ORIGINAL ARTICLE

Parallel detection of single nucleotide variants and copy number variants with exome analysis: Validation in a cohort of 700 undiagnosed patients

Hisato Suzuki[1] | Mamiko Yamada[1] | Tomoko Uehara[1] |
Toshiki Takenouchi[2] | Kenjiro Kosaki[1]

[1]Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

[2]Department of Pediatrics, Keio University School of Medicine, Tokyo, Japan

Correspondence
Kenjiro Kosaki, MD, PhD, Center for Medical Genetics, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo 160-8582, Japan.
Email: kkosaki@z3.keio.jp

Abstract

Copy number variants (CNVs) are significant causes of rare and undiagnosed diseases. Parallel detection of single nucleotide variants (SNVs) and CNVs with exome analysis, if feasible, would shorten the diagnostic closure in a timely manner. We validated such "parallel" approach through a cohort study of 791 undiagnosed patients. In addition to routine exome analysis, we applied an innovative algorithm EXCAVATOR2 which enhances sensitivity by paradoxically exploiting read depth data that covers nonexonic regions where baits were not originally intended to hybridize. About 48 patients had copy number variations, 42 deletions, and 6 duplications with a resolution of 0.51–14.7 mega base pairs. Importantly from a clinical standpoint, we identified three patients with "dual diagnosis" due to concurrent pathogenic CNV and SNV. We suggest "hitting two birds with one stone" approach to exome data is an efficient strategy in deciphering undiagnosed patients and may well be considered as a first-tier genetic test.

KEYWORDS
chromosomal microarray, copy number variants, exome sequencing, undiagnosed diseases

1 | INTRODUCTION

Copy number variants (CNVs) including chromosomal deletions and duplications and single nucleotide variants (SNVs) along with small insertions/deletions (indels) represent major causes of intellectual disability with or without multiple congenital anomalies. CNVs are typically detected using chromosomal microarrays (CMA), whereas SNVs and small indels are typically detected using an exome analysis performed on next-generation sequencers (NGS). Usually, CMA and exome analyses are performed in a serial manner, rather than in a parallel manner. Several algorithms have been developed in the hope of obtaining CNV data from aligned exome data that are originally

acquired for SNV detection during conventional exome analyses. Representative examples include XHMM (Miyatake et al., 2015), Conifer (Pfundt et al., 2017), and CNVnator (Abyzov, Urban, Snyder, & Gerstein, 2011). Each algorithm has been partially successful but sensitivity and specificity are less than optimal.

Recently, the innovative algorithm EXCAVATOR2 has been developed; this algorithm calculates the copy number information through the active use of read data from exome baits that have been mapped to off-target regions (i.e., introns and intergenic regions), in contrast to previously developed software that discard information such off-target reads (D'Aurizio et al., 2016). EXCAVATOR2 allows copy number estimations of some, if not all, regions outside of the

coding regions against which the exome bait probes have been designed. The performance of EXCAVATOR2 has been evaluated using simulated alignment map file (bam file) data, but the effectiveness of the algorithm has not yet been clinically validated in a large set of patient data. Here, we document the clinical utility of EXCAVATOR2 through a reanalysis of alignment map files generated during regular exome analyses of 791 patients with intellectual disability and multiple congenital anomalies from a nationwide project on rare and undiagnosed diseases.

2 | MATERIALS AND METHODS

2.1 | Subjects

The research protocol was approved by the institutional review board (IRB) of Keio University School of Medicine. Undiagnosed patients were enrolled in a Japanese project known as the "Initiative on Rare and Undiagnosed Diseases (IRUD)" (Adachi et al., 2017). Since July 2015, a total of 791 undiagnosed patients have been enrolled in this project, and trio-exome sequencing has identified causative SNVs or indels in known human disease-related genes in 304 patients (38.4%). In most of the patients, CMA testing had not been performed prior to entry in the presently reported study because CMA testing has not been reimbursed by the national health insurance program in Japan.

2.2 | EXCAVATOR2 analysis

The EXCAVATOR2 analysis was performed as originally described (D'Aurizio et al., 2016). The detailed setting is summarized in Supporting Information. Pathogenic significance of CNVs detected by the EXCAVATOR2 analysis was determined by clinical geneticists who evaluate whether CNVs could account for the phenotypes of the patients. CNVs observed in more than 3 unrelated patients among the 73 patients were excluded as artifacts or polymorphisms.

