Check for updates

Incorporating CNV analysis improves the yield of exome sequencing for rare monogenic disorders—an important consideration for resource-constrained settings

Nadja Louw[1]*, Nadia Carstens[1,2], Zané Lombard[1] and
for DDD-Africa as members of the H3Africa Consortium

[1]Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of
Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, [2]Genomics Platform, South
African Medical Research Council, Cape Town, South Africa

Exome sequencing (ES) is a recommended first-tier diagnostic test for many rare monogenic diseases. It allows for the detection of both single-nucleotide variants (SNVs) and copy number variants (CNVs) in coding exonic regions of the genome in a single test, and this dual analysis is a valuable approach, especially in limited resource settings. Single-nucleotide variants are well studied; however, the incorporation of copy number variant analysis tools into variant calling pipelines has not been implemented yet as a routine diagnostic test, and chromosomal microarray is still more widely used to detect copy number variants. Research shows that combined single and copy number variant analysis can lead to a diagnostic yield of up to 58%, increasing the yield with as much as 18% from the single-nucleotide variant only pipeline. Importantly, this is achieved with the consideration of computational costs only, without incurring any additional sequencing costs. This mini review provides an overview of copy number variant analysis from exome data and what the current recommendations are for this type of analysis. We also present an overview on rare monogenic disease research standard practices in resource-limited settings. We present evidence that integrating copy number variant detection tools into a standard exome sequencing analysis pipeline improves diagnostic yield and should be considered a significantly beneficial addition, with relatively low-cost implications. Routine implementation in underrepresented populations and limited resource settings will promote generation and sharing of CNV datasets and provide momentum to build core centers for this niche within genomic medicine.

KEYWORDS

copy number variation, exome sequencing, low-middle-income countries, variant calling, rare disease, monogenic disorders

1 Introduction

Exome sequencing (ES) is a widely used molecular approach and is recommended as a first-tier test for diagnostic purposes for rare monogenic disorders (Stark et al., 2016; Hu et al., 2018; Srivastava et al., 2019). ES identifies variants within coding exonic regions and is predominantly centered around single-nucleotide variant (SNV) discovery. Recent computational advances have made it possible to incorporate copy number variant (CNV) analysis from ES data, making it more practical and cost-effective, especially for disorders where both SNVs and CNVs are involved in disease etiology.

CNVs attribute to the pathogenesis of up to 15% of rare monogenic cases (Truty et al., 2019; Testard et al., 2022) and tend to have a more severe consequence on phenotype compared to SNVs due to their large size and effect on entire coding regions (Park et al., 2019). Progress has been made regarding joint SNV and CNV investigations in low-middle-income countries (LMICs); however, the gold standard for CNV detection remains chromosomal microarray (CMA) despite its inability to detect SNVs or smaller insertions and deletions. Thus, CMA does not facilitate the efficient use of resources when applied exclusively within already resource-limited settings (Park et al., 2019).

ES has proven to be a cost-effective first-tier test in developed countries, predicting a cost saving between $1,484 and $3,242 per diagnosis (Schwarze et al., 2018). Implementing a diagnostic exome is still thought to be higher in LMICs due to the lack of established infrastructure, high cost of reagents, and the need for personnel training (Wiener et al., 2023); however, studies show that traditional genetic testing and pre-ES investigations can cost up to six times more than local ES costs (Cordoba et al., 2018; Masri and Hamamy, 2021). ES as a first-line investigation would thus be beneficial for many patients and a worthwhile investment in a limited resource setting (Wiener et al., 2023). Despite the advances in ES, it is still not routinely used, especially in countries where genetic testing is limited. The overall diagnostic rate of ES is estimated at ~25% (Yang et al., 2013; Lee et al., 2014; Fung et al., 2020); however, yields as high as 36% (Srivastava et al., 2019) and 41% (Dong et al., 2020; Wright et al., 2023) have been reported in patients with rare monogenic developmental disorders. In studies involving consanguineous patients, a yield of up to 86% has been reported (Hiz Kurul et al., 2022).

In this mini review, we present evidence to show that integrating CNV detection tools into a standard ES analysis pipeline should be considered since it is cost-effective, improves diagnostic yield, and shortens the diagnostic odyssey for patients.

2 Bioinformatic considerations and CNV calling tools

Many bioinformatic tools have been developed for the identification of CNVs from genome and ES data. While no additional sequencing costs are involved in the exome CNV analysis, computational costs relating to additional data analysis should be considered. Comparing computational costs of exome CNV tools, the average expected central processing unit usage was 5.68 GHz and an average of 267,55 Mb of space was used for a 11.2 Mb series of datasets with ×100 coverage (Zhao et al., 2020).

Implementing CANOES on 285 samples took ~6 min per sample using a 2.3 GHz central processing unit core (Backenroth et al., 2014) and for CLAMMS, (Packer et al., 2016) an estimate of ~50 MB random-access memory is required per process. The four main approaches to detect CNVs (Figure 1) are read depth-based, paired-end mapping, split read-based, and assembly-based approaches (Zhao et al., 2013). A combination or ensemble approach is also commonly used as none of the methods alone detect all CNVs with high specificity and sensitivity. Here, we will focus on the most recent and most widely used CNV tools (Gabrielaite et al., 2021) divided into categories according to these different detection approaches (Table 1).

