European Review for Medical and Pharmacological Sciences 2021; 25 (1 Suppl): 7-13

CNV analysis in a diagnostic setting using target panel

E. SORRENTINO[1], M. DAJA[1], F. CRISTOFOLI[1], S. PAOLACCI[2], M. BERTELLI[1,2], G. MARCEDDU[1]

[1]MAGI EUREGIO, Bolzano, Italy
[2]MAGI'S LAB, Rovereto (TN), Italy

Abstract. – OBJECTIVE: Copy-number variation (CNV) is an important source of genetic diversity in humans. It can cause Mendelian or sporadic traits or be associated with complex diseases by various molecular mechanisms, including gene dosage, gene disruption, gene fusion and position effects. In clinical diagnostics, it is therefore fundamental to be able to identify such variations. The preferred techniques for CNV detection are MLPA, aCGH and qPCR, which have proven to be valuable, and they are complex, costly and require prior knowledge of the region to analyze. CNV calling from NGS data still suffers from data variability. Coverage can vary greatly from one region of the genome to another, depending on many factors like complexity, GC content, repeated regions and many others. In this paper, we describe how we developed a method for CNV detection.

MATERIALS AND METHODS: Our method exploits CoNVaDING to detect single- and multiple-exon CNVs in targeted NGS data.

RESULTS: We demonstrated that our CNV analysis has 100% specificity and 99.998% sensitivity. We also show how we evaluated the performance of this method based on internal analysis.

CONCLUSIONS: The results indicate that the method can be used to screen prior to standard labs technologies, thus reducing the number of analyses, as well as costs, and increasing test conclusiveness.

Key Words:
Copy number variant, Diagnostics, Target panel.

Introduction

Next Generation Sequencing (NGS), or massive parallel sequencing, is a DNA sequencing technology which has revolutionized genomic research. NGS can detect single-nucleotide variants and small deletions and insertions, but detection of large rearrangements, such as copy-number variants (CNVs), remains challenging. The method has several intrinsic issues, including short read lengths and GC-content bias[1]. Germline CNVs are known to be a source of genetic diversity in humans but also the cause of various hereditary diseases, both common and complex[2], influencing a variety of Mendelian and somatic genetic disorders.

To detect CNVs in genetic diagnostics, the methods used are usually multiplex ligation-dependent probe amplification (MLPA)[3], array comparative genomic hybridization (aCGH)[4] and qPCR[5]. These methods have many drawbacks, being complex, costly and requiring prior knowledge of the region to analyze. Thus, testing is usually only performed on a subset of genes.

Capacity to identify CNVs, in particular from NGS data, would be fundamental for increasing diagnostic yield and improving clinical management[6]. The possibility of preliminary screening, prior to testing with more complex and costly techniques, could improve diagnostic conclusiveness for certain diseases, while reducing health expenditure due to late diagnosis and the use of standard techniques.

Since there are many tools for CNV detection in NGS data, we decided to search the literature to find the best tool for our purposes.

In 2020, Koboldt[7] discussed several tools for CNV detection in NGS data, such as cn.MOPS[8], CONTRA[9], CoNVEX[10], ExomeCNV[11] and XHMM[12]. Zhao et al[13] described CoNIFER[14] and XHMM as good tools for rare CNV detection. To choose the best tool for our purpose, we extracted the performances declared by the developers of each tool. We also considered other aspects, such as type of NGS data (WES or gene panels), the biological samples required to obtain the data, and how the tools can be integrat-

Corresponding Author: Stefano Paolacci, Ph.D; e-mail: stefano.paolacci@assomagi.org

ed into our existing pipeline. Cn.MOPS, which is distributed as an R or Bioconductor package, declares a recall of 88% for gains and 96% for losses. CONTRA calculates its performance on exome capture data and declares 100% specificity and 96.4% sensitivity. CoNVEX reduces the variability of coverage ratios, and then, uses HMM to detect CNVs; it has been tested with tumor data vs. control WES data, obtaining a sensitivity of 92% and a precision of 50%. ExomeCNV declares sensitivity and specificity of more than 90% on melanoma WES data, but it also declares that the tool does not perform optimally when cases and controls have significantly different coverages. CoNIFER was tested on three or more consecutive exons using HapMap WES samples, obtaining a precision of 94% and an accuracy of 78%. XHMM was tested on WES data and reports a sensitivity of 79%.

