Copy-number Analysis by Base-level Normalization (CABANA): An Intuitive Visualization Tool for Confirming True Copy Number Variations

Hongkyung Kim

Department of Laboratory Medicine, Yonsei University College of Medicine, Severance Hospital

Yeeun Shim

Brain Korea 21 PLUS Project for Medical Science, Yonsei University

Taek Gyu Lee

Brain Korea 21 PLUS Project for Medical Science, Yonsei University

Dongju Won

Department of Laboratory Medicine, Yonsei University College of Medicine, Severance Hospital

Jong Rak Choi

Department of Laboratory Medicine, Yonsei University College of Medicine, Severance Hospital

Saeam Shin (✉ saeam0304@yuhs.ac)

Department of Laboratory Medicine, Yonsei University College of Medicine, Severance Hospital

Seung-Tae Lee

Department of Laboratory Medicine, Yonsei University College of Medicine, Severance Hospital

Research Article

Keywords:

Posted Date: February 4th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1292720/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Next-generation sequencing (NGS) facilitates comprehensive molecular analyses, allowing accurate diagnosis of unsolved disorders. In addition to detecting single-nucleotide variations and small insertions/deletions, bioinformatics tools can identify copy number variations (CNVs) in NGS data, which improves the diagnostic yield. However, due to the possibility of false positives, subsequent confirmation tests are generally performed. Here, we introduce Copy-number Analysis by BAsE-level NormAlization (CABANA), a computational tool that can intuitively visualize true CNVs using the normalized single-base-level read depth calculated from NGS data. To demonstrate how CABANA works, NGS data were obtained from 474 patients with neuromuscular disorders. CNVs were screened using a conventional bioinformatics tool, ExomeDepth, and then we normalized and visualized those data at the single-base level using CABANA, followed by manual inspection by geneticists to exclude false-positives. In this way, we identified 31 true-positive CNVs (7%) in 474 patients and subsequently confirmed all of them to be true using multiplex ligation-dependent probe amplification. The performance of CABANA was deemed acceptable by comparing its diagnostic yield with previous data about neuromuscular disorders. Despite some limitations, we expect CABANA to help researchers accurately identify CNVs and reduce the need for subsequent confirmation testing.

Introduction

Next-generation sequencing (NGS), a massively parallel sequencing technology, is one of the most important analytical tools in molecular genetics^{1–4}. NGS technology enables rapid, cost-effective, comprehensive molecular analyses and contributes significantly to detection of a broad range of pathogenic variants, especially small variants such as single-nucleotide variants (SNVs) and small insertions/deletions (INDELs)^{5–7}. Therefore, NGS can help diagnose patients who have phenotypically and genetically diverse disorders, such as neuromuscular disorders (NMDs)^{2,8,9}. In addition to the small variants, copy number variations (CNVs), structural variations that usually range from 1 kb to 3 Mb, can be detected using NGS data and can play an important role in diagnosing patients^{9–12}.

Various computational tools have been developed to enhance the sensitivity of CNV detection in NGS data^{7,13}. However, most tools produce many false-positive CNV calls, necessitating subsequent confirmatory tests such as multiplex ligation-dependent probe amplification (MLPA) and array comparative genomic hybridization^{11,14,15}. For SNVs, the need for additional confirmatory tests has been much reduced by powerful visualization tools such as the Integrative Genomics Viewer, but few such tools are available for CNVs^{15,16}.

Here, we introduce Copy-number Analysis by BAsE-level NormAlization (CABANA), a CNV visualization tool using the normalized single-base-level read depth to identify true CNVs. In a comparison with gold-standard conventional testing, we confirmed that CABANA is both effective and accurate in determining true CNVs.

Methods

Patients and data.

