의학박사 학위논문

Clinical application of the algorithm for detecting copy number variations at the exon level using next-generation sequencing data

NGS 데이터를 이용한 엑손 수준 복제수변이 분석 알고리즘의 임상적 적용

August 2021

Department of Medicine

Seoul National University

College of Medicine

Man Jin Kim

A thesis of Degree of Doctor of Philosophy

NGS 데이터를 이용한 엑손 수준 복제수변이 분석 알고리즘의 임상적 적용

Clinical application of the algorithm for detecting copy number variations at the exon level using next-generation sequencing data

August 2021

Department of Medicine

Seoul National University

College of Medicine

Man Jin Kim

NGS 데이터를 이용한 엑손 수준 복제수변이 분석 알고리즘의 임상적 적용

지도 교수 박 성 섭

이 논문을 의학박사 학위논문으로 제출함
2021년 7월

서울대학교 대학원
의학과 검사의학 전공
김 만 진

김만진의 의학박사 학위논문을 인준함
2021년 7월

위 원 장 _____

부위원장 _____

위 원 _____

위 원 _____

위 원 _____

Clinical application of the algorithm for detecting copy number variations at the exon level using next-generation sequencing data

By Man Jin Kim
(Directed by Sung Sup Park, M.D., Ph.D.)

April 2021

Approved by Thesis Committee:

Professor _____

Professor _____

Professor _____

Professor _____

Professor _____

Abstract

Clinical application of the algorithm for detecting copy number variations at the exon level using next-generation sequencing data

.

Man Jin Kim

Department of Medicine

The Graduated School

Seoul National University

Introduction: Despite the importance of exonic copy number variations (CNVs) in human genetic diseases, reliable Next-generation sequencing (NGS)-based methods for detecting them are unavailable. We developed an expandable and robust exonic CNV detection tool called consistent count region (CCR)-CNV. **Methods**: In total, 1,000 samples of the truth set were used for validating CCR-CNV. A custom targeted gene panel containing

hundreds of genes as well as exome sequencing data was included in the truth set.

Results: The overall sensitivity of our method was 99.7%, which was superior to that of other CNV tools, such as DECoN, Atlas-CNV, and CNV-RF. Importantly, the false discovery rate of our method was comparable to that of other tools. CCR-CNV also showed a high concordance rate with the chromosomal microarray analysis data. Moreover, genome-wide CNV screening by using low-coverage genome sequencing showed comparable performance to that of chromosomal microarray analysis.

Conclusion: Here, we present a novel diagnostic tool that allows the identification of exonic CNVs with high confidence using various reagents and clinical NGS platforms. We validated this method using the largest multiple ligation-dependent probe amplification (MLPA)-confirmed dataset, including sufficient copy-normal control data.

Keyword : copy-number variation; germ-line; molecular genetics; targeted gene panel clinical sequencing

Student Number : 2016-21960

Table of Contents

List of Tables

List of Figures

Introduction

Copy number variation (CNV) is a type of genomic structural variation involving segmental duplications or deletions of a DNA fragment. The size of CNV varies from 50 bp to several megabases.[1,2] An increasing number of studies have reported that large-scale CNVs encompassing several genes are related to many human genetic diseases, including autism, Alzheimer's disease, epilepsy, and schizophrenia.[3,4] The intragenic or exonic CNVs of smaller size account for up to 3–35% of pathogenic or likely pathogenic variants.[5] Exonic CNV accounts for a high proportion of causative variants in certain diseases. For example, exonic CNV accounts for 70% cases of Duchenne muscular dystrophy.[6,7] Therefore, reliable methods for detecting exonic CNVs are crucial for the diagnosis of human genetic diseases.

Previously, large-scale CNVs were detected using conventional karyotyping or chromosome microarray (CMA).[8] In contrast, multiple ligation-dependent probe amplification (MLPA) is currently considered the gold standard for diagnosing exonic CNVs owing to its high sensitivity and specificity.[9,10] Next generation sequencing (NGS)-based targeted sequencing and exome sequencing are becoming increasingly common in clinical genetic testing.[11-13] Single nucleotide variants and small insertion and

deletion variants have been accurately detected using NGS. However, proper diagnoses of exonic deletions or duplications using targeted NGS data have proved to be challenging.[14-26] In addition, MLPA cannot be performed individually on all genes in a NGS panel because of the associated expenses.

There are two major types of NGS-based CNV detection algorithms—read-depth and paired-end mapping[8]—and both use statistical models and clustering approaches, respectively, for CNV detection. The depth of coverage can be affected by many factors such as the type of enrichment methods and sequencing platform, and panel size.[27] Various methods have been developed to identify exonic CNVs in gene panel sequencing data.[14-26] These methods often use complex data normalization algorithms that reduce sensitivity for exonic CNVs and provide highly segmented copy-number regions, resulting in high false-positive rates. The use of sophisticated algorithms such as machine learning in clinical settings is difficult. Furthermore, strict quality criteria of input data have restricted the robustness of these tools. These barriers have forced many diagnostic laboratories to exclude exonic CNVs from clinical NGS-based genetic testing. In addition, these algorithms were not validated using sufficient MLPA-confirmed data. Statistically sufficient negative data can be used to accurately

determine the false discovery rate (FDR). FDR is an important factor, as it indicates the extent to which reconfirmation using alternative methods such as MLPA is necessary. Therefore, development of a reliable, sensitive, and efficient exonic CNV detection tool using NGS data is important for clinical molecular diagnostics.

Here, we developed an expandable and robust CNV detection method called "consistent count region (CCR)-CNV". We validated this method using the largest MLPA-confirmed data, including sufficient copy-normal control data, which enabled calculation of clinically relevant FDR. Furthermore, the performance of our methods on several NGS platforms and enrichment methods were also evaluated.

Materials and Methods

Samples and patient consent

Data from 1,100 samples were included, of which 200 samples were in the control set and 900 samples were in the clinical validation set. Data were generated on lymphocyte DNA extracted from the peripheral blood of patients and healthy individuals using a Chemagic 360 instrument (Perkin Elmer, Baesweiler, Germany). Although the study was conducted retrospectively, informed consent for NGS testing was obtained from all the enrolled patients.

MLPA

For confirming exonic deletions and duplications, MLPA was performed using the appropriate probe kits and protocols from MRC Holland (Amsterdam, The Netherlands). In total, 1,068 exonic CNVs were confirmed using MLPA. DNA denaturation, probe–target sequence hybridization, probe ligation, and polymerase chain reaction (PCR) of the ligated probes were performed according to the manufacturer's instructions. The products were loaded onto an ABI PRISM 3130xl DNA analyzer (Applied Biosystems, Foster City, CA, USA) and analyzed using GeneMarker software version 1.51 (SoftGenetics).

ICR96 exon validation series

The ICR96 exon validation series was used for evaluating CCR-CNV (www.icr.ac.uk/icr96.).[28] This dataset consists of data from a targeted NGS assay comprising 96 independent samples confirmed using MLPA. Thirty samples had normal copy number, as confirmed using MLPA for 26 genes. Among the 30 negative samples, 25 samples were selected as controls for CCR-CNV. As the ICR96 exon validation series included no gender information, two X-linked genes (*GPC3* and *FANCB*) were excluded from analysis.

Targeted NGS sequencing: Custom panel I, II-1, II-2

Library preparation was performed according to SureSelectXT Target Enrichment protocol (Agilent, Santa Clara, CA, USA). The target genes are listed in Table 1. Paired-end 150-bp sequencing was performed using the MiSeq platform for custom panels I and II-1, and the NextSeq platform for custom panel II-2 (Illumina, San Diego, CA, USA). The raw data of targeted sequencing was obtained in the FASTQ format. The sequencing data were mapped to the human reference genome sequence (GRCh37/hg19), and per-base coverage was calculated using the NextGENe software v2.4.0.1. (SoftGenetics).