2.3 | Chromosomal microarray testing

CMA testing was performed using a SurePrint G3 Human CGH 1×1 M microarray (Agilent Technologies, Santa Clara, CA) or a SurePrint G3 Human CGH 4×180 K microarray (Agilent Technologies). Microarray slides were scanned using Agilent CytoGenomics 4.9.3.12.

3 | RESULTS

3.1 | Exploratory cohort

As an exploratory cohort, we evaluated 73 patients who had not been clinically diagnosed despite previous exome analyses and who had not yet undergone CMA testing. We performed both a CMA analysis and an EXCAVATOR2 analysis. Two clinical geneticists independently performed the CMA and EXCAVATOR2 analyses for the 73 patients. CMA testing identified 11 CNVs among the 73 undiagnosed patients (12.3%). Eight of these 11 patients had a microdeletion (0.2–9.7 Mb), and the remaining 3 patients had a microduplication (0.51–1.3 Mb). For the EXCAVATOR2 analysis, we used the standard criteria for the definition of pathogenic CNVs (Vermeesch, Brady, Sanlaville, Kok, & Hastings, 2012; Wyandt, Wilson, & Tonk, 2017). The EXCAVATOR2 analysis detected candidate CNVs in 11 patients. Nine CNVs were shared by both analysis results. Representative output images of the EXCAVATOR2 results are shown in Figure 1.

The sensitivity and specificity of the EXCAVATOR2 analysis was determined by defining the CMA test as the gold-standard. The receiver operating characteristic (ROC) curve is shown in Figure 2. The area under the curve (AUC) was 0.904. Hence, the EXCAVATOR2 algorithm had a CNV detection rate comparable to that of CMA (Akobeng, 2007). The ROC analysis indicated that the accuracy (true positive + true negative/overall) of the analysis was maximized when the threshold was set at 0.87 Mb. When that threshold was applied, the sensitivity and specificity were 73 and 100%, respectively.

3.2 | Validation cohort

We further performed EXCAVTOR2 analyses for the remaining samples (i.e., 718 patients; validation cohort), including those patients who had previously shown to have a pathogenic SNVs per conventional exome analysis. This cohort was distinct from the exploratory cohort. The EXCAVATOR2 algorithm detected CNVs in 39 patients (5.4%) among the 718 patients. All the CNVs were confirmed using CMA testing except for one CNV, which was confirmed using fluorescence in situ hybridization (FISH). The results of the CMA and
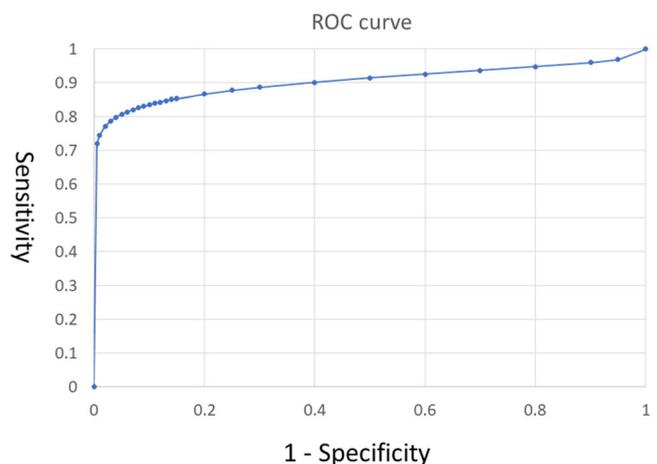
The area under the curve (AUC) was 0.904