We retrospectively collected NGS data from 474 patients who underwent targeted NGS related to NMDs in our laboratory between July 2016 and December 2020. Using electronic medical records, the disorders suspected when the NGS tests were ordered were largely classified into muscular disorders and neurological disorders and then subsequently subdivided into specific disorders. Three targeted NGS panels were used in this cohort: a muscular panel, neurological panel, and neuromuscular panel consisting of 212, 172, and 599 genes, respectively (Fig. 1; Supplementary Table S1). This study protocol was approved by the Institutional Review Board/ Ethics Committee of the Severance Hospital, Seoul, Korea (IRB No. 4-2020-0715), and the requirement for informed consent was waived by the Institutional Review Board/ Ethics Committee of the Severance Hospital due to the retrospective study design. All methods were performed in accordance with the relevant guidelines and regulations.

Processing of NGS data.

NGS data were generated using a NextSeq 550Dx System (Illumina, San Diego, CA, USA) with 2×151 bp reads. HaplotypeCaller and MuTect2 in the GATK package (3.8-0) and VarScan2 (2.4.0) were used to detect SNVs and small INDELs. Read depth metrics, including allele depth per sample and overall depth of coverage, were obtained through the GATK package. All samples showed a mean depth of at least 500×, and the target region coverage was greater than 99.9% at 30×. ExomeDepth (version 1.1.10), a read depth–based algorithm with high sensitivity, was used to screen exonic CNVs in the target regions^{17–19}.

Calculation of base-level normalized depths and graphical visualization by CABANA.

Using depths of coverage at the base level from the GATK package, the adjusted read depth (ARD) at base position i of sample j ($ARD[i, j]$) was calculated by multiplying the read depth at base position i of sample j by an adjustment factor (retrieved from the sum of the read depths of all samples in the same batch at base position i divided by the sum of the read depths at all base positions in sample j). A *case sample* is defined as the sample whose copy numbers are to be examined, and *control samples* are all the other samples in the same batch. The normalized read depth (NRD) at base position i of the *case* sample ($NRD[i, case]$) was calculated by subtracting the mean of the ARDs at base position i of all *control* samples ($ARD[i, controls]$) from the ARD at base position i of the *case* sample ($ARD[i, case]$) and then dividing that result by $ARD[i, controls]$. The NRDs are depicted on the y-axis with a zero-based scale, so that copy numbers of zero, one, two, three, and four are presented as -1.0, -0.5, 0.0, 0.5, and 1.0, respectively (Fig. 1). A CNV is presumed to be a true positive when the NRDs of the controls show very low variability and those of the case are greatly deviated from those of controls in specific regions or exons. All CNVs called by ExomeDepth software were visually inspected with CABANA. The pathogenicity of CNVs was determined according to the technical standards for interpreting and reporting constitutional CNVs in a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Clinical Genome Resource²⁰.

Confirmation of CNVs.

Among the CNVs visually identified by CABANA, all pathogenic CNVs that could affect the final diagnosis were confirmed by MLPA testing. Commercial SALSA MLPA Probemix P034 DMD mix 1, P035 DMD mix 2, P268 DYSF, P071 LMNB1-PLP1-NOTC3, P033 CMT1, P021-B1 SMA, and P213 HSP mix-2 (MRC Holland, Amsterdam, Netherlands) were used to validate CNVs in *DMD*, *DYSF*, *PLP1*, *PMP22*, *SMN1*, *SMN2*, and *SPAST*, respectively. In brief, 100 ng of genomic DNA was used for ligation. Denaturation at 98°C for 5 minutes, hybridization with each SALSA Probemix at 60°C for 16 hours, ligation reaction by Ligase-65 (MRC-Holland) at 54°C for 15 minutes, and ligase inactivation by incubation at 98°C for 5 minutes were performed sequentially. Multiplex PCR was conducted using fluorescence-labeled universal primers, dNTPs, PCR buffer, and polymerase for 35 cycles (95°C for 30 seconds, 60°C for 30 seconds, 72°C for 1 minute) in a C1000 Thermal Cycler (BioRad, Cressier, Switzerland). The fragments were analyzed using an ABI 3130 Genetic Analyzer (Applied Biosystems, Foster, CA, USA) and GeneMarker 2.2 software (SoftGenetics, State College, PA, USA). Copy numbers were determined as the signal ratio between case and control samples.

Results

Graphical visualization with CABANA.