Table 1. List of panels and covered genes used in this study.

Panel Name	No. of tested genes	Covered genes
ICR96	94	*AIP, ALK, APC, ATM, BAP1, BLM, BMPR1A, BRCA1, BRCA2, BRIP1, BUB1B, CDC73, CDH1, CDK4, CDKN1C, CDKN2A, CEBPA, CEP57, CHEK2, CYLD, DDB2, DICER1, DIS3L2, EGFR, EPCAM, ERCC2, ERCC3, ERCC4, ERCC5, EXT1, EXT2, EZH2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FH, FLCN, GATA2, GPC3, HNF1A, HRAS, KIT, MAX, MEN1, MET, MLH1, MSH2, MSH6, MUTYH, NBN, NF1, NF2, NSD1, PALB2, PHOX2B, PMS1, PMS2, PRF1, PRKAR1A, PTCH1, PTEN, RAD51C, RAD51D, RB1, RECQL4, RET, RHBDF2, RUNX1, SBDS, SDHAF2, SDHB, SDHC, SDHD, SLX4, SMAD4, SMARCB1, STK11, SUFU, TMEM127, TP53, TSC1, TSC2, VHL, WRN, WT1, XPA, XPC*
Custom Panel I	148	*AGL, ALS2, ANG, ANO5, ANOS1, ATL1, ATXN2, BRCA1, BRCA2, BRIP1, BSCL2, C9orf72, CAPN3, CAV3, CHCHD10, CHD7, CHMP2B, COL5A1, COL5A2, CYP7B1, DAG1, DES, DMD, DNAJB6, DYSF, ENO3, ERBB4, ERCC4, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FBP1, FEZF1, FGF17, FGF8, FGFR1, FH, FHL1, FIG4, FKRP, FKTN, FSHB, FUS, G6PC, GAA, GBE1, GLE1, GMPPB, GNRH1, GNRHR, GYS1, GYS2, HNRNPA1, HNRNPDL, HSPD1, IL17RD, ISPD, ITGA7, KIF1B, KIF5A, KISS1, KISS1R, L1CAM, LAMP2, LHB, LIMS2, LMNA, MATR3, MAX, MDH2, MYOT, NEFH, NF1, NIPA1, NR0B1, NSMF, OPTN, PALB2, PFKM, PFN1, PGAM2, PGM1, PHKA2, PHKB, PHKG2, PLEC, PLP1, POMGNT1, POMT1, POMT2, PRKAG2, PROK2, PROKR2, PYGL, PYGM, RAD51C, REEP1, RET, SDHA, SDHAF2, SDHB, SDHC, SDHD, SETX, SGCA, SGCB, SGCD, SGCG, SIGMAR1, SLC16A2, SLC2A2, SLC37A4, SLX4, SOD1, SPAST, SPG11, SPG20, SPG21, SPG7, SQSTM1, TAC3, TACR3, TAF15, TARDBP, TBK1, TCAP, TMEM127, TNPO3, TRAPPC11, TRIM32, TTN, UBQLN2, VAPB, VCP, VHL, WASHC5, WDR11, XRCC2, ZFYVE26, ZFYVE27*

Table 1. Continued

Panel Name	No. of tested genes	Covered genes
Custom Panel II−1	200	*ABCB11, ABCC8, ABCD1, ACADM, ACADS, ACTA2, ACVR1, ACVRL1, AGL, ALAS2, ALB, APC, ASAH1, ASL, ASS1, ATL1, ATP7A, ATP7B, ATP8B1, ATRX, BCKDHB, BMPR2, BRAT1, BRCA1, BSCL2, BTK, CACNA1A, CACNA1S, CALM1, CAPN3, CASP10, CASP8, CBS, CD3G, CFH, CFTR, CHN1, CLCN1, CLN3, COL1A1, COL1A2, COL2A1, COL3A1, COMP, CPS1, CPT1A, CREBBP, CSF1R, CSF3R, CYBB, CYP21A2, CYP4V2, DPYD, DYSF, ELN, ENG, EXT1, EXT2, F11, F12, F8, FAH, FANCA, FBN1, FECH, FGFR1, FGFR2, FGFR3, FLCN, FOXP3, FRMD7, FUS, G6PC, G6PD, GAA, GALC, GALE, GALT, GARS, GBA, GBE1, GCDH, GCK, GLA, GLB1, GLUD1, GNAS, GNE, GNPTAB, GPI, HADHB, HEXB, HMBS, HNF1A, IDUA, IKBKG, INSR, IVD, JAG1, KCNH2, KCNQ1, KMT2D, LDLR, LIPA, LPL, MAPT, MARS, MCCC1, MCCC2, MEN1, MFN2, MKS1, MLH1, MSH2, MTM1, MUT, MYL3, NF1, NF2, NIPBL, NOTCH3, NPC1, NPHS1, NPR2, NSD1, NTRK1, OCRL, OPA1, OTC, PAFAH1B1, PAH, PAX6, PCCA, PCCB, PDHA1, PEX1, PHEX, PHKA2, PKD1, PKD2, PKHD1, PKLR, PLG, POR, PRF1, PRKN, PRODH, PROS1, PSEN1, PTCH1, PTEN, PTPN11, PYGM, RAF1, RB1, RET, RP1L1, RPE65, SBF1, SCN4A, SCN5A, SDHB, SDHD, SGCE, SLC12A3, SLC22A12, SLC25A13, SLC26A4, SLC2A1, SLC2A2, SLC3A1, SLC6A19, SLC7A7, SMAD4, SOS1, SPAST, SPTLC1, STAT3, STK11, TGFBI, THRB, TMEM67, TMPRSS6, TNNI3, TP53, TPP1, TRDN, TSC1, TSC2, TSHR, TTR, TYMP, UBE3A, UMOD, UNC13D, VHL, VWF, WAS, WNT10A, YARS*
Custom Panel II−2		
Whole exome sequencing	> 20,000	All human 20,00 genes

Exome sequencing

The following method of exome sequencing was utilized for gene analysis. The SureSelectXT Human All Exon V5 and Human All Exon V6 kits (Agilent) were used to enrich the exon regions of the genome. Paired-end 100-bp sequencing was performed using the Illumina HiSeq platform (Illumina). The produced sequencing data were aligned to the human reference genome sequence (GRCh37/hg19). Per-base coverage was obtained using NextGENe's distribution coverage report function.

Low-coverage whole genome sequencing

The sequencing libraries were prepared according to the manufacturer's instructions of TruSeq DNA PCR-free sample preparation kit (Illumina). Paired-end (2 × 100 bp) sequencing was performed by Macrogen (Macrogen, Seoul, Korea) using the NovaSeq6000 platform. Per-base coverage was calculated using the NextGENe software v2.4.0.1. (SoftGenetics).

CMA

CMA analysis was performed using CytoScan 750K (Affymetrix, Santa Clara, CA, USA). The array contained >750,436 CNV markers, including 200,436 genotypable SNP probes and >550,000 non-SNP probes. The overall average marker space was 4,127 base pairs. All data were visualized and analyzed using the Chromosome Analysis Suite (ChAS) software package (Affymetrix) with the human genome (hg19) sequence. The software CytoScan 750K was designed to detect minimum 200 kb aberrations.

Statistical methods

Cut-off values for determining CNVs were searched based on receiver-operating characteristic (ROC) curve analysis. The Youden index method was used to determine the optimal cut-off level on the ROC curve.[29,30] *P* values in this study were uniformly computed using Wilcoxon's rank-sum test. All statistical analyses were conducted using R version 3.6.0 (The R Foundation, www.r-project.org).