FIGURE 1 Comparison of the results of EXCAVATOR2 and chromosomal microarray (CMA) testing. Top: copy number variations (CNV) detected using EXCAVATOR2. Bottom: CNV detected using CMA [Color figure can be viewed at wileyonlinelibrary.com]
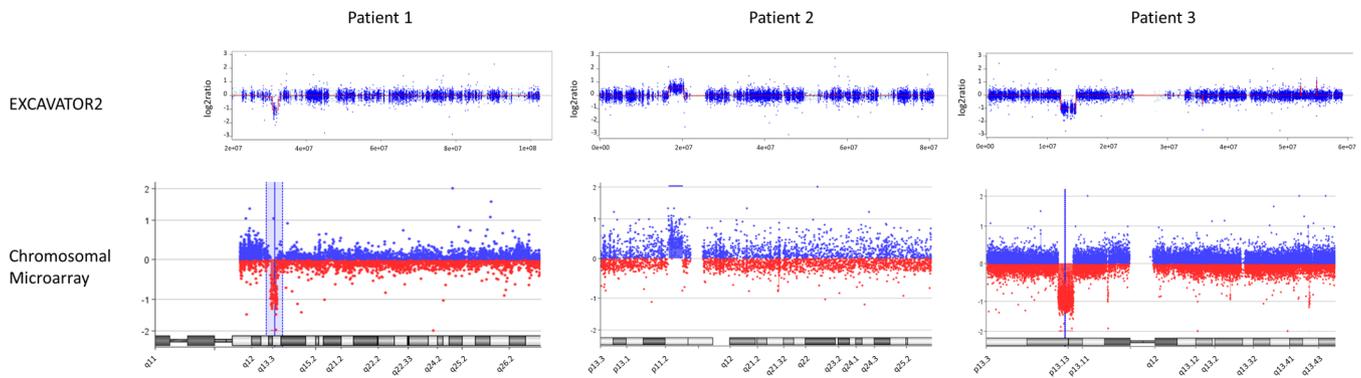
FIGURE 2 Receiver operating characteristic (ROC) curve for EXCAVATOR2 to predict the threshold for pathogenic copy number variation (CNV). The ROC was determined for 76 undiagnosed patients. CNVs were detected in 11 patients using EXCAVATOR2 and in 11 patients using chromosomal microarray testing. The results were the same for 9 patients. The area under the ROC curve was 0.904 [Color figure can be viewed at wileyonlinelibrary.com]

EXCAVATOR2 analyses for both the exploratory and validation cohorts are summarized in Table S1. All CNVs were confirmed to be de novo.

Importantly from a clinical standpoint, we identified three patients with "dual diagnosis" due to concurrent pathogenic CNV and SNV. Patient 1 was an 11-year-old girl with mild intellectual disability (IQ = 71 at the age of 8), congenital scoliosis, and chronic thrombocytopenia (<100,000 platelets per microliter of blood). The trio exome analysis identified a de novo frameshift variant in the *ETV6* (NM_001987.4: c.1153-1_1165del) gene, which reportedly causes thrombocytopenia 5 (OMIM:#616216). Additionally, the EXCAVATOR2 analysis identified a 1.6-Mb deletion in 15q13.1q13.3 which resides within the critical region of 15q13.3 microdeletion syndrome (OMIM:#612001). The thrombocytopenia was ascribed to the *ETV6* mutation, and the intellectual disability was ascribed to the 15q13.3 microdeletion. Patient 2 was a 5-year-old boy with feeding difficulty and intellectual disability. Dysmorphic features included thick hair and eyebrows, low-set ears, a high palate, micrognathia, and short fifth fingers. The exome analysis identified a de novo frameshift variant in *ARID1B* (NM_020732.3: c.2687del, p.Leu896Argfs*18), which causes Coffin-Siris syndrome (OMIM:#135900). The EXCAVATOR2 analysis identified a 3.84-Mb duplication in 17p11.2. The 17p11.2 duplication syndrome is known as Potocki-Lupski syndrome (OMIM:#610883). The intellectual disability and dysmorphic facial features overlapped with those seen in Coffin-Siris syndrome and Potocki-Lupski syndrome. The short fifth fingers were ascribed to Coffin-Siris syndrome. Patient 3 was a 1-year-old boy with hypotonia, feeding difficulty, and developmental delay. Dysmorphic features included frontal bossing and a prominent occiput. The exome analysis identified a hemizygous missense variant in *MED12* (NM_005120.3:c.3067A>G, p.Ile1023Val), which reportedly causes Lujan-Fryns syndrome (Yamamoto & Shimojima, 2015). The EXCAVATOR2 analysis identified a 2.5-Mb deletion in 19p13.2p13.12 which included two disease-associated genes: *NFIX* (Sotos syndrome 2, OMIM:#614753), and *CACNA1A* (epileptic encephalopathy, early infantile 42, OMIM:#617106). The intellectual disability and hypotonia were ascribed to both the *MED12* mutation and the 19p13.2p13.12

microdeletion. The dysmorphic features, including frontal bossing, were ascribed to the deletion of *NFIX* contained in 19p13.2p13.12.

4 | DISCUSSION

We demonstrated the clinical utility of the EXCAVATOR2 algorithm as an effective screening method for detecting chromosomal deletions and duplications using exome data. From the standpoint of the cost of genetic testing, the application of the EXCAVATOR2 algorithm requires no additional consumables, other than the computational cost, once a standard exome analysis has been performed.