2.1 Read depth-based approach

This approach relies on the depth of coverage to estimate the copy number that the genomic region is correlated with. A higher depth of coverage at a specific region indicates a gain, whereas a lower depth of coverage indicates a loss of copy number. This approach performs well in complex genomic regions (Yoon et al., 2009). CLAMMS (Packer et al., 2016), CoNIFER (RRID:SCR_013213) (Krumm et al., 2012), ExomeDepth (RRID:SCR_002663) (Plagnol et al., 2012), XHMM (Fromer et al., 2012), cn.MOPS (RRID:SCR_013036) (Klambauer et al., 2012), and GATK-gCNV (Babadi et al., 2022) are amongst the most recent and often used read depth-based tools. As only the exonic regions are sequenced, some considerations need to be addressed for these tools to function optimally. ES is considered more appropriate for this approach since it has higher coverage than whole genome sequencing. The majority of tools developed to date for the identification of CNVs from ES data are thus based on this approach. When using read depth-based CNV detection, one should take into consideration that most tools require the use of a reference panel of samples. An assumption of the read depth approach is that reads are distributed uniformly across the genome; however, this is not the case for exome sequencing. Reference samples are thus used to control these biases created by regions of variable depth across exons by establishing a baseline for CNV calling, which ensures the accurate detection of CNVs. These samples should ideally be matched in terms of preparation and sequencing platform and even sequencing batch if possible to limit technical biases which might hinder CNV detection. A number of tools require matched case–control samples as inputs; however, many tools use multiple test samples as a cohort to serve as reference samples for the analysis. The number of samples to be used ranges from below ten to hundreds of samples; for instance, cn.MOPS requires a minimum of six samples, whereas XHMM has a minimum of 50 samples. Several read depth-based tools have been developed and implemented on ES data (Krumm et al., 2012; Tan et al., 2014; Pfundt et al., 2017; Zhao et al., 2020; Gordeeva et al., 2021).

2.2 Split read-based approach

This approach detects unmatched read pairs; thus, one read aligns to the reference genome, while the other read fails to map or aligns only partially to the genome. This potentially identifies the
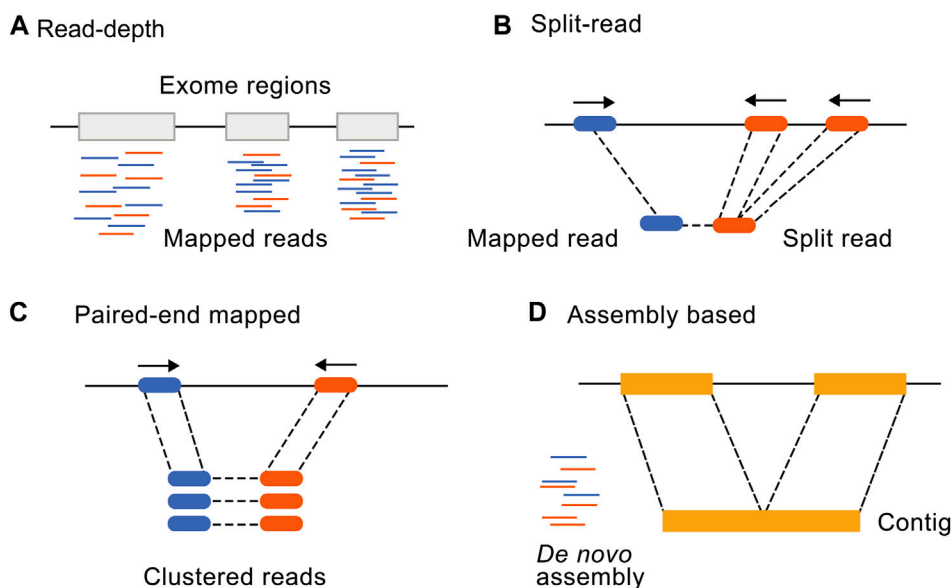
FIGURE 1
Illustration of the four main CNV calling methods from NGS data. Adapted from Zhao et al. (2013), licensed under CC BY 2.0. **(A)** Read depth-based,
(B) Split read-based, **(C)** Paired-end mapping, and **(D)** assembly-based approaches.

breakpoints for CNVs. A few recent tools developed based on this approach are PINDEL (RRID:SCR_000560) (Ye et al., 2018), PRISM (RRID:SCR_005375) (Jiang et al., 2012), SVseq2 (Zhang et al., 2012), and Gustaf (Trappe et al., 2014). One tool developed specifically for ES data is INDELible (Gardner et al., 2021) which was designed to target smaller structural variations (21–500 bp) mostly missed by other CNV calling tools.

2.3 Paired-end mapping approach

This approach was the first approach to put forth the possibility of using next-generation sequencing (NGS) data to detect CNVs (Tuzun et al., 2005; Korbel et al., 2009). It relies on the insert size from the library preparation process and identifies any decreased insert size or swapped read directions between read pairs to identify a CNV or mobile element, insertions, inversions, and tandem duplications. In regions of low complexity containing segmental duplications, this approach seems to be limited. A number of tools have been developed, such as BreakDancer (RRID:SCR_001799) (Fan et al., 2014), HYDRA (RRID:SCR_005260) (Quinlan et al., 2010), PEMer (RRID:SCR_005263) (Korbel et al., 2009), and Ulysses (Gillet-Markowska et al., 2015) being the most widely used (Zhao et al., 2013; Gabrielaite et al., 2021).