Results

CCR-CNV overview

An overview of CCR-CNV was illustrated by obtaining the CCRs of a target exon (Figure 1). Assuming that the target exon was exon 1 of gene 1 (yellow square), the depth of the target exon was called α; +20% and -20% from α are indicated by transverse dotted lines. The exons that consistently belonged to this range among three controls were defined as the CCR of the target exon (shaded yellow squares). Furthermore, 'r' is the value of DCtarget divided by the mean value of DC_{CCR1} and DC_{CCR2}. If 'r' was 0.744 or less, the CNV status of the target exon was classified as a heterozygous deletion. If DC_N was 1.273 or more, it was classified as a heterozygous duplication. The rest were classified as a normal copy.

To determine thresholds for CCR candidates, the percentage of CCR calling for ICR96 data was plotted against the threshold ranging from 10% to 95%. The cut-offs for deletion and duplication were predetermined for each threshold point according to the results of ROC analysis (Figure 2). At the 20% threshold, the percentage of CCR calling was 85.1% (Figure 3A). However, the specificity for duplication was the highest (98.98%) at this threshold (Figure 2, 3E). Most of the false positive calls were due to duplication. As most of the false positive calls were due to

duplication, specificity for duplication was important for reducing the FDR. Therefore 20% was selected as the optimal threshold. We plotted a ROC curve and calculated the Youden index for determining the cut-off value for heterozygous deletion and duplication. The optimal cut-offs for heterozygous deletion and duplication were 0.744 and 1.273, respectively. The best cut-points for hemizygous deletion and duplication were 0.0081 and 1.69, respectively (Figure 4). The ROC curve was generated using CCR-CNV analysis data on the ICR96 exon CNV validation series.[28]

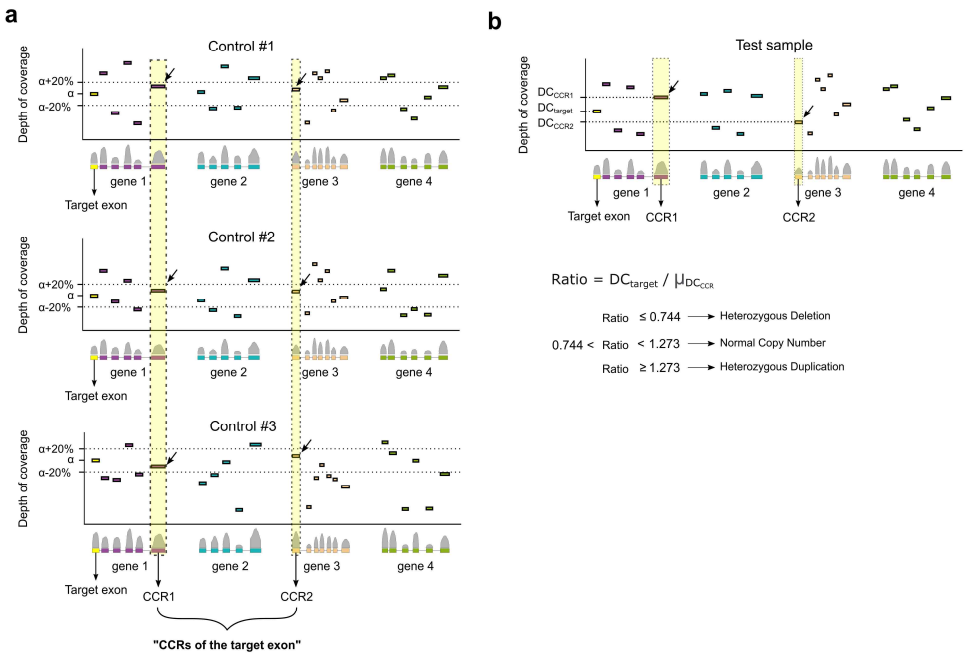
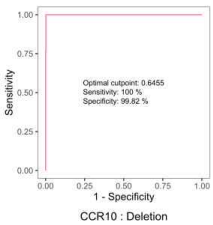


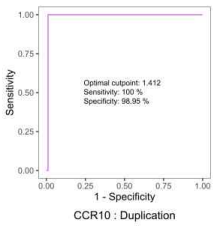
Figure 1. Schematic overview of the CCR-CNV method.

Schematic overview of the CCR-CNV method. (a) Assuming that the target exon is exon 1 of gene 1 (yellow square), the depth of the target exon is called α. +20% and -20% from α are indicated by transverse dotted lines. The exons that consistently belong to this range among three controls are defined as the CCR of the target exon (shaded yellow squares). (b) How to determine CNV of test sample is shown. 'Ratio' is the value of DCtarget divided by the mean value of DCCCR1 and DCCCR2. If the ratio was 0.744 or less, the CNV status of the target exon was classified as a heterozygous deletion. If the ratio was 1.273 or more, it was classified as a heterozygous duplication. The rest were classified as a normal copy.

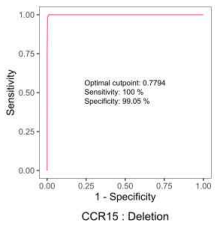
Abbreviations: CCR, consistent count region; CNV, copy number variation; DC, depth of coverage.