Raw depths, NRDs, and CNV calls were plotted together. Base-level read depths before normalization were plotted on top, with those of case samples in light green and those of controls in light yellow. In the middle space, NRDs were plotted, with those of case samples in violet and those of controls in light yellow. The CNVs of exons called from ExomeDepth are visualized below the NRD plot, with red and blue boxes representing copy number gains and losses, respectively (Fig. 2). A true-positive CNV typically shows a distinct deviation between the NRD lines of cases and controls, with steady NRD lines of controls around zero with very low variability across all positions in the regions of interest (Fig. 2A). In contrast, false-positive CNVs show an unclear deviation between the NRD line of cases and the highly variable and widely fluctuating NRD lines of the controls (Fig. 2B, Supplementary Fig. S1).

CABANA effectively discriminated a heterozygous deletion in a single or a few exons (Fig. 3A and B). For hemizygous deletions, CABANA provided a clearer and more intuitive illustration of the absent reads in the deleted region (Fig. 3C). For genes with homologous regions or genes, such as *SMN1* and *SMN2*, detecting CNVs can be very tricky using conventional CNV algorithms that adopt read-depth comparisons. Despite that limitation, CABANA successfully discriminated some CNVs in homologous regions, as shown in a case with a homozygous deletion of exon 7 in *SMN1*, which has highly homology with *SMN2* (Fig. 3D). Examples of CNV visualization with explanation are provided in Supplementary Fig. S2.

Small deletions and partial exon deletion.

Because it plots CNVs per base, CABANA can distinguish small deletions undetectable by conventional CNV algorithms. As shown in Figure 4A, a deletion of 3 bases in *PRX* (NM_181882.2:c.1389_1403del) showed a sharp and distinct decrease in the NRD in CABANA, and a deletion of 15 bases in the same gene (NM_181882.2:c.416_418del) showed an apparent decrease in the NRD in the middle of an exon. In addition, CABANA can identify partial exon deletions. Figure 4B shows a distinct deviated line with constant and decreased NRDs in an exon, which indicates the partial exon 4 deletion of *FLCN* (NM_144997.5:c.-24-1345_35delinsG).

Proportion of CNVs with the true-positive pattern.

Among all CNVs called by ExomeDepth, only about 5% showed true-positive CNV patterns when visualized by CABANA. ExomeDepth called CNVs in *TTN*, *NEB*, *TNXB*, *SMN2*, *PIEZO2*, and *ASAH1* in more than 40% of the patients tested with the neuromuscular panel, but only one of them in *SMN2* showed the true-positive pattern when visualized by CABANA. The proportion of true-positive CNVs identified by CABANA was high in *DMD*, where hemizygous mutations in males are frequent. Interestingly, 11% of the CNVs that ExomeDepth detected in *SMN1*, which is highly homologous with *SMN2*, were visualized as true-positive CNVs by CABANA and confirmed as true by MLPA^{21,22}.

Diagnostic utility of CABANA.

Among 474 patients with NMD, 169 (36%) were confirmed to have pathogenic or likely pathogenic mutations (Fig. 5A). Among them, 31 CNVs (7%) were identified by CABANA: 16, 5, 4, 4, 1, and 1 in *DMD*, *PMP22*, *SMN1*, *SPAST*, *DYSF*, and *PLP1*, respectively, all of which were confirmed to be true-positive by MLPA (Supplementary Table S2). The diagnostic yields and proportions of pathogenic CNVs varied by disease entity (Fig. 5B). Of the patients with muscular dystrophies, 14% had pathogenic CNVs, with the most commonly mutated gene being *DMD*. Pathogenic CNVs were observed in 4%, 8%, and 4% of patients with myopathies, spastic paraplegia, and peripheral neuropathy, respectively. In Charcot-Marie-Tooth disease (CMT) and spinal muscular atrophy (SMA), CNVs accounted for 33% and 20% of pathogenic variants, respectively. In addition to the pathogenic or likely pathogenic CNVs, 33 of uncertain significance detected by CABANA in the 474 patients with NMD are listed in Supplementary Table S3.