2.4 Assembly-based approach

The assembly-based approach assembles reads *de novo* and does not align to a reference genome. Overlapping reads are assembled, and these contigs are then compared to the reference genome, identifying regions with contradictory copy numbers. A minimum read coverage is required for tools based on this

approach to be used successfully. The most commonly used assembly-based tool is Magnolya (RRID:SCR_000164) (Nijkamp et al., 2012).

2.5 The ensemble approach

None of the abovementioned methods alone detects the full spectrum of CNVs with high sensitivity and specificity, and thus it is recommended to use an ensemble approach. In this regard, several tools have been developed to integrate multiple approaches and increase performance. These include DELLY (RRID:SCR_004603) (Rausch et al., 2012), LUMPY (RRID:SCR_003253) (Layer et al., 2014), Manta (Chen et al., 2016), CNVer (RRID:SCR_010820) (Medvedev et al., 2010), and GenomeSTRiP (Handsaker et al., 2011). Although this approach is recommended, there is still no gold standard for CNV detection, especially from ES data. A review of recent publications making use of these bioinformatic tools will thus provide a clearer indication of what approach to consider when calling CNVs from ES data.

3 Best approaches for CNV calling from ES data

As there are many tools available for CNV detection from ES data, recommendations have been made focusing on the use of these tools for optimal results. In a recent comparative analysis of ES-focused CNV tools (Zhao et al., 2020), the recommendations for obtaining the best results were related to the specific dataset. In terms of accuracy, it was recommended to use CNVkit (Talevich et al., 2016) if CNV size is small (<100 kb), whereas cn.MOPS seems to be optimal if CNV size is larger. If the dataset presents with more

TABLE 1 Summary of bioinformatic tools for CNV detection using next-generation sequencing data.

Tool	Method	Advantages/Attributes	Reference
CANOES	Read depth	Rare CNVs from ES data	Backenroth et al. (2014)
CLAMMS	Read depth	Large population ES studies and easy integration into automated variant-calling pipeline	Packer et al. (2016)
CoNIFER	Read depth	More insertions identified and low memory required. ES-specific tool	Krumm et al. (2012)
ExomeDepth	Read depth	Good control of technical variability between samples and effective across a wider range of exome datasets. Small and heterozygous deletions	Plagnol et al. (2012)
XHMM	Read depth	Explores novel classes of CNVs from ES data	Fromer et al. (2012)
Cn.MOPS	Read depth	Larger CNVs identified from ES and WGS data. Fast average running time	Klambauer et al. (2012)
GATK-gCNV	Read depth	Detects rare CNVs and determines copy number biases and CNVs from WGS and ES data	Babadi et al. (2022)
HMZDelFinder/HMZDelFinder _opt	Read depth	Rare, intragenic homozygous and hemizygous deletions from ES data	Gambin et al. (2017)
EXCAVATOR2	Read depth	Detects CNVs from ES data with genome-wide resolution	D'aurizio et al., 2016
CODEX	Read depth	No matched normal controls required. Designed for ES data.	Jiang et al. (2012)
CN-Learn	Read depth	Small sample size needed, good precision, and high-confidence CNVs from ES data	Pounraja et al. (2019)
CNVkit	Read depth	<100 kb CNVs from WGS and ES data and more effective for deletions	Talevich et al. (2016b)
PINDEL	Split read	Large deletions and medium-sized insertions from WGS	Ye et al. (2018)
PRISM	Split read	SVs and precise breakpoints from WGS	Jiang et al. (2012)
SVseq2	Split read	Indel from low coverage WGS data and exact breakpoints	Zhang et al. (2012)
Gustaf	Split read	Identifies size and location of dispersed duplications and translocations from WGS data; 30–100 bp and >500 bp	Trappe et al. (2014)
INDELible	Split read	Smaller SVs (20–500 bps) from ES data missed by conventional methods	Gardner et al. (2021)
BreakDancer	Paired-end	Various SVs including indels, inversions, and translocations from WGS	Chen et al. (2016)
HYDRA	Paired-end	Diverse classes of SVs, including those involving repetitive elements such as transposons and segmental duplications from WGS	Quinlan et al. (2010)
PEMer	Paired-end	~3 kilobases or larger SVs from WGS	Korbel et al. (2009)
Ulysses	Paired-end	High specificity over a complete spectrum of variants	Gillet-Markowska et al. (2015)
Magnolya	Assembly based	No mapping of reads to reference genome and de novo CNV detection	Nijkamp et al. (2012)
DELLY	Ensemble approach	Full spectrum of genomic rearrangements from WGS, including complex events	Rausch et al. (2012)
LUMPY	Ensemble approach	Increased sensitivity of SV detection from WGS	Layer et al. (2014)
Manta	Ensemble approach	Less computational time/space, intense large-scale SVs, medium-sized indels, and large insertions from WGS and ES data	Chen et al. (2016)
GenomeSTRiP	Ensemble approach	Whole-genome discovery and genotyping of deletions	Handsaker et al. (2011)
CNVer	Ensemble approach	Better mitigates the sequencing biases causing uneven local coverage and accurately predicts CNVs	Medvedev et al. (2010)

*Whole-genome sequencing (WGS).
*Structural variation (SV).

insertions, using CoNIFER is recommended, and CNVkit is seemingly the best for identifying deletions. If there is no prior knowledge of the dataset, then using cn.MOPS and CoNIFER together is recommended (Guo et al., 2013; Zhao et al., 2020).