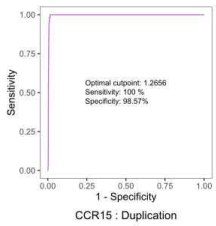
CCR10 : Deletion



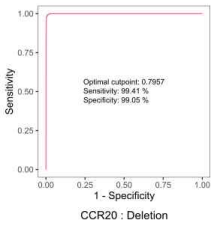
CCR10 : Duplication



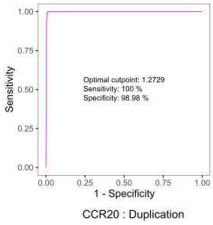
CCR15 : Deletion



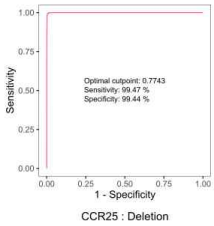
CCR15 : Duplication



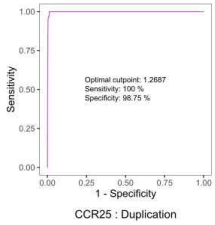
CCR20 : Deletion



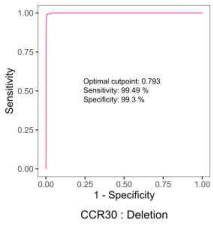
CCR20 : Duplication



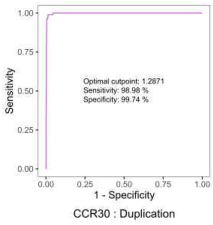
CCR25 : Deletion



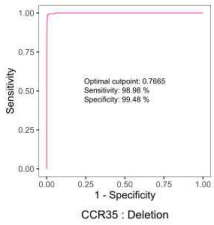
CCR25 : Duplication



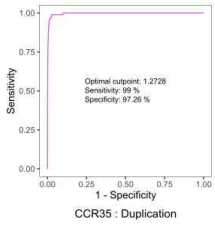
CCR30 : Deletion



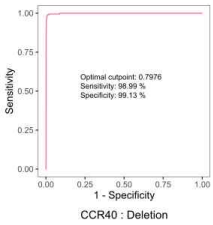
CCR30 : Duplication



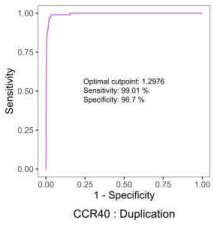
CCR35 : Deletion



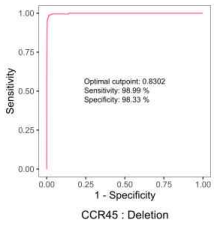
CCR35 : Duplication



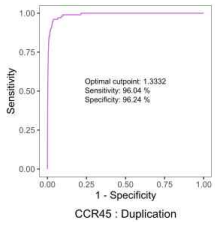
CCR40 : Deletion



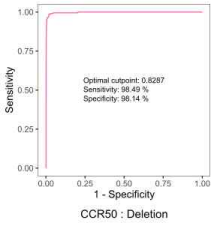
CCR40 : Duplication



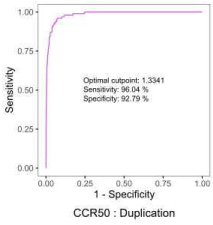
CCR45 : Deletion



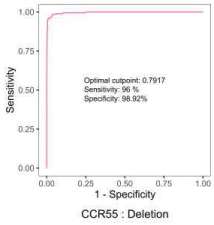
CCR45 : Duplication



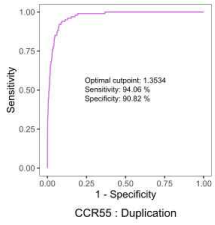
CCR50 : Deletion



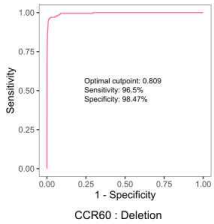
CCR50 : Duplication



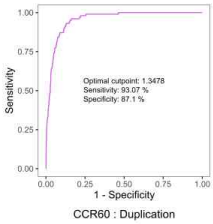
CCR55 : Deletion



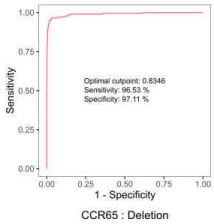
CCR55 : Duplication



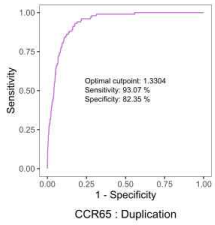
CCR60 : Deletion



CCR60 : Duplication



CCR65 : Deletion



CCR65 : Duplication

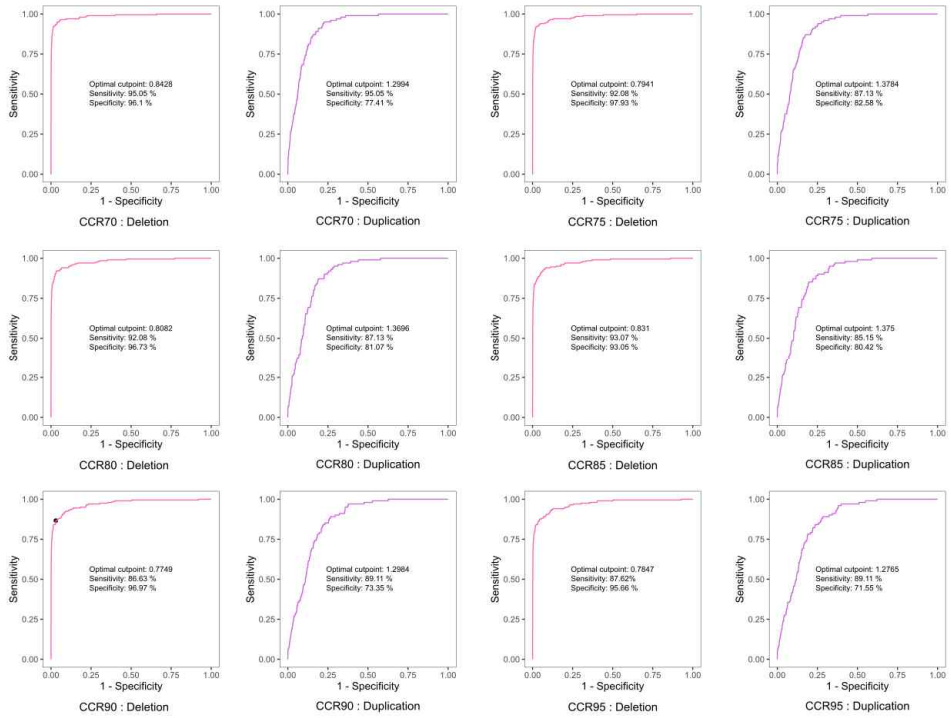


Figure 2. The ROC analysis to select a threshold for CCR candidates.

To determine thresholds for CCR candidates, the percent of CCR calling for ICR96 data was plotted for threshold ranged 10% to 95%. For each threshold, optimal cutpoints for deletion and duplication were determined according to the results of ROC analysis. Then sensitivity and specificity were calculated based on these cut-offs. Abbreviations: ROC, receiver-operating characteristic; CCR, consistent count region.

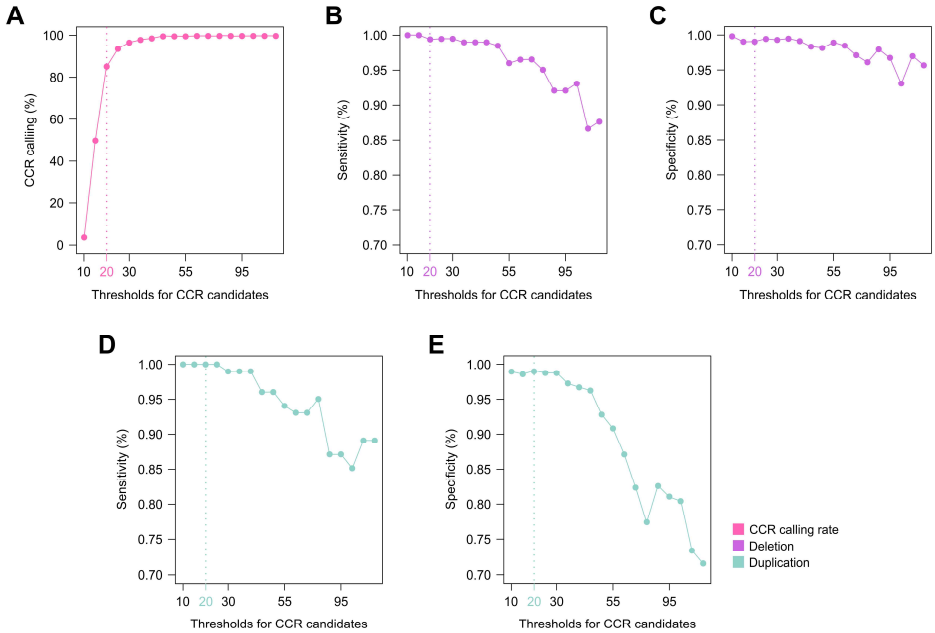


Figure 3. Selection of optimal thresholds for including CCR candidates.

The percentage of CCR calling (A), sensitivity (B,D), and specificity (C,E) were plotted against each threshold point. The accuracy analysis was performed using the ICR96 exon validation series (https://www.icr.ac.uk/icr96).

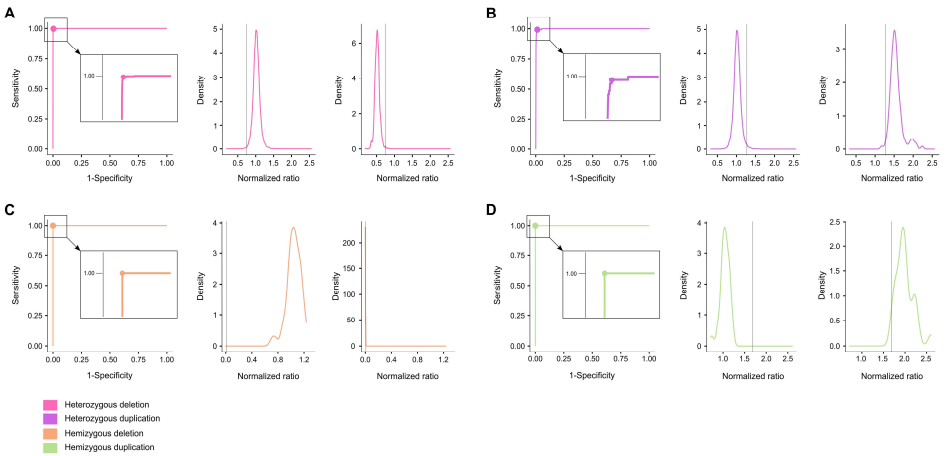


Figure 4. ROC analysis for determining the optimal cut-offs for deletions and duplications.