Discussion

Some patients with genetic disorders have structural abnormalities, of which CNVs are the most common type. NGS can detect both SNVs and CNVs, and many detection algorithms for CNVs have been developed using different principles. One popular method involves comparing the depths of coverage among cases and controls. The performance of those tools has been evaluated in several benchmark studies and was found to depend significantly on the dataset^{12,15,23,24}. Roca et al. reported that ExomeDepth, DeCoN, and ExomeCNV had high sensitivity²³. Moreno-Cabrera et al. conducted a benchmark evaluation of five CNV detection tools (DECoN, CoNVaDING, panelcn.MOPS, ExomeDepth, and CODEX2) using targeted panel NGS data and found that DECoN and panelcn.MOPS had high

sensitivity, and ExomeDepth had the most balanced performance¹⁵. Zhao et al. conducted a performance evaluation of four CNV tools (CoNIFER, cn.MOPS, CNVkit, and exomeCopy) with whole exome sequencing data and found that performance differed according to targeted CNV size and type²⁴. Of note, CNV algorithms based on the depth of coverage method commonly have problems with false positive calls, which are mainly affected by high GC content and poor mappability. Plus, inaccurate detection of small CNVs remains challenging because most target regions of whole exome sequencing and targeted NGS data are small and noncontiguous^{12,23}.

To improve the performance of CNV detection, some CNV visualization tools have been developed (Table 1). Users can recognize true CNVs more intuitively by visually inspecting the depth of coverage in the regions of interest. Most CNV visualization tools use a window of a specific length to reduce variability in read depths, and they generally visualize CNVs at the chromosome or gene level^{25–32}. On the contrary, CABANA visualizes CNVs with high resolution based on normalized single-base-level read depth. To the best of our knowledge, only one previous tool visualizes CNVs using normalized read depths per base¹⁴. With that higher resolution, users can efficiently discriminate true CNVs, both small and large, from false CNVs. Unlike other CNV visualization tools, CABANA produces uniform, steady lines plotted using the NRDs, which are an important factor in filtering out false-positive calls and greatly increase specificity.

Table 1

Copy number variation detection tools with a visualization function. *CNV* Copy number variation; *NGS* Next-generation sequencing; *WGS* Whole genome sequencing.

Name	Window size	Visualization level	Features	Reference
ReadDepth	> 500 bp	Chromosome	Improves resolution in low-coverage data	Miller CA et al. (2011) ²⁵
CNView	100bp–1 kb	Chromosome	Preliminary CNV screening tool for large WGS datasets	RL Collins et al. (2016) ²⁶
iCopyDAV	User-defined (default 100 bp)	Chromosome User-defined region	Integrated platform for CNV detection Functional annotations for CNVs	Dharanipragada P et al. (2018) ²⁷
(untitled)	Amplicon-dependent	Gene	CNV visualization method for amplicon-based sequencing data	SY Nishio et al. (2018) ²⁸
CNVkit	Region-dependent	Chromosome, Gene	Combination of read depths from both the on- and off-target regions Visualization of segmented copy ratios generated from the algorithm	Talevich E et al. (2016) ²⁹
VisCap	Exon-dependent	Chromosome, Exon (low resolution)	CNV visualization for quality control and manual inspection A visual scoring system for filtration of false-positive calls	TJ Pugh et al. (2016) ³⁰
CNspector	User-defined	Chromosome, Exon (low resolution)	Multi-scale CNV visualization with clinically contextual data Web-based tool	Markham JF et al. (2019) ³¹
DeviCNV	Probe-dependent	Chromosome, Exon	Detection tool for exon-level CNVs in targeted NGS data Visualization of CNV candidates with statistical information	Y Kang et al. (2018) ³²
(untitled)	Not required	Gene, Exon (low resolution)	CNV visualization using the normalized reads per nucleotide Potential to replace CNV confirmation tests	Kerkhof J et al. (2017) ¹⁴

Name	Window size	Visualization level	Features	Reference
CABANA	Not required	Exon (high resolution)	<p>Visualization of CNVs screened by a conventional bioinformatics tool at the single-base level</p> <p>An efficient method for detecting exon-level CNVs</p> <p>Potential to replace CNV confirmation tests</p>	Present study