A tool that can perform CNV detection for NGS panel data is CoNVaDING[15]. Its developers compared it with XHMM, CoNIFER, CONTRA and CODEX, using 320 gene panel data samples, achieving a sensitivity of 100% and a specificity of 99.998% and outperforming the other tools on the same samples. The four tools used for comparison considered control samples equally informative but differed in PCR and capturing efficiency. This increases the risk of low sensitivity and specificity for single-exon CNV detection or limits analysis to detection of variations that span multiple exons[15]. With stringent quality control and selection of samples based on coverage patterns, CoNVaDING selects controls most similar to the sample, improving performance. Based on these characteristics, we chose this tool, which we tested it by including it in the pipeline we use for analysis of NGS data in our laboratory. Considering the high coverage standardization requirements of CNV detection, we added a prior PCA to identify control samples with variability most similar to the sample under analysis. This method was integrated in our diagnostic pipeline, as a screening test. Therefore, to be compliant with the stringent diagnostic requirements, we also evaluated the performance of this method using internal data.

Materials and Methods

In this section we describe the CoNVaDING workflow, and the procedure used to obtain the coverage data of different samples. The method is composed of sequencing to capture target regions, bioinformatics analysis to obtain coverage data, and analysis of the coverage data with CoNVaDING.

Sequencing

We captured genomic regions of interest using Illumina Nextera Flex for Enrichment, a solution enrichment system, according to the supplier's protocol. Briefly, 50 ng genomic DNA was fragmented by the enzyme method (Tagmentation), which uses the enrichment Bead-Linked Transposomes (eBLT) system. This system fragments the DNA in a single step and adds the adapters necessary for the subsequent amplification steps to the ends of fragments. The DNA library is then purified and amplified in nine PCR cycles. During amplification, the IDT for Illumina Nextera Unique Dual Indexes (sequences necessary for sequencing), and the common adapters (P5 and P7, fundamental for cluster generation and sequencing) were added. Subsequently, multiple libraries of different samples (containing different indexes) were combined, and then, hybridized in solution with the probes (Ocular panel Id: 140115) to capture the target regions in a single hybridization and capture step. Finally, a second PCR step of 15 cycles for enrichment of the captured DNA fragments is performed.

Bioinformatic Analysis

Bioinformatic analysis of the sequences generated to identify variants with a possible pathogenic role involves the coding regions and splice-site-flanking regions (±5 bp flanking each exon) of the specific subset of genes of the suspected diagnosis, as indicated by the test. The details of bioinformatics analysis are available in Marceddu et al[16]. Briefly, the generated sequences are mapped against the reference sequence to obtain a list of variants. The position, amino acid change and predicted effect on protein function (SIFT, PolyPhen-2) of the variants identified are recorded. Other information on the variants is obtained from various public sources, including NCBI, PubMed and other specific databases.

Once annotation is complete, the variants are filtered to distinguish common benign ones from rare, possibly pathogenic variants. Population frequency data, such as dbSNP[17], NHLBI Exome Sequencing Project[18] and 1000 Genomes Project[19], available on the web, is used to classify benign variants.

The Magi APP Bioinformatics Pipeline uses assembly hg38 as reference genome. The positions of the variants selected from the public data are therefore converted from assembly hg19 to hg38, using UCSC lift over[20], an online or command line tool that converts genome coordinates and genome annotation files from one assembly to another.

CoNVaDING

CoNVaDING[15] is a tool that not only detects (single- and multiple-exon) CNVs in targeted NGS data but also provides a stringent quality control metric that distinguishes high-quality samples and targets from low-quality ones. Its workflow includes several steps: selecting controls, calculating average coverage, coverage analysis and CNV prediction.

For in-depth comparison, CoNVaDING uses control samples captured with the same panel and that went through the same bioinformatic analysis as the samples to analyze. The tool selects control samples with similar coverage patterns to the ones in analysis. Then, it calculates the average coverage for each target region and normalizes it using two methods in parallel: average coverage of all autosomal targets and of the same gene. Based on the normalizations, ratio and distribution analysis using Z-scores is performed to obtain the relative differences in average coverage of the targets of samples and controls, from which CNV is predicted[15].

We decided to add a Principal Component Analysis (PCA) selection step on control samples before input to CoNVaDING in order to identify those with variability most similar to the samples in analysis (Figure 1). PCA is a technique for reducing the dimensions of feature space, facilitating interpretation while minimizing information loss.

Implementation

Analysis of CNV in NGS gene panel data was performed by entering CoNVaDING, running in Ubuntu 16.04, in the existing PipeMAGI pipeline. We used CoNVaDING version 1.2.0[21], which is written in Perl[22] and depends on specific Perl libraries, as well as on Samtools[23] version 1.3 or higher. The method is written in Python 2.7[24] and requires several modules, such as Pandas[25] for data manipulation, Matplotlib[26] for data plotting and Scikit-learn[27] for PCA selection. The code of our method can be found in our GitLab repository at https://gitlab.com/magieuregio2016/cnv_analysis.