ROC curve analysis was performed on heterozygous (A, B) and hemizygous (C, D) CNVs. The ROC analysis was performed using CCR-CNV analysis data on the ICR96 exon validation series (https://www.icr.ac.uk/icr96).

Clinical validation of CCR-CNV

The reagent and platforms of NGS assays used to evaluate the CCR-CNV algorithm are summarized in Table 2. The ICR96 exon CNV validation series includes high-quality sequencing data from the TruSight Cancer Panel (Illumina) for 96 independent samples. It contained 94 genes and the panel size was 0.25 MB. Eleven male patients and female carriers with *DMD* were tested using the custom panel I. Custom panel II included 200 rare diseases-related genes. Among these, the CNV of 33 genes were confirmed using MLPA. Custom panel II was used with the same capture probes as in NextSeq 550 (Illumina), while custom panel I was tested on MiSeq (Illumina). Exome sequencing was performed using SureSelect Human All Exon V5 and V6 (Agilent). MLPA was performed on 28 genes.

Using the MLPA results on each exon as the truth set, we determined the number of true positive, true negative, false positive, and false negative calls using CCR-CNV (Table 3). The ICR96 had 193 false positive calls and 5 false negative calls, showing 98.9% sensitivity and 98.6% specificity. Custom panel I, which was confirmed using MLPA only for *DMD*, had no positive calls and two false negative calls in heterozygous deletion, and no false negative calls in hemizygous deletion. The sensitivity was 100% and

specificity was 99.4% for heterozygote calling, while they were 100% for hemizygous calling. In total, 391 exons were misclassified as positive, and one exon was misclassified as negative, showing 99.8% sensitivity and 97.4% specificity when custom panel II-1 data were analyzed using CCR-CNV. The sensitivity and specificity of CCR-CNV on custom panel II-2 were 100% and 99.5%, respectively. Finally, the sensitivity and specificity of CCR-CNV on exome sequencing were 100% and 96.0%, respectively.

Table 2. Characteristics of gene panels for validation.

Panel name	Genes tested	Target size (Mb)	No. of MLPA-confirmed genes	Mean depth of controls	CCR calling (%)	No. of test samples	Mean depth of test samples	Enrichment	Sequencing platform
ICR96[a]	94	0.25	33	1331.7	85.1	67	1316.3	Trusight Rapid Capture	Illumina HiSeq 2500
Custom Panel I	148	0.45	1	243.8	74.2	11	431.1	Custom SureSelect, Agilent	Illumina Miseq
Custom Panel II-1	199	0.57	54	170.3	58.5	537	174.9	Custom SureSelect XT, Agilent	Illumina Miseq
Custom Panel II-2	199	0.57	54	1004.5	91.6	303	1274.6	Custom SureSelect XT, Agilent	Illumina Nextseq 550
Exome	> 20,000	50 / 60[b]	28	160	73.7	42	149.7	Human All Exon V5 / V6, Agilent	Illumina HiSeq 2500

Abbreviations: MLPA, multiple ligation-dependent probe amplification; CCRs, consistent count regions.

[a]ICR96 exon validation series (www.icr.ac.uk/icr96.)

[b]50 Mb for Human All Exon V5, 60 Mb for Human All Exon V6

Table 3. Accuracy of CCR-CNV.

Panel Name	Enrichment	Sequencing platform	No. of tested genes	No. of MLPA-confirmed genes	CCR calling (%)	No. of tested samples	Mean depth of tested samples	Subcategory I	Subcategory II	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	F1
Custom Panel I	Custom SureSelect, Agilent	Illumina MiSeq	148	1	74.2	11	431.1	*DMD* heterozygote	Del	180	320	2	0	1	0.9938	0.989	1	0.9945
									Dup	26	320	2	0	1	0.9938	0.9286	1	0.963
								DMD hemizygote	Del	39	52	0	0	1	1	1	1	1
									Dup	26	52	0	0	1	1	1	1	1
Custom Panel II-1	Custom SureSelect, Agilent	Illumina MiSeq	199	54	58.5	537	174.9		Del	631	14,765	391	0	1	0.9742	0.6174	1	0.7635
									Dup	2	14,765	391	1	0.6667	0.9742	0.0051	0.9999	0.0101
Custom Panel II-2	Custom SureSelect, Agilent	Illumina NextSeq 550	199	54	91.6	303	1274.6		Del	15	8,391	42	0	1	0.995	0.2632	1	0.4167
									Dup	3	8,391	42	0	1	0.995	0.0667	1	0.125
Exome	Human All Exon V5 / V6, Agilent	Illumina HiSeq 2500	> 20,000	28	73.7	42	149.7		Del	13	799	33	0	1	0.9603	0.2826	1	0.4407
									Dup	2	799	33	0	1	0.9603	0.0571	1	0.1081
Total									Del	878	24,327	468	1	0.9989	0.9811	0.6523	1	0.7892
									Dup	59	24,327	468	1	0.9833	0.9811	0.112	1	0.201

Abbreviations: MLPA, multiple ligation-dependent probe amplification; CCR, consistent count region; CNV, copy number variation; TP, true positive; TN, true negative; FP, false positive; FN, false negative; Del, deletion; Dup, duplication.

Various applications of CCR-CNV

Heterozygous deletion in exons 5 and 6 of *STX16* was initially identified using MLPA (Figure 5A). Exon 7 deletion, which was not detected using MLPA, was suspected using the CCR-CNV method. Gap PCR was subsequently performed. Finally, we confirmed a deletion of 2,979 bp spanning exons 5 to 7. CCR-CNV analysis was performed on the exome sequencing data of five patients who underwent clinical CMA testing to estimate the clinical sensitivity of our algorithm for large-scale CNVs (Table 4). The cases included 4 pathogenic genomic losses and 1 pathogenic genomic gain. The range of CNVs varied from 909 kb to 6,558 kb. The sensitivity of CMA-positive calls ranged from 98.1% to 100% using CCR-CNV. A 0.9 Mb deletion on the X chromosome encompassing *ZC4H2* was identified (arr [hg19] Xq11.2 (63,603,040 - 64,512,427) × 1) (Figure 5B). Furthermore, we applied the CCR-CNV on the two low-coverage whole genome sequencing data for detecting large-scale CNVs (Figure 5C,D). Approximately 6 MB copy number loss in chr 4: 71,552 - 6,164,267 was identified using CCR-CNV analysis, in addition to approximately 6.8 MB copy number gain in chr 7: 42,976 - 6,861,119.