In support of that specificity, 31 pathogenic CNVs determined as true by CABANA were all confirmed to be true by MLPA. Therefore, CABANA visualization can decrease the need for additional confirmatory testing to increase the cost-effectiveness of NGS and reduce the burden on laboratories. In addition, small deletions and partial exon deletions that were not identified by the conventional CNV algorithm were detected by CABANA. Because visual inspection with CABANA is very intuitive, even inexperienced users can easily identify true CNVs. Nonetheless, we recommend that confirmation tests be applied in specific instances, such as a single exon deletion.

CNVs in *TTN*, *NEB*, and *TNXB* were frequently called by ExomeDepth, but none of them showed a true-positive pattern in CABANA. The presence of tandem repeat regions in *TTN* and *NEB* and a highly homologous pseudogene in *TNXB* might have influenced the performance of CABANA^{33–35}. Similar to other bioinformatic tools that use the read-depth approach, CABANA seems to have difficulty in determining true CNVs in regions with high GC content, where highly variable NRDs tend to appear^{7,13,36}. Although the CNVs in *TTN*, *NEB*, and *TNXB* called by ExomeDepth were not confirmed by MLPA, the similar patterns recurrently observed in specific regions of those genes in normal healthy controls suggest a very low likelihood that they are true pathogenic CNVs.

In this study, we found that about 36% of patients with NMD harbored molecular abnormalities on the targeted NGS panels. Previous studies reported that clinically significant variants were detected in 20–49% of NMD patients, but that diagnostic yield varied with the NGS panel and cohort group tested^{9,37,38}. A large-scale study on the diagnosis of NMD using multigene panels showed that pathogenic CNVs were identified in 7.6% of NMD patients, with the majority being on *SMN1*, *PMP22*, *DMD*, and *SPAST*⁹. Using our bioinformatics pipelines with CABANA, we found pathogenic CNVs in 7% of patients with NMD; in concordance with the previous large-scale study, most of them were in *DMD*, *PMP22*, *SPAST*, and *SMN1*.

The most commonly mutated gene in our cohort was *DMD*, a causative gene for Duchenne muscular dystrophy and a major cause of inherited muscular disorders in Korea³⁹. Of 39 pathogenic variants in *DMD*, 16 (41%) were CNVs. Considering that approximately 70% of Duchenne/Becker muscular dystrophy patients with molecular defects had pathogenic variants of *DMD* in the form of CNVs, the proportion of CNVs detected by CABANA seems to be low^{40,41}. However, there might have been selection bias in our

patient cohort because some patients had proven to be negative for CNVs by MLPA or quantitative PCR before NGS testing. *SPAST*, a major causative gene for autosomal dominant spastic paraplegia⁴², was the third most commonly mutated gene in our patients with NMD, with about 29% being pathogenic CNVs. Previous studies on hereditary spastic paraplegia reported that the proportion of pathogenic CNVs was 2.5–37.5% depending on the characteristics of each cohort^{43,44}. *PMP22* is related to CMT type 1A and hereditary neuropathy with pressure palsies, with most patients having deletion and duplication CNVs, respectively⁴⁵. Consistent with that, all the pathogenic variants found in *PMP22* in this study were CNVs. Collectively, the mutation spectrum and proportion of CNVs in these disorders found using CABANA were concordant with the literature. In most patients, phenotype was consistent with disorders related to the gene with pathogenic CNVs. This evidence supports our CABANA algorithm as robust and accurate.

Our study has some limitations. First, the performance of CABANA could not be thoroughly evaluated due to the limited availability of confirmatory tests and practical considerations, such as an uncertain false-negative rate. Nonetheless, its performance was deemed to be acceptable compared with previously reported CNV data in patients with NMD and clinical correlations with our patients' results. Second, CNV visualization with CABANA was performed only on CNVs called by ExomeDepth, which might have missed true CNVs¹⁵. Third, as described above, it can be challenging for CABANA to identify true CNVs in repeat regions, highly homologous regions, and GC content-rich regions^{7,13,36}.

