > 100{,}000$) on a desktop computer without the need of computing clusters. Current limitations of ParseCNV2 include the lack of support for analysis of smaller retrotransposons SVA, ALU, LINE elements, inversions, small insertions, short tandem repeats (STRs), and mosaic CNVs (MOS).

ParseCNV2 is the direct evolution from its predecessor, ParseCNV [5], which was developed in 2008 and has been continuously maintained. It was the first known CNV association tool that incorporates quality metrices to help users filter for high confidence associations accounting for factors such as CNV overlapping regions, probe intensities and the number of probes supporting each CNV (Fig. 1C). It is also worth mentioning that ParseCNV2 has been developed by the same creators of PennCNV [6], a CNV detection software that accurately detects CNVs at high-resolution through a Hidden Markov Model-based approach from SNP genotyping data by incorporating the global signal intensity at each SNP (logR ratio), adjacent SNP distance and allele specific intensity (B allele frequency). Unlike many tools at the time,

[1]Division of Nephrology, Department of Medicine, Columbia University, New York, NY, USA. ✉email: ss2517@cumc.columbia.edu

Received: 13 December 2022 Accepted: 19 December 2022
Published online: 11 January 2023



	ParseCNV	ParseCNV2
Input compatibility	Genotyping data in PennCNV format, SNP positions MAP file	Genotyping and sequencing data in VCF format
Quality score for CNV curation	Yes	Yes
Association test	Fisher's exact test	Fisher's exact test, Linear or logistic regression RVTESTS (Rare variant tests for genetic association studies, Zhan et al, 2016)
Limitations	Memory intensive No VCF input No support for SINE, VNTR and Alu elements (SVA), Alu elements, Long Interspersed Element -1 (LINE-1), Inversions, Insertions, Short Tandem Repeats (STR), Mosaic CNV	No support for SINE, VNTR and Alu elements (SVA), Alu elements, Long Interspersed Element -1 (LINE-1), Inversions, Insertions, Short Tandem Repeats (STR), Mosaic CNV

Fig. 1 Structural variant detection and analysis over time and applications and characteristics of ParseCNV2. A Advances in SV/CNV detection methods over the past century and how these technologies enabled CNVs to be extracted at higher resolution and in large cohort studies. **B** Graphical example of a study supported by ParseCNV2 from CNV extraction, association, to downstream analyses. **C** Comparisons between the new ParseCNV2 and its predecessor, ParseCNV. 1KGP The 1000 Genomes Project, CGH comparative genomic hybridization, EtBr Ethidium Bromide, FISH Fluorescence in situ hybridization, GIAB Genome in a Bottle, MAPH Multiplex Amplifiable Probe Hybridization, MLPA Multiplex ligation-dependent probe amplification, qPCR quantitative Polymerase Chain Reaction, SMRT Single Molecule Real time, VCF Variant Call Format. Citations for figures are available in the supplementary reference file. Images created with BioRender.com.

PennCNV applied a "six-state" definition which was originally pioneered by QuantiSNP [7] to provide a more precise description of each CNV event ranging from deletions, copy neutral Loss-of-Heterozygosity to duplications. The tool's remarkable performance has supported approximately 2,000 publications in peer reviewed journals indexed in Pubmed, which use PennCNV as a genesis for further software expansion and CNV association studies, conducting case or case-control discoveries across human diseases. These include large-scale studies that implicated rare recurrent CNVs for multiple traits including autism [8], congenital heart disease [9], kidney and urinary tract malformations [10] and many others. While ParseCNV2 does not provide new improved CNV calling algorithms, which are highly needed especially for other forms of SVs detectable from either markers hybridization signals (microarrays) or sequencing data, it focuses on the filtering, cleaning, merging and association testing in CNV dataset, thus solving a myriad of technical problems and offering a robust framework for defining CNV regions and providing a powerful set of statistical tests for associations.

In summary, the recent unprecedented technology advances allowing cost-effective and accurate deep sequencing of human genomes is opening new venues and resolution in size, complexity and allelic frequency for structural variants analysis. This rapid evolution meets with challenges in integration of structural variants generated using different technologies and from different populations, requiring significant efforts in software development. The future of CNV association tools will therefore require easy-to-use bioinformatic platforms that integrate array-derived and sequencing-derived data with efficient curation on all input data. ParseCNV2 is an example of a much-needed effort towards the development of methods that are geared for dataset harmonization and association analyses using datasets from legacy and present technologies.

REFERENCES

1. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of copy-number variants on complex human traits. Am J Hum Genet. 2022;109:647–68.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature 2006;444:444–54.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–53.
4. Glessner JT, Li J, Liu Y, Khan M, Chang X, Sleiman PMA, et al. ParseCNV2: efficient sequencing tool for copy number variation genome-wide association studies. Eur J Hum Genet. 2022. https://doi.org/10.1038/s41431-022-01222-7.
5. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res. 2013;41:e64.
6. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17:1665–74.
7. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an objective Bayes hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35:2013–25.
8. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87:1215–33.
9. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circulation Res. 2014;115:884–96.
10. Verbitsky M, Westland R, Perez A, Kiryluk K, Liu Q, Krithivasan P, et al. The copy number variation landscape of congenital anomalies of the kidney and urinary tract. Nat Genet. 2019;51:117–27.

AUTHOR CONTRIBUTIONS

TYL drafted the manuscript and prepared the figures; MV edited and helped conceptualizing the manuscript; SSC conceptualized the manuscript and conducted the final edits.

FUNDING

SSC is supported by grants (P20 DK116191, RO1 DK103184, and R01 DK115574) from the National Institutes of Health/NIDDK.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-022-01280-x.

Correspondence and requests for materials should be addressed to Simone Sanna-Cherchi.