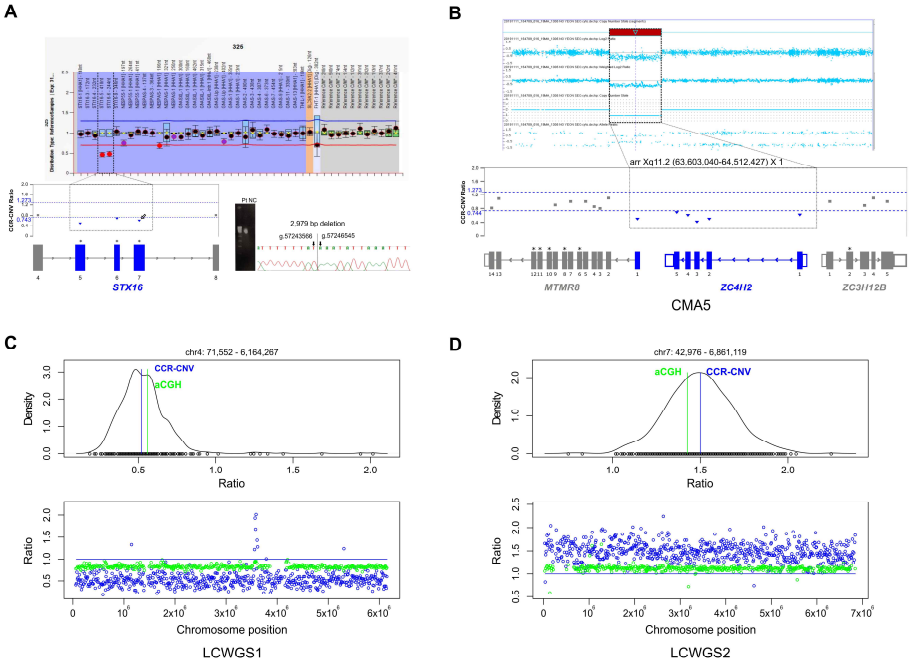


Figure 5. Various applications of CCR-CNV.

(A) Heterozygous deletions in exons 5 and 6 of *STX16* was initially identified using MLPA. Exon 7 deletion, which was not detected using MLPA, was suspected using the CCR-CNV method. Gap PCR was subsequently performed. Finally, we confirmed the 2,979 bp deletion spanning exons 5 to 7. (B) CMA identified a 0.9 Mb deletion on the X chromosome (arr [hg19] Xq11.2 (63,603,040 – 64,512,427) × 1) containing *ZC4H2*. The results of CCR analysis were in good agreement with the CMA results. Identification of approximately 6 Mb copy loss on chr 4: 71,552 – 6,164,267 (C) and approximately 6.8 Mb copy gain on chr 7: 42,976 – 6,861,119 (D) using CCR-CNV analysis on low-pass whole genome

sequencing data.

Abbreviations: CCR-CNV, consistent count region-copy number variation; NGS, next-generation sequencing; MLPA, multiplex ligation-dependent probe amplification; CMA, chromosomal microarray.

Table 4. Performance of CCR-CNV for low coverage whole genome sequencing data to detect large-scale CNVs.

Sample No.	Gender	Diagnosis	CMA results	Size (kb)	Gain or Loss	No. of 10 kb windows	Ratio
LCWGS1	Male	Wolf-Hirschhorn syndrome	arr[hg19] 4p16.3p16.1 (71,552-6,164,267) x1	6092.7	Loss	599	0.52
LCWGS2	Male	7p22.2 microduplication syndrome	arr[hg19] 7p22.3p22.1 (42,976-6,861,119) x3	6818.1	Gain	674	1.5

Abbreviations: CCR, consistent count region; CNV, copy number variation. CMA, chromosomal microarray; LCWGS, low coverage whole genome sequencing.

Comparison between the ratios of true positive and false positive calls from CCR-CNV

The ratios of false CNVs and true CNVs were compared (Figure 6). The ratios of the true CNV and false CNV groups differed significantly in all panels excluding false duplication from the exome sequencing data. The ratio of the true call did not differ significantly from that of the false call in duplication from exome sequencing data.

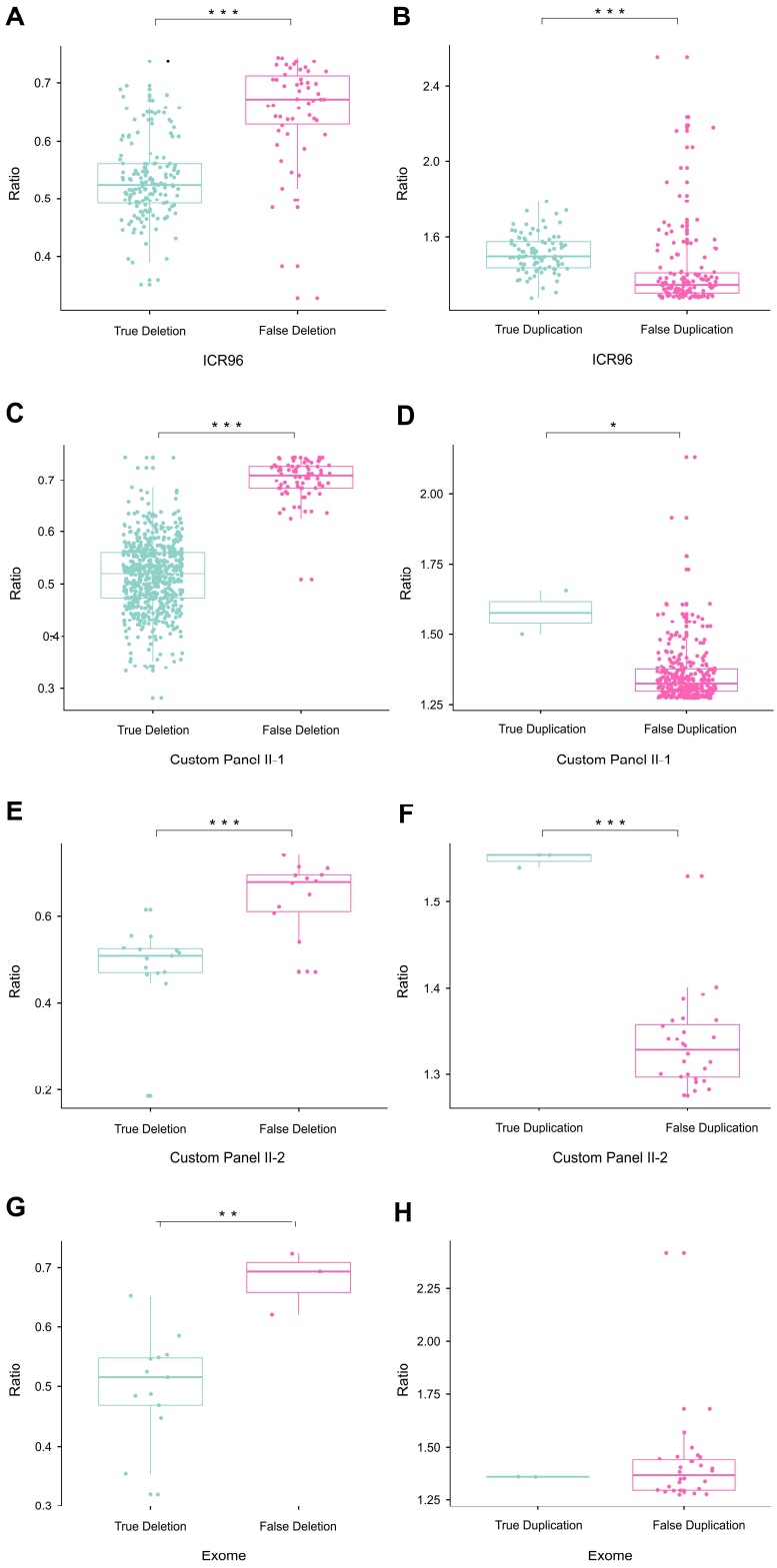

A
B
C
D
E
F
G
H
Ratio
True Deletion
False Deletion
True Duplication
False Duplication
ICR96
Custom Panel II-1
Custom Panel II-2
Exome
* * *
* *
*

Figure 6. Comparison between the ratios of true and false CNVs.