In summary, we developed a base-level visualization software, CABANA, as a confirmatory tool for CNVs called by other algorithms. With its high resolution, CABANA showed excellent fidelity and specificity and could help exclude false CNVs and identify true CNVs without additional confirmation tests. In patients with NMD, CABANA effectively detected pathogenic CNVs, demonstrating its high utility with clinical samples.

Declarations

Author contributions statement

Conceptualization: S.-T.L. Data Curation: H.K. Funding acquisition: S.S., S.-T.L. Investigation: H.K., Y.S., T.G.L., D.W., S.S. Methodology: S.-T.L. Resources: S.S. Software: S.-T.L. Supervision: J.R.C. Visualization: H.K. Writing - Original Draft: H.K. Writing - Review & Editing: S.S., S.-T.L.

Data availability statement

The formula used in this study is included in this article, and all copy number variants identified in this study are presented in the Supplementary Information.

Funding

This work was supported by National Research Foundation of Korea [grant numbers NRF-2021R1I1A1A01045980, NRF-2016R1D1A1B01010566].

Competing interests

The authors declare no competing interests.

References

1. Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet* **17**, 95-115, doi:10.1146/annurev-genom-083115-022413 (2016).
2. Laing, N. G. Genetics of neuromuscular disorders. *Crit Rev Clin Lab Sci* **49**, 33-48, doi:10.3109/10408363.2012.658906 (2012).
3. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).
4. Tucker, T., Marra, M. & Friedman, J. M. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* **85**, 142-154, doi:10.1016/j.ajhg.2009.06.022 (2009).
5. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46, doi:10.1038/nrg2626 (2010).
6. Tian, X. *et al.* Expanding genotype/phenotype of neuromuscular diseases by comprehensive target capture/NGS. *Neurol Genet* **1**, e14, doi:10.1212/nxg.0000000000000015 (2015).
7. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711-2718, doi:10.1093/bioinformatics/bts535 (2012).
8. von der Hagen, M. *et al.* Facing the genetic heterogeneity in neuromuscular disorders: linkage analysis as an economic diagnostic approach towards the molecular diagnosis. *Neuromuscul Disord* **16**, 4-13, doi:10.1016/j.nmd.2005.10.001 (2006).
9. Winder, T. L. *et al.* Clinical utility of multigene analysis in over 25,000 patients with neuromuscular disorders. *Neurol Genet* **6**, e412, doi:10.1212/nxg.00000000000000412 (2020).
10. Välipakka, S. *et al.* Copy number variation analysis increases the diagnostic yield in muscle diseases. *Neurol Genet* **3**, e204, doi:10.1212/nxg.00000000000000204 (2017).
11. Giugliano, T. *et al.* Copy Number Variants Account for a Tiny Fraction of Undiagnosed Myopathic Patients. *Genes (Basel)* **9**, doi:10.3390/genes9110524 (2018).
12. Zhang, L., Bai, W., Yuan, N. & Du, Z. Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput Biol* **15**, e1007069, doi:10.1371/journal.pcbi.1007069 (2019).
13. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**, S1, doi:10.1186/1471-2105-14-s11-s1 (2013).