Different tools have been designed to obtain optimal sensitivity and specificity focused on rare or common CNVs as well as population size. Previous limitations, for instance, only being able to identify CNVs spanning at least two or more exons, GC content, or mappability biases as well as sequencing noise have been addressed, and many tools have been developed to try and overcome these difficulties (Pounraja et al., 2019; Bigio et al., 2020; Filer et al., 2021; Babadi et al., 2022). CLAMMS was developed to be more suitable for large population studies (Packer et al., 2016) and integrated more easily into an automated variant calling pipeline. In order to more accurately identify rare and intragenic homozygous and hemizygous deletions, HMZDelFinder (Gambin et al., 2017) was developed, and the newer HMZDelFinder_opt (Bigio et al., 2020) outperforms the older version in terms of accuracy and specifically identifies partial exon deletions. ExomeDepth has also been widely used (Marchuk et al., 2018; Rajagopalan et al., 2020; Zhai et al., 2021) and designed to control technical variability between samples. CANOES (Backenroth et al., 2014) is complementary to methods like XHMM and CoNIFER, and accuracy can be improved when using CANOES in combination with one of these methods. CN-Learn identifies true CNVs with higher precision and recall rates without compromising performance even with as little as 30 samples (Pounraja et al., 2019). This tool uses CNVs predicted by four different CNV callers (CANOES, CODEX, XHMM, and CLAMMS) which were found to enhance performance instead of using the tools as standalone methods. Another study also merged results from CANOES and HMZDelFinder after each tool was applied separately (Dong et al., 2020). It was also suggested to combine GATK-gCNV, LUMPY, DELLY, and cn.MOPS which have the best recall and capture different CNVs (Gabrielaite et al., 2021). While LUMPY and DELLY have been developed for whole-genome sequencing data, GATK-gCNV and cn.MOPS should be used with ES data. In a recent study, CNVkit, XHMM, EXCAVATOR2, and ExomeDepth were used for ES-based CNV calling in order to maximize the sensitivity and make ES a more powerful tool to diagnose neurodevelopmental disorders (Zhai et al., 2021).

It is clearly demonstrated that the ensemble approach yields optimal results while increasing the sensitivity and specificity of CNV detection (Välipakka et al., 2020). Individual implementation strategies could still be helpful and lead to an increased diagnostic yield but is largely influenced by the available computing infrastructure in specific environments as well as adequate representation of the different calling strategies. CNV calling from ES data should be particularly attractive in resource-constrained settings with reduced capital expenditure and required infrastructure.

4 Value of CNV calling from ES in resource-constrained countries

In a recent study (Dong et al., 2020), an overall yield of 41.4% was reported by the simultaneous analysis of SNV and CNV, of which 12.0% can be attributed to CNVs. Another study based in China found that SNV and exome-based CNV calling yielded an overall diagnostic rate of 58.8%, of which diagnostic CNVs accounted for 17.6% (Xiang et al., 2021). A comprehensive method was used for CNV identification which included combining XHMM and principal component analysis with CNVKit. Similarly, it was found that incorporating exome-based CNV detection into conventional SNV analysis for a single trio-ES test significantly improved the diagnostic rate (Zhai et al., 2021). When combining SNV and CNV analyses, an overall diagnostic yield of 54% was obtained, which included 18.9% from CNV analysis alone. CNVs in this study were detected using CNVkit, XHMM, EXCAVATOR2, and ExomeDepth, which were all retained and annotated thereafter.

In an effort to identify the cause of congenital heart disease in 96 child participants from Nigeria, a combined approach was applied by making use of ES from patient and parents (where available) and performing XHMM CNV analysis on the data (Ekure et al., 2021). Assessing the genomic etiology of autism spectrum disorder in India, a diagnostic yield of 29.7% of individuals in total was obtained for exome sequencing, of which CNVs contributed 3%, and interestingly, CMA analysis carried out on the same cohort yielded a diagnostic rate of 2.9% (Sheth et al., 2023). Thus, combined CNV and SNV analysis from ES data significantly increased the diagnostic yield versus only using CMA (29.7% vs. 2.9%). The combined SNV and CNV analysis from the discussed literature has been shown to increase the diagnostic yield by as much as 18% (Figure 2), which is an additional diagnosis for ~2 out of every 100 individuals. The average increased yield attributed to CNVs from the discussed research is 10.7% without additional testing costs involved. Therefore, implementing ES as a first tier for diagnosis, especially when incorporating CNV analysis, should be considered because it is efficient and cost-effective and shortens the diagnostic odyssey for patients who would not have otherwise necessarily received a molecular diagnosis.

As is the case for most resource-limited settings, the cost of sequencing a trio and availability of both parents are always the important limiting factors. A study carried out in India (Pranav Chand et al., 2023) on children with neurodevelopmental delay found that a proband-only ES approach obtained a diagnostic yield of 31.5% of these children. Addition of parental samples increased this yield by only 3%, and CNVs contributed to 6.5% of the diagnoses. Another study conducted in China had an overall diagnostic rate of 28.8% after analyzing 1,323 pediatric patients, which proved to be a relatively efficient and cost-effective approach in a developing country (Hu et al., 2018). A South African study (Moosa et al., 2022) found that proband-only ES is a very valuable tool for diagnosis, especially if CNV analysis is included. A diagnostic yield of 51% was obtained with 46% of patients presenting with SNVs and 5% with CNVs. Even though trio-ES has been shown to have the best outcome for a positive diagnosis (Wright et al., 2023), proband-only exome analysis has proven to be a feasible option for diagnosis in settings with limited resources or difficulty in obtaining parental samples.