The ratios of false and true CNVs were compared. (A–G) The ratios of the true CNV group differed significantly from those of the false CNVs group in all panels, excluding false duplication from exome sequencing data. (H) The ratio of the true call did not differ significantly from that of the false call in duplication from exome sequencing data. (A–G): (*) $P < 0.05$; (**) $P < 0.01$; (***) $P < 0.001$. Abbreviations: CNVs, Copy number variations.

Discussion

We report here an expandable and robust algorithm to detect germline CNVs from diverse types of NGS data. CCR-CNV was designed for implementation in clinical molecular diagnostic laboratories. Approximately 1,000 MLPA-confirmed NGS data were enrolled in this study. As MLPA is currently considered the "reference standard" for detecting CNVs in humans, we validated our method using the largest number of reported MLPA-confirmed data.[14-26]

CCR-CNV is robust. It works effectively on panels of various size, commonly used capture systems, and sequencing platforms. Furthermore, CCR-CNV can be applied irrespective of the quality of the input data. We included the various factors that may affect NGS data, such as enrichment method and sequencing platform. Both Illumina and Agilent capture methods were validated in this study. The Agilent SureSelect technology uses RNA molecules as probes, while Illumina Truseq technology utilizes DNA as capture probe molecules. This difference can result in a significant alteration in target coverage.[31] The type of sequencing platform can also affect mean depth of coverage. The depth of coverage of custom panel II-2 on NextSeq 550 (Illumina) was higher than that of the custom panel II-1 on MiSeq (Illumina). Furthermore, the

CCR calling rate dramatically increased and specificity also improved in custom panel II-2. The performance can be expected to improve with increase in read depth. In this context, the low efficiency of CCR-CNV on exome sequencing data is expected to improve if depth of coverage increases.

We evaluated the performance of CCR-CNV and compared it with those of three CNV tools validated with MLPA-confirmed data (Table 5). CCR-CNV was applied to data from various Illumina platforms, while DECoN,[16] Atlas-CNV[15] and CNV-RF[22] were validated using data generated on Hiseq 2500 (Illumina). CCR-CNV was superior in terms of sensitivity. The sensitivity of CCR-CNV was 99.7% (95% confidence interval (CI) = 99.1% - 99.9%) while that of the other methods ranged from 86.0% to 96.8%. Moreover, the specificity of our method was slightly lower than that of the other methods. Of note, we included more negative samples to determine a more reliable specificity, which may justify our lower specificity values. As the positive results given by this CNV tools are yet to be confirmed using a reference method, it is important to determine the extent of FDR.[32] Even in the context of similar specificity values, FDR generally increases as more negative samples are included. Therefore, to determine a clinically relevant FDR, we assumed that the yield of genetic testing in the clinical

laboratories is 30% and that CNV accounts for 5% of them. As a result, the estimated positive rate of CNV is 1.5%. Under these assumptions, the FDR in this study was 53.1% (95% CI = 51.2% – 55.0%). The FDRs of DECoN and CNV-RF were 18.1% (95% CI = 3.0% – 60.9%) and 94.8% (86.8% – 98.7%), respectively. Of note, the FDR of DECoN was lower than that of our method; however, the CI was wide since the value was obtained from a small number of samples. These results show that the FDR of our method is comparable to those of three other methods. Furthermore, most of the false positive-results came from incorrectly-called single exon duplications. Excluding single duplication events, our method's FDR was reduced to 23.7% (95% CI = 21.2% – 26.4%).

Our study has some limitations. CNV analysis is not possible for exons without CCR. Comparison between custom panels II-1 and II-2 revealed that increasing the mean depth of coverage may improve the CCR calling rate (Figure 7). In addition, we can design NGS panels containing high probability CCRs after further evaluation of the common characteristics of CCRs, such as GC ratio or exon size.

NGS consists of numerous quality control points, namely, input DNA quality, fragmentation options such as enzyme and ultrasound, end-repair adapter ligation, enrichment methods such as amplicon

or capture, and several sequencing issues such as cluster density.[33] These numerous step-by-step issues are reflected in the read depth, resulting in variability in NGS data. Therefore, the variability of each step cannot be dealt with individually, but as a whole. CCR-CNV can efficiently unmask the fluctuation of all these steps and uncover the true signal.

In summary, CCR-CNV can be determined using simple calculations, circumventing the need of a complicated algorithm. This method can potentially be used to detect CNVs in a clinical molecular diagnostics laboratory. In conclusion, CCR-CNV is a simple and robust CNV diagnosis method, which was clinically validated using numerous types of MLPA-confirmed NGS data.

Table 5. Comparison of exonic CNV tools validated using MLPA-confirmed data.

Tool name	Capture system	Sequencing platform	Positive set		Negative set		Sensitivity (95% CI) (%)	Specificity (95% CI) (%)	Estimated FDR** (95% CI) (%)
			No. of MLPA-confirmed Positive CNVs	No. of Tested genes	No. of MLPA-confirmed Negative CNVs	No. of Tested genes			
CCR-CNV	Illumina Rapid Capture, Agilent SureSelect	MiSeq, NextSeq 550, Hiseq 2500	118	37	2018	88	99.7 (99.1 - 99.9)	98.3 (98.1 - 98.4)	53.1 (51.2 - 55.0
DECoN(Fowler et al. 2016)	Illumina Rapid Capture	HiSeq 2500	31	10	308	2	96.8 (83.3 - 99.9)	99.7 (98.2 - 100.0)	18.1 (3.0 - 60.9)
Atlas-CNV(Chiang et al. 2019)	Illumina Rapid Capture, Roche-Nimblegen methods	HiSeq 2500	64	22	NA*	NA*	86.0	NA*	NA*
CNV-RF(Onsongo et al. 2016)	Illumina Rapid Capture, Agilent SureSelect	HiSeq 2500	13	9	3	3	92.3 (64.0 - 99.8)	100 (29.2 - 100.0)	94.8 (86.8 - 98.7)

* The specificity of Atlas-CNV is calculated based on the Illumina HumanExome-12v array (Illumina) results rather than MLPA.

** The prevalence of CNV is assumed to be 1.5%

Abbreviations: CNV, copy number variation; MLPA, multiple ligation-dependent probe amplification; CI, confidence interval; NA, not assessed.

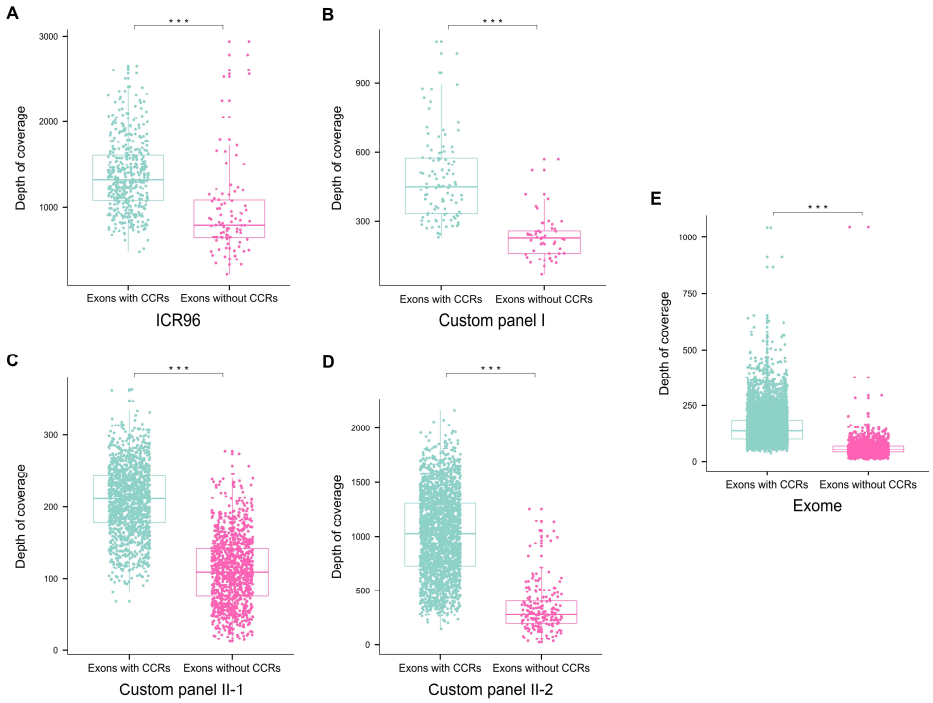


Figure 7. Comparison of depths based on the presence of CCRs.