14. Kerkhof, J. *et al.* Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. *J Mol Diagn* **19**, 905-920, doi:10.1016/j.jmoldx.2017.07.004 (2017).
15. Moreno-Cabrera, J. M. *et al.* Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet* **28**, 1645-1655, doi:10.1038/s41431-020-0675-z (2020).
16. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
17. Plagnol, V. *et al.* 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 (2012).
18. Sadedin, S. P., Ellis, J. A., Masters, S. L. & Oshlack, A. Ximmer: a system for improving accuracy and consistency of CNV calling from exome data. *Gigascience* **7**, doi:10.1093/gigascience/giy112 (2018).
19. Samarakoon, P. S. *et al.* cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data. *BMC Genomics* **17**, 51, doi:10.1186/s12864-016-2374-2 (2016).
20. Riggs, E. R. *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* **22**, 245-257, doi:10.1038/s41436-019-0686-8 (2020).
21. Chen, X. *et al.* Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet Med* **22**, 945-953, doi:10.1038/s41436-020-0754-0 (2020).
22. Feng, Y. *et al.* The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic SMN1 copy-number and sequence variant analysis by massively parallel sequencing. *Genet Med* **19**, 936-944, doi:10.1038/gim.2016.215 (2017).
23. Roca, I., González-Castro, L., Fernández, H., Couce, M. L. & Fernández-Marmiesse, A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res* **779**, 114-125, doi:10.1016/j.mrrev.2019.02.005 (2019).
24. Zhao, L., Liu, H., Yuan, X., Gao, K. & Duan, J. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics* **21**, 97, doi:10.1186/s12859-020-3421-1 (2020).
25. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6**, e16327, doi:10.1371/journal.pone.0016327 (2011).
26. Collins, R. L., Stone, M. R., Brand, H., Glessner, J. T. & Talkowski, M. E. CNView: a visualization and annotation tool for copy number variation from whole-genome sequencing. *bioRxiv*, 049536, doi:10.1101/049536 (2016).
27. Dharanipragada, P., Vogeti, S. & Parekh, N. iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. *PLoS One* **13**, e0195334,

- doi:10.1371/journal.pone.0195334 (2018).
28. Nishio, S. Y., Moteki, H. & Usami, S. I. Simple and efficient germline copy number variant visualization method for the Ion AmpliSeq™ custom panel. *Mol Genet Genomic Med* **6**, 678-686, doi:10.1002/mgg3.399 (2018).
 29. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873, doi:10.1371/journal.pcbi.1004873 (2016).
 30. Pugh, T. J. *et al.* VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet Med* **18**, 712-719, doi:10.1038/gim.2015.156 (2016).
 31. Markham, J. F. *et al.* CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. *Sci Rep* **9**, 6426, doi:10.1038/s41598-019-42858-8 (2019).
 32. Kang, Y. *et al.* DeviCNV: detection and visualization of exon-level copy number variants in targeted next-generation sequencing data. *BMC Bioinformatics* **19**, 381, doi:10.1186/s12859-018-2409-6 (2018).
 33. Savarese, M. *et al.* The complexity of titin splicing pattern in human adult skeletal muscles. *Skelet Muscle* **8**, 11, doi:10.1186/s13395-018-0156-z (2018).
 34. Zenagui, R. *et al.* A Reliable Targeted Next-Generation Sequencing Strategy for Diagnosis of Myopathies and Muscular Dystrophies, Especially for the Giant Titin and Nebulin Genes. *J Mol Diagn* **20**, 533-549, doi:10.1016/j.jmoldx.2018.04.001 (2018).
 35. Morissette, R. *et al.* Broadening the Spectrum of Ehlers Danlos Syndrome in Patients With Congenital Adrenal Hyperplasia. *J Clin Endocrinol Metab* **100**, E1143-1152, doi:10.1210/jc.2015-2232 (2015).
 36. Szatkiewicz, J. P., Wang, W., Sullivan, P. F., Wang, W. & Sun, W. Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic Acids Res* **41**, 1519-1532, doi:10.1093/nar/gks1363 (2013).
 37. Park, J. *et al.* Usefulness of comprehensive targeted multigene panel sequencing for neuromuscular disorders in Korean patients. *Mol Genet Genomic Med* **7**, e00947, doi:10.1002/mgg3.947 (2019).
 38. Gonzalez-Quereda, L. *et al.* Targeted Next-Generation Sequencing in a Large Cohort of Genetically Undiagnosed Patients with Neuromuscular Disorders in Spain. *Genes (Basel)* **11**, doi:10.3390/genes11050539 (2020).
 39. Park, H. J. *et al.* Discovery of pathogenic variants in a large Korean cohort of inherited muscular disorders. *Clin Genet* **91**, 403-410, doi:10.1111/cge.12826 (2017).
 40. Takeshima, Y. *et al.* Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. *J Hum Genet* **55**, 379-388, doi:10.1038/jhg.2010.49 (2010).
 41. Prior, T. W. & Bridgeman, S. J. Experience and strategy for the molecular testing of Duchenne muscular dystrophy. *J Mol Diagn* **7**, 317-326, doi:10.1016/s1525-1578(10)60560-0 (2005).