Another advantage of identifying CNVs in underrepresented populations is the expansion of variant representation in predominantly European-focused public data repositories. Recent
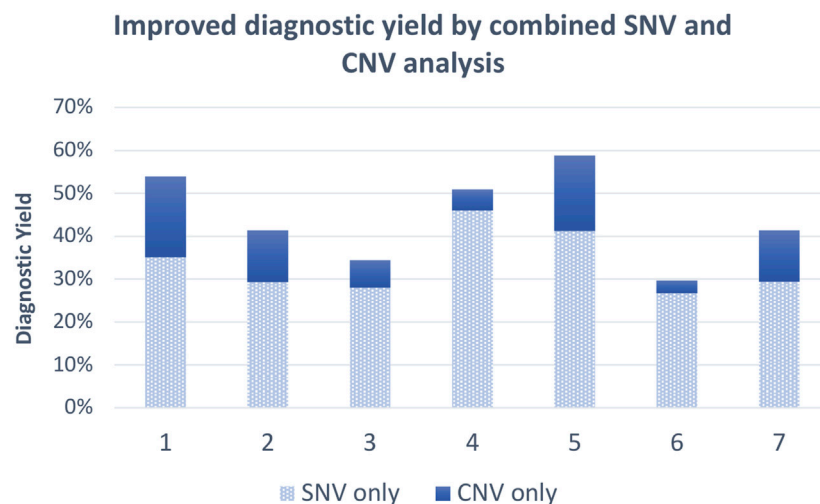
FIGURE 2
Average diagnostic yield of exome SNVs and combined SNVs/CNVs from 1. Zhai et al. (2021), 2. Truty et al. (2019), 3. Pranav Chand et al. (2023), 4. Moosa et al. (2022), 5. Xiang et al. (2021), 6. Sheth et al. (2023), and 7. Dong et al. (2020).

progress has been made to contribute CNVs from African population groups to variant databases (Nyangiri et al., 2020; Romdhane et al., 2021; Yilmaz et al., 2021) as the lack of diversity of high-quality genomic data, specifically from Africa, hampers the implementation of appropriate genetic services and brings forth healthcare inequalities (Baine-Savanhu et al., 2023). The lack of representation in population frequency databases has also made clinical interpretation and classification of CNVs more challenging in LMICs. Standardized CNV reporting is possible by using specific ACMG and Clinical Genome (ClinGen) Resource guidelines for CNV classification (Riggs et al., 2020), but careful evaluation of CNVs is encouraged to ensure that only likely disease-causing CNVs matching the patient phenotype are reported. Resolving VUSs remains challenging if population representation is inadequate. Furthermore, additional investigations including validation and functional experiments are often not available in LMIC laboratories. Distinguishing benign CNVs from pathogenic CNVs can be challenging, and thus a number of tools have been developed to enable a more convenient manner of CNV annotation and interpretation. These tools provide support for annotation and/ or classification of CNVs, and many tools are web-based, easy to use, and freely available. A recent review have summed up these tools comprehensively to make it easier for clinicians, laboratory scientists, and genetic counselors to make a decision as to which tool would work best in their setting (Pös et al., 2021).

5 Discussion

Overall, simultaneous analysis of CNVs and SNVs through ES shows potential as a first-tier investigation for diagnosing rare monogenic disorders. Novel candidate genes and variants have been identified, representing the first step in genomic studies within understudied populations. The diagnostic yield of the current gold standard for CNV detection (CMA) is 15%–20%

(Miller et al., 2010), which is significantly lower than ES. In a recent scoping review, it was shown that ES for diagnosing neurodevelopmental disorders outperforms CMA by 10%–28% (Srivastava et al., 2019), further supporting the combined SNV and CNV analysis approach from ES data. Although it would be ideal to validate exome CNVs with methods such as microarray, it is costly and often not feasible in LMICs. Accurate CNV calling incorporating thorough quality control can help limit false-positive and false-negative results. This is evidenced by eliminating the need for Sanger sequencing validation of SNVs when proper quality control is carried out (Strom et al., 2014). It is also important to note that ES CNV tools have limitations due to their inability to detect specific types of variations, for instance, balanced structural variants (translocations and inversions), mosaicism, and smaller CNVs (<50 bp). Although analyses and technologies are improving to address these shortcomings, it should be considered when implementing these tools. Whole-genome sequencing and array-based techniques can be used to identify these variations; however, this will incur additional costs. Long-read sequencing has the ability to detect these structural variations as well as SNVs, making it ideal to implement as a single assay to replace all the above methods. At present, this technology is too expensive to use as a first-tier test; however, as costs decrease, this might be a possibility for future consideration.

It is evident that a more diverse reference genome representing a larger range of population groups is required to improve CNV calling and classification. Improved diversity in population frequency databases will also provide access to key data needed for the clinical interpretation of CNVs. Although there is still limited data and genetic services in most of Africa, making it difficult to translate research into clinical healthcare services (Kamp et al., 2021), a current and thorough analysis of the cost-benefit for ES would be beneficial toward motivating the adoption of ES as a first-tier test in resource-constrained environments. This review highlights the need for