Our ROC analysis of the exploratory cohort, consisting of blinded CMA and EXCAVATOR2 analyses, showed that the optimized threshold for detecting deletions was 0.87 Mb. This size is slightly smaller than the threshold value of 1 Mb recommended for the detection of pathogenic microdeletions (Gijsbers, Schoumans, & Ruivenkamp, 2011). Moreover, our validation study demonstrated that all the microdeletions detected by the EXCAVATOR2 algorithm were also detected using CMA testing, suggesting a high specificity for EXCAVATOR2 analyses. It is significant to note that we identified 3 patients among 304 diagnosed patients (0.99%) who had both pathogenic SNV and CNV. A conventional array-first approach would have truncated the diagnostic journey of the three patients without pinpointing the additional pathogenic SNVs/small indels.

The EXCAVATOR2 method has some disadvantages compared with existing standard methods, such as array comparative genomic hybridization (array CGH) or SNP microarrays. First, array CGH analysis and SNP arrays have a better resolution than the EXCAVATOR2 analysis of exome data. Second, SNP arrays can readily detect uniparental disomy. Reanalysis using dedicated software, such as UPDio (King et al., 2014), would be helpful for detecting UPD. Despite these limitations, the prime advantage of the EXCAVATOR2 analysis is that this method allows the reutilization of existing data with minimal additional analytic costs.

In conclusion, we suggest "hitting two birds with one stone" approach to exome data, using EXCAVATOR2, is an efficient strategy

in deciphering undiagnosed patients and may well be considered as a first-tier genetic test.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Hisato Suzuki: Material preparation, data collection and analysis, writing of first draft of the manuscript; Mamiko Yamada: Material preparation, data collection and analysis; Tomoko Uehara: Material preparation, data collection and analysis; Toshiki Takenouchi: Material preparation, data collection and analysis; Kenjiro Kosaki: Writing of first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors contributed to the study conception and design. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author [KK].

ORCID

Hisato Suzuki https://orcid.org/0000-0002-8122-7180
Mamiko Yamada https://orcid.org/0000-0002-4039-8899
Tomoko Uehara https://orcid.org/0000-0002-1497-7686
Toshiki Takenouchi https://orcid.org/0000-0002-7311-4135
Kenjiro Kosaki https://orcid.org/0000-0002-6798-8151

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6), 974–984.

Adachi, T., Kawamura, K., Furusawa, Y., Nishizaki, Y., Imanishi, N., Umehara, S., ... Suematsu, M. (2017). Japan's initiative on rare and undiagnosed diseases (IRUD): Towards an end to the diagnostic odyssey. *European Journal of Human Genetics*, 25(9), 1025–1028.

Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica*, 96(5), 644–647.

D'Aurizio, R., Pippucci, T., Tattini, L., Giusti, B., Pellegrini, M., & Magi, A. (2016). Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Research*, 44(20), e154.

Gijsbers, A. C., Schoumans, J., & Ruivenkamp, C. A. (2011). Interpretation of array comparative genome hybridization data: A major challenge. *Cytogenetic and Genome Research*, 135(3–4), 222–227.

King, D. A., Fitzgerald, T. W., Miller, R., Canham, N., Clayton-Smith, J., Johnson, D., ... Study, D. D. D. (2014). A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Research*, 24(4), 673–687.

Miyatake, S., Koshimizu, E., Fujita, A., Fukai, R., Imagawa, E., Ohba, C., ... Matsumoto, N. (2015). Detecting copy-number variations in whole-exome sequencing data using the eXome hidden Markov model: An 'exome-first' approach. *Journal of Human Genetics*, 60(4), 175–182.

Pfundt, R., Del Rosario, M., Vissers, L., Kwint, M. P., Janssen, I. M., de Leeuw, N., ... Hehir-Kwa, J. Y. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine*, 19(6), 667–675.

Vermeesch, J. R., Brady, P. D., Sanlaville, D., Kok, K., & Hastings, R. J. (2012). Genome-wide arrays: Quality criteria and platforms to be used in routine diagnostics. *Human Mutation*, 33(6), 906–915.

Wyandt, H. E., Wilson, G. N., & Tonk, V. S. (2017). Array-comparative genomic hybridization/microarray analysis: Interpretation of copy number variants. In H. E. Wyandt, G. N. Wilson, & V. S. Tonk (Eds.), *Human chromosome variation: Heteromorphism, polymorphism and pathogenesis* (pp. 191–234). Singapore: Springer.

Yamamoto, T., & Shimojima, K. (2015). A novel MED12 mutation associated with non-specific X-linked intellectual disability. *Human Genome Variation*, 2, 15018.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.