42. Solowska, J. M. & Baas, P. W. Hereditary spastic paraplegia SPG4: what is known and not known about the disease. *Brain* **138**, 2471-2484, doi:10.1093/brain/awv178 (2015).

43. Kadnikova, V. A., Rudenskaya, G. E., Stepanova, A. A., Sermyagina, I. G. & Ryzhkova, O. P. Mutational Spectrum of Spast (Spg4) and At11 (Spg3a) Genes In Russian Patients With Hereditary Spastic Paraplegia. *Sci Rep* **9**, 14412, doi:10.1038/s41598-019-50911-9 (2019).

44. Boone, P. M. *et al.* The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am J Hum Genet* **95**, 143-161, doi:10.1016/j.ajhg.2014.06.014 (2014).

45. van Paassen, B. W. *et al.* PMP22 related neuropathies: Charcot-Marie-Tooth disease type 1A and Hereditary Neuropathy with liability to Pressure Palsies. *Orphanet J Rare Dis* **9**, 38, doi:10.1186/1750-1172-9-38 (2014).

Figures

Figure 1

CABANA workflow using next-generation sequencing data obtained from patients with neuromuscular disorders. The numbers in parentheses indicate the number of samples involved. The formula for

normalizing read depth is described in the blue box, where i, j , and n represent a certain base position, a certain sample, and the number of samples in a batch, respectively. RD, ARD, and NRD stand for read depth, adjusted read depth, and normalized read depth, respectively. A case sample is one whose copy numbers are to be tested, and control samples are all the samples in a single batch excluding the case sample. In calculating ARD, the sum of read depths from the start to end positions indicates the sum of read depths in all the targeted regions examined.

Figure 2

Examples of true- and false-positive copy number variations visualized with CABANA. (A) A true-positive CNV pattern visualized by CABANA is shown in exon 63–78 of *DMD*. The normalized read depth line of the case deviates significantly from those of controls, which have very low variability around zero. Exon 63–78 duplication of *DMD* was confirmed as true by MLPA. (B) A false-positive CNV pattern visualized by CABANA is shown in exon 11 of *MLC1*, where the copy number loss signal was generated by ExomeDepth. In this region, the normalized read depth line of the case deviates unevenly from the base line, and those of the controls exhibit high variability. The copy number loss signal of exon 11 was confirmed as false by MLPA.

Figure 3

Examples of copy number loss visualized by CABANA with MLPA confirmation. (A) Heterozygous exon 3 deletion in *DMD*, (B) Heterozygous exon 12–13 deletion in *DYSF*, (C) Hemizygous exon 45–55 deletion in *DMD*, (D) Homozygous exon 7 deletion in *SMN1*. All CNVs visualized by CABANA were confirmed as true by MLPA (MLPA data are not shown for the hemizygous exon 51–55 deletion in *DMD* in Figure 3C).

Figure 4

Examples of small deletions and partial exon deletions visualized by CABANA. (A) Two small deletions of *PRX* (NM_181882.2:c.1389_1403del, NM_181882.2:c.416_418del). Drastic declines in normalized read depths are shown for these two variants. (B) A partial exon deletion in exon 4 of *FLCN* (NM_144997.5:c.-24-1345_35delinsG). A steady line with decreased normalized read depth is observed in a specific part of the exon, where no copy number loss signal was called by ExomeDepth.

Figure 5

Diagnostic yield of three next-generation sequencing panels associated with neuromuscular disorders. (A) Overall diagnostic yield of small variants and copy number variations. Proportions of genes with true-positive copy number variations are shown in the stacked bar chart. (B) The frequency and proportion of small variants and copy number variations identified according to disorder suspected when next-generation sequencing was ordered. The percentage of true-positive copy number variations in each disorder is indicated at the top of the bar.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure202201242.pdf](#)
- [SupplementaryTable20220124.docx](#)