incorporating not only efficient but appropriate exome pipelines in LMICs to further implement genomic medicine and make it more attainable for all. A wider adoption of CNV calling from ES data and use over time will allow for more opportunities to achieve this. Reanalysis of data should be considered for patients without a definite diagnosis as this has proven to increase diagnostic yield (Liu et al., 2019; Schuermans et al., 2022). More CNV publications and ClinVar submissions (Landrum et al., 2014; Landrum et al., 2016; Landrum et al., 2017) from understudied populations will expand the size and scope and improve the resolution of clinically relevant CNVs in the public domain. Public data repositories like ClinVar and DECIPHER (Bragin et al., 2014) have contributed to diversifying data; however, more effective production and sharing of genomic datasets are needed to advance genomic medicine globally. Recent initiatives have been established to facilitate African data-sharing and empower health experts by availing tools, training, and coordination to strengthen laboratory and bioinformatic capacity (Mulder et al., 2016; Mulder et al., 2018; Makoni, 2020; Lumaka et al., 2022). International collaborations and training could be crucial to resolve the true impact of CNVs and build strong core groups with expertise, experience, and technical competence to accurately report on CNVs in a diagnostic context within LMICs. CNV calling from existing ES datasets from non-European individuals may therefore be an important analysis to invest in.

Author contributions

NL: conceptualization, visualization, data curation and writing–original draft. NC: conceptualization, funding acquisition, supervision and writing–review and editing. ZL: conceptualization, funding acquisition, supervision and writing–review and editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Babadi, M., Fu, J. M., Lee, S. K., Smirnov, A. N., Gauthier, L. D., Walker, M., et al. (2022). GATK-gCNV: a rare copy number variant discovery algorithm and its application to exome sequencing in the UK biobank. bioRxiv.

Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., et al. (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.* 42, e97. doi:10.1093/nar/gku345

Baine-Savanhu, F., Macaulay, S., Louw, N., Bollweg, A., Flynn, K., Molatoli, M., et al. (2023). Identifying the genetic causes of developmental disorders and intellectual disability in Africa: a systematic literature review. *Front. Genet.* 14, 1137922. doi:10.3389/fgene.2023.1137922

Bigio, B., Seeleuthner, Y., Kerner, G., Migaud, M., Rosain, J., Boisson, B., et al. (2020). Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing. bioRxiv.

Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P., et al. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* 42, D993–D1000. doi:10.1093/nar/gkt937

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer applications. *Bioinformatics* 32, 1220–1222. doi:10.1093/bioinformatics/btv710

Cordoba, M., Rodriguez-Quiroga, S. A., Vega, P. A., Salinas, V., Perez-Maturo, J., Amartino, H., et al. (2018). Whole exome sequencing in neurogenetic odysseys: an effective, cost- and time-saving diagnostic approach. *PLoS One* 13, e0191228. doi:10.1371/journal.pone.0191228

D'aurizio, R., Pippucci, T., Tattini, L., Giusti, B., Pellegrini, M., and Magi, A. (2016). Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* 44, e154.

Dong, X., Liu, B., Yang, L., Wang, H., Wu, B., Liu, R., et al. (2020). Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: a Chinese cohort. *J. Med. Genet.* 57, 558–566. doi:10.1136/jmedgenet-2019-106377

Ekure, E. N., Adeyemo, A., Liu, H., Sokunbi, O., Kalu, N., Martinez, A. F., et al. (2021). Exome sequencing and congenital heart disease in sub-saharan Africa. *Circulation Genomic Precis. Med.* 14, e003108. doi:10.1161/CIRCGEN.120.003108

Fan, X., Abbott, T. E., Larson, D., and Chen, K. (2014). BreakDancer: identification of genomic structural variation from paired-end read mapping. *Curr. Protoc. Bioinforma.* 45 (6), 1–11. doi:10.1002/0471250953.bi1506s45

Filer, D. L., Kuo, F., Brandt, A. T., Tilley, C. R., Mieczkowski, P. A., Berg, J. S., et al. (2021). Pre-capture multiplexing provides additional power to detect copy number variation in exome sequencing. *BMC Bioinforma.* 22, 374. doi:10.1186/s12859-021-04246-w

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* 91, 597–607. doi:10.1016/j.ajhg.2012.08.005

Fung, J. L. F., Yu, M. H. C., Huang, S., Chung, C. C. Y., Chan, M. C. Y., Pajusalu, S., et al. (2020). A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *npj Genomic Med.* 5, 37. doi:10.1038/s41525-020-00144-x

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., et al. (2021). A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers (Basel)* 13, 6283. doi:10.3390/cancers13246283

Gambin, T., Akdemir, Z. C., Yuan, B., Gu, S., Chiang, T., Carvalho, C. M. B., et al. (2017). Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.* 45, 1633–1648. doi:10.1093/nar/gkw1237

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., et al. (2021). Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *Am. J. Hum. Genet.* 108, 2186–2194. doi:10.1016/j.ajhg.2021.09.010

Gillet-Markowska, A., Richard, H., Fischer, G., and Lafontaine, I. (2015). Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics* 31, 801–808. doi:10.1093/bioinformatics/btu730

Gordeeva, V., Sharova, E., Babalyan, K., Sultanov, R., Govorun, V. M., and Arapidi, G. (2021). Benchmarking germline CNV calling tools from exome sequencing data. *Sci. Rep.* 11, 14416. doi:10.1038/s41598-021-93878-2

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., et al. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *Biomed. Res. Int.* 2013, 915636. doi:10.1155/2013/915636

Handsaker, R. E., Korn, J. M., Nemesh, J., and Mccarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* 43, 269–276. doi:10.1038/ng.768

Hiz Kurul, S., Oktay, Y., Töpf, A., Szabó, N. Z., Güngör, S., Yaramis, A., et al. (2022). High diagnostic rate of trio exome sequencing in consanguineous families with neurogenetic diseases. *Brain* 145, 1507–1518. doi:10.1093/brain/awab395