For each panel, the read depth was compared between exons with and without CCRs. Each dot represents an individual exon constituting each panel. (green dots : exons with CCRs, pink dots : exons without CCRs). (***) $P < 0.001$.

Abbreviations: CCR, consistent count region.

References

1. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-992.
2. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172-183.
3. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-455.
4. Brouwers N, Van Cauwenberghe C, Engelborghs S, et al. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. Mol Psychiatry. 2012;17(2):223-233.
5. Truty R, Paul J, Kennemer M, et al. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet Med. 2019;21(1):114-123.
6. Takeshima Y, Yagi M, Okizuka Y, et al. Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. J Hum Genet. 2010;55(6):379-388.

7. Magri F, Govoni A, D'Angelo MG, et al. Genotype and phenotype characterization in a large dystrophinopathic cohort with extended follow-up. J Neurol. 2011;258(9):1610-1623.
8. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol. 2019;15(5):e1007069.
9. Schenkel LC, Kerkhof J, Stuart A, et al. Clinical Next-Generation Sequencing Pipeline Outperforms a Combined Approach Using Sanger Sequencing and Multiplex Ligation-Dependent Probe Amplification in Targeted Gene Panel Analysis. J Mol Diagn. 2016;18(5):657-667.
10. Stuppia L, Antonucci I, Palka G, Gatta V. Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases. Int J Mol Sci. 2012;13(3):3245-3276.
11. Alfares AA, Kelly MA, McDermott G, et al. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. Genet Med. 2015;17(11):880-888.
12. Kurian AW, Hare EE, Mills MA, et al. Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk

assessment. J Clin Oncol. 2014;32(19):2001-2009.

13. Tang W, Qian D, Ahmad S, et al. A low-cost exon capture method suitable for large-scale screening of genetic deafness by the massively-parallel sequencing approach. Genet Test Mol Biomarkers. 2012;16(6):536-542.

14. Chang LC, Das B, Lih CJ, et al. RefCNV: Identification of Gene-Based Copy Number Variants Using Whole Exome Sequencing. Cancer Inform. 2016;15:65-71.

15. Chiang T, Liu X, Wu TJ, et al. Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. Genet Med. 2019;21(9):2135-2144.

16. Fowler A, Mahamdallie S, Ruark E, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016;1:20.

17. Kang Y, Nam SH, Park KS, et al. DeviCNV: detection and visualization of exon-level copy number variants in targeted next-generation sequencing data. BMC Bioinformatics. 2018;19(1):381.

18. Kong J, Shin J, Won J, Lee K, Lee U, Yoon J. ExCNVSS: A Noise-Robust Method for Copy Number Variation Detection in Whole Exome Sequencing Data. Biomed Res Int. 2017;2017:9631282.

19. Malekpour SA, Pezeshk H, Sadeghi M. MSeq-CNV: accurate detection of Copy Number Variation from Sequencing of Multiple samples. Sci Rep. 2018;8(1):4009.

20. Marchuk DS, Crooks K, Strande N, et al. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. PLoS One. 2018;13(12):e0209185.

21. Markham JF, Yerneni S, Ryland GL, et al. CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. Sci Rep. 2019;9(1):6426.

22. Onsongo G, Baughn LB, Bower M, et al. CNV-RF Is a Random Forest-Based Copy Number Variation Detection Method Using Next-Generation Sequencing. J Mol Diagn. 2016;18(6):872-881.

23. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019;29(7):1134-1143.

24. Pugh TJ, Amr SS, Bowser MJ, et al. VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. Genet Med. 2016;18(7):712-719.

25. Roca I, Gonzalez-Castro L, Maynou J, et al. PattRec: An easy-to-use CNV detection tool optimized for targeted NGS assays with diagnostic purposes. Genomics. 2020;112(2):1245-1256.

26. Yao R, Zhang C, Yu T, et al. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. Mol Cytogenet. 2017;10:30.

27. Garcia-Garcia G, Baux D, Faugere V, et al. Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep. 2016;6:20948.

28. Mahamdallie S, Ruark E, Yost S, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome Open Res. 2017;2:35.

29. Unal I. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. Comput Math Methods Med. 2017;2017:3762651.

30. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J. 2005;47(4):458-472.

31. Chilamakuri CS, Lorenz S, Madoui MA, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014;15:449.

32. Kerkhof J, Schenkel LC, Reilly J, et al. Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. J Mol Diagn. 2017;19(6):905-920.

33. Endrullat C, Glokler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genom. 2016;10:2-9.

국문초록

서론: 차세대 염기서열분석 (Next-generation sequencing, NGS) 는 한 번에 수많은 유전자를 한 번에 검사하면서도 우수한 성능으로 임상현장에서 활발히 사용되고 있다. 하지만, 데이터를 활용한 복제수변이 분석에 대한 표준 알고리즘은 현재까지 부재한 상태다. 본 연구자는 NGS 데이터로부터 엑손 수준의 복제수변이를 검출하는 알고리즘을 확립하고 이를 Multiple ligation-dependent probe amplification (MLPA)로 확인된 임상검체를 이용해 검증하여 임상적용이 가능하도록 하고자 하였다.

방법: 본 연구자는 NGS 데이터로부터 엑손 수준의 복제수변이를 검출하기 위해 "Consistent Count Region" (CCR)라는 알고리즘을 고안하였다. 본 알고리즘은 대조군에서 일정하게 타겟의 리드뎁스와 정해진 범위 안에 존재하는 엑손을 CCR로 정의한다. CCR 리드뎁스 평균값으로 타겟의 리드뎁스를 나눠준 값을 이용해 복제수변이를 검출한다. 우선 CCR 알고리즘의 각 파라미터와 컷오프를 Receiver operating characteristic (ROC) 분석과 컷오프분석 (Youden Index)을 통해 최적화한다. CCR 알고리즘을 MLPA로 복제수변이의 유무가 확인된 다양한 NGS 데이터에서 임상적 효용성을 검증한다.

결과: CCR-CNV 알고리즘의 전체 민감도는 99.7%로 기존에 알려진 DECoN, Atlas-CNV, CNV-RF과 같은 알고리즘의 민감도보다 높았다. 특이도는 98.1%로 타 알고리즘과 비슷하거나 더 높은 수치를 보였다.

특히, CCR-CNV의 false discovery rate (FDR)은 기존 알고리즘과 비슷한 수준을 보였다.

결론: 본 연구자는 엑손 수준의 복제수 변이를 높은 정확도로 검출할 수 있는 알고리즘을 고안하였다. 본 알고리즘은 다양한 시약과 플랫폼에서 만들어진 NGS 데이터의 복제수 변이를 검출할 수 있음을 보였다.

Keyword : 복제수변이; germ-line; 분자유전; targeted gene panel clinical sequencing

학번 : 2016-21960