Hu, X., Li, N., Xu, Y., Li, G., Yu, T., Yao, R. E., et al. (2018). Proband-only medical exome sequencing as a cost-effective first-tier genetic diagnostic test for patients without prior molecular tests and clinical diagnosis in a developing country: the China experience. *Genet. Med.* 20, 1045–1053. doi:10.1038/gim.2017.195

Jiang, Y., Wang, Y., and Brudno, M. (2012). PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* 28, 2576–2583. doi:10.1093/bioinformatics/bts484

Kamp, M., Krause, A., and Ramsay, M. (2021). Has translational genomics come of age in Africa? *Hum. Mol. Genet.* 30, R164–R173. doi:10.1093/hmg/ddab180

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40, e69. doi:10.1093/nar/gks003

Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., et al. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 10, R23. doi:10.1186/gb-2009-10-2-r23

Krumm, N., Sudmant, P. H., Ko, A., O'roak, B. J., Malig, M., Coe, B. P., et al. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22, 1525–1532. doi:10.1101/gr.138115.112

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi:10.1093/nar/gkv1222

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2017). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi:10.1093/nar/gkx1153

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi:10.1093/nar/gkt1113

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 15, R84. doi:10.1186/gb-2014-15-6-r84

Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 312, 1880–1887. doi:10.1001/jama.2014.14604

Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., et al. (2019). Reanalysis of clinical exome sequencing data. *N. Engl. J. Med.* 380, 2478–2480. doi:10.1056/NEJMc1812033

Lumaka, A., Carstens, N., Devriendt, K., Krause, A., Kulohoma, B., Kumuthini, J., et al. (2022). Increasing African genomic data generation and sharing to resolve rare and undiagnosed diseases in Africa: a call-to-action by the H3Africa rare diseases working group. *Orphanet J. Rare Dis.* 17, 230. doi:10.1186/s13023-022-02391-w

Makoni, M. (2020). Africa's $100-million pathogen genomics initiative. *Lancet Microbe* 1, e318. doi:10.1016/S2666-5247(20)30206-8

Marchuk, D. S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L. V., Brandt, A., et al. (2018). Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One* 13, e0209185. doi:10.1371/journal.pone.0209185

Masri, A., and Hamamy, H. (2021). Cost effectiveness of whole exome sequencing for children with developmental delay in a developing country: a study from Jordan. *J. Pediatr. Neurology* 20, 020–023. doi:10.1055/s-0040-1722265

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Res.* 20, 1613–1622. doi:10.1101/gr.106344.110

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* 86, 749–764. doi:10.1016/j.ajhg.2010.04.006

Moosa, S., Coetzer, K. C., Lee, E., and Seo, G. H. (2022). Undiagnosed disease program in South Africa: results from first 100 exomes. *Am. J. Med. Genet. A* 188, 2684–2692. doi:10.1002/ajmg.a.62847

Mulder, N., Abimiku, A., Adebamowo, S. N., De Vries, J., Matimba, A., Olowoyo, P., et al. (2018). H3Africa: current perspectives. *Pharmgenomics Pers. Med.* 11, 59–66. doi:10.2147/PGPM.S141546

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., et al. (2016). H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res.* 26, 271–277. doi:10.1101/gr.196295.115

Nijkamp, J. F., Van Den Broek, M. A., Geertman, J. M., Reinders, M. J., Daran, J. M., and De Ridder, D. (2012). *De novo* detection of copy number variation by co-assembly. *Bioinformatics* 28, 3195–3202. doi:10.1093/bioinformatics/bts601

Nyangiri, O. A., Noyes, H., Mulindwa, J., Ilboudo, H., Kabore, J. W., Ahouty, B., et al. (2020). Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics* 21, 289. doi:10.1186/s12864-020-6669-y

Packer, J. S., Maxwell, E. K., O'dushlaine, C., Lopez, A. E., Dewey, F. E., Chernomorsky, R., et al. (2016). CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics* 32, 133–135. doi:10.1093/bioinformatics/btv547

Park, K. B., Nam, K. E., Cho, A. R., Jang, W., Kim, M., and Park, J. H. (2019). Effects of copy number variations on developmental aspects of children with delayed development. *Ann. Rehabil. Med.* 43, 215–223. doi:10.5535/arm.2019.43.2.215

Pfundt, R., Del Rosario, M., Vissers, L., Kwint, M. P., Janssen, I. M., De Leeuw, N., et al. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.* 19, 667–675. doi:10.1038/gim.2016.163

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 28, 2747–2754. doi:10.1093/bioinformatics/bts526

Pös, O., Radvanszky, J., Styk, J., Pös, Z., Buglyó, G., Kajsik, M., et al. (2021). Copy number variation: methods and clinical applications. *Appl. Sci.* 11, 819. doi:10.3390/app11020819

Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N., and Girirajan, S. (2019). A machine-learning approach for accurate detection of copy-number variants from exome sequencing. bioRxiv.

Pranav Chand, R., Vinit, W., Vaidya, V., Iyer, A. S., Shelke, M., Aggarwal, S., et al. (2023). Proband only exome sequencing in 403 Indian children with neurodevelopmental disorders: diagnostic yield, utility and challenges in a resource-limited setting. *Eur. J. Med. Genet.* 66, 104730. doi:10.1016/j.ejmg.2023.104730

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., Hurles, M. E., et al. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* 20, 623–635. doi:10.1101/gr.102970.109

Rajagopalan, R., Murrell, J. R., Luo, M., and Conlin, L. K. (2020). A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med.* 12, 14. doi:10.1186/s13073-020-0712-0

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333–i339. doi:10.1093/bioinformatics/bts378

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257. doi:10.1038/s41436-019-0686-8

Romdhane, L., Mezzi, N., Dallali, H., Messaoud, O., Shan, J., Fakhro, K. A., et al. (2021). A map of copy number variations in the Tunisian population: a valuable tool for medical genomics in North Africa. *npj Genomic Med.* 6, 3. doi:10.1038/s41525-020-00166-5

Schuermans, N., Hemelsoet, D., Terryn, W., Steyaert, S., Van Coster, R., Coucke, P. J., et al. (2022). Shortcutting the diagnostic odyssey: the multidisciplinary program for undiagnosed rare diseases in adults (UD-PrOZA). *Orphanet J. Rare Dis.* 17, 210. doi:10.1186/s13023-022-02365-y

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* 20, 1122–1130. doi:10.1038/gim.2017.247

Sheth, F., Shah, J., Jain, D., Shah, S., Patel, H., Patel, K., et al. (2023). Comparative yield of molecular diagnostic algorithms for autism spectrum disorder diagnosis in India: evidence supporting whole exome sequencing as first tier test. *BMC Neurol.* 23, 292. doi:10.1186/s12883-023-03341-0

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet. Med.* 21, 2413–2421. doi:10.1038/s41436-019-0554-6

Stark, Z., Tan, T. Y., Chong, B., Brett, G. R., Yap, P., Walsh, M., et al. (2016). A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet. Med.* 18, 1090–1096. doi:10.1038/gim.2016.1

Strom, S. P., Lee, H., Das, K., Vilain, E., Nelson, S. F., Grody, W. W., et al. (2014). Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genet. Med.* 16, 510–515. doi:10.1038/gim.2013.183

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016b). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.* 12, e1004873. doi:10.1371/journal.pcbi.1004873

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., et al. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.* 35, 899–907. doi:10.1002/humu.22537

Testard, Q., Vanhoye, X., Yauy, K., Naud, M. E., Vieville, G., Rousseau, F., et al. (2022). Exome sequencing as a first-tier test for copy number variant detection: retrospective evaluation and prospective screening in 2418 cases. *J. Med. Genet.* 59, 1234–1240. doi:10.1136/jmg-2022-108439

Trappe, K., Emde, A.-K., Ehrlich, H.-C., and Reinert, K. (2014). Gustaf: detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics* 30, 3484–3490. doi:10.1093/bioinformatics/btu431

Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L., et al. (2019). Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet. Med.* 21, 114–123. doi:10.1038/s41436-018-0033-5

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., et al. (2005). Fine-scale structural variation of the human genome. *Nat. Genet.* 37, 727–732. doi:10.1038/ng1562

Välipakka, S., Savarese, M., Sagath, L., Arumilli, M., Giugliano, T., Udd, B., et al. (2020). Improving copy number variant detection from sequencing data with a combination of programs and a predictive model. *J. Mol. Diagnostics* 22, 40–49. doi:10.1016/j.jmoldx.2019.08.009

Wiener, E. K., Buchanan, J., Krause, A., Lombard, Z., and DDD-Africa Study, as members of the H3Africa Consortium (2023). Retrospective file review shows limited genetic services fails most patients - an argument for the implementation of exome

sequencing as a first-tier test in resource-constraint settings. *Orphanet J. Rare Dis.* 18, 81. doi:10.1186/s13023-023-02642-4

Wright, C. F., Campbell, P., Eberhardt, R. Y., Aitken, S., Perrett, D., Brent, S., et al. (2023). Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* 388, 1559–1571. doi:10.1056/NEJMoa2209046

Xiang, J., Ding, Y., Yang, F., Gao, A., Zhang, W., Tang, H., et al. (2021). Genetic analysis of children with unexplained developmental delay and/or intellectual disability by whole-exome sequencing. *Front. Genet.* 12. doi:10.3389/fgene.2021.738561

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* 369, 1502–1511. doi:10.1056/NEJMoa1306555

Ye, K., Guo, L., Yang, X., Lamijer, E. W., Raine, K., and Ning, Z. (2018). Split-read indel and structural variant calling using PINDEL. *Methods Mol. Biol.* 1833, 95–105. doi:10.1007/978-1-4939-8666-8_7

Yilmaz, F., Null, M., Astling, D., Yu, H.-C., Cole, J., Santorico, S. A., et al. (2021). Genome-wide copy number variations in a large cohort of Bantu African children. *BMC Med. Genomics* 14, 129. doi:10.1186/s12920-021-00978-z

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi:10.1101/gr.092981.109

Zhai, Y., Zhang, Z., Shi, P., Martin, D. M., and Kong, X. (2021). Incorporation of exome-based CNV analysis makes trio-WES a more powerful tool for clinical diagnosis in neurodevelopmental disorders: a retrospective study. *Hum. Mutat.* 42, 990–1004. doi:10.1002/humu.24222

Zhang, J., Wang, J., and Wu, Y. (2012). An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinforma.* 13, S6. doi:10.1186/1471-2105-13-S6-S6

Zhao, L., Liu, H., Yuan, X., Gao, K., and Duan, J. (2020). Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinforma.* 21, 97. doi:10.1186/s12859-020-3421-1

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinforma.* 14, S1. doi:10.1186/1471-2105-14-S11-S1