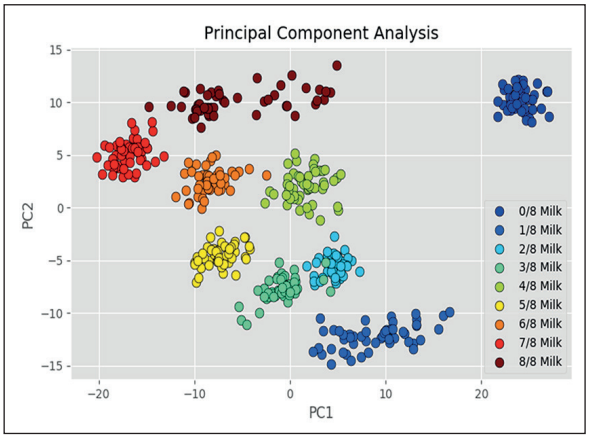

Figure 1. Plot of the first two principal components of the control group.

Results

Data and Design

To test the method with CoNVaDING added to our pipeline, we performed a validation with 12 genotyped samples containing deletions and duplications confirmed by MLPA or obtained from the NIGMS Human Genetic Cell Repository, a collection of well-characterized, high-quality human cells for use in biomedical research, made available by the Coriell Institute[28]. The specimens, equally from males and females, were acquired from individuals with inherited diseases, apparently healthy individuals and individuals with different geographic origins.

Table I shows the validation design: of the 12 samples analyzed, eight were confirmed with MLPA while four were Coriell samples. One sample had a mono-exon deletion while the others were known to have multi-exon deletions or duplications. We used 389 samples as control group.

Parameters

To evaluate the performance of our method, we chose the parameters sensitivity, specificity and accuracy. To calculate the parameters, we used two modes to measure the ability of the method to find deletions and duplications on single exons and entire duplications or deletions. Single exons or entire variants were divided into four classes: *false positive* (FP), *false negative* (FN), *true positive* (TP) and *true negative* (TN) in order to calculate the parameters.

Table I. Validation design of 12 samples analysed, eight were confirmed by MLPA and the other four were Coriell samples. One sample had a mono-exon deletion, while the others were known to have multi-exon deletions or duplications.

ID	Origin	Gene	Indel
RX26.2019	Internal MLPA	EYS	Deletion exons 16-19
RX27.2019	Internal MLPA	EYS	Deletion exons 14-15
RX28.2019	Internal MLPA	USH2A	Deletion exon 13
RX29.2019	Internal MLPA	EYS	Deletion exons 14-22
RX30.2019	Internal MLPA	EYS	Deletion exons 14-22
RX31.2019	Internal MLPA	USH2A	Deletion exons 14-15
RX32.2019	Internal MLPA	ABCA4	Deletion exons 1-5
R2325.2020	NA00214	CRB1	Deletion of entire gene
R2326.2020	HG01802	LCA5	Duplication of entire gene
R2327.2020	HG02397	CRX	Duplication of entire gene
R2328.2020	NA10946	LCA5	Deletion of entire gene
RE1628.2020	Internal MLPA	USH2A	Deletion exons 5-10

Sensitivity is the ability to correctly identify variants that really exist in the sample and its formula is:

$$\frac{TP}{TP + FN}$$

Specificity is the ability *not* to call variants that do not exist in the sample; its formula is:

$$\frac{TN}{TN + FP}$$

Accuracy is the ability of the method to identify or exclude the existence of a variant, and its formula is:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

For variant-call analysis, we assigned a variant to the true positive group even if it was called partially. We assigned it to the false negative group if it was not called. Variants erroneously called were considered false positives and variants correctly *not* called were considered true negatives.

Figures 2 and 3 show the results of validation, reporting parameter calculations by the single-exon method and by calling of entire duplications and deletions, respectively. The specificity and accuracy of the method reached 99% for both methods. Sensitivity reached 100% for entire variant calling and only 83% for single-exon variants. This performance makes the method suitable for multi-exon CNV screening.

Discussion

Copy number variation is an important source of genetic diversity in humans, and can cause Mendelian or sporadic traits, or be associated with complex diseases. It is therefore fundamental in clinical diagnostics to be able to identify such variations. The techniques of choice for CNV detection are complex, costly, and require prior knowledge of the region to analyze, while CNV calling from NGS data still suffers from the variability of such data.

Here we described how we integrated CoNVaDING in our bioinformatic pipeline. CoNVaDING is a tool that detects single- and multiple-exon CNVs in targeted NGS data, after performing stringent quality control.

As screening in a diagnostic setting requires good performance, we tested the performance of our method on internal data. We also describe how we tested the tool for single- and multiple-exon CNV detection. The method was tested on 12 samples (with 389 internal control samples), whose deletions and duplications were confirmed by standard techniques. We calculated the sensitivity, specificity and accuracy of the tool by two methods: one measured the ability of the method to find deletions and duplications

	VP	VN	FP	FN	Sensitivity	Specificity	Accuracy
RX26.2019	4	3998	0	0	1.0000	1.0000	1.0000
RX27.2019	2	3998	2	0	1.0000	0.9995	0.9995
RX28.2019	0	4000	1	1	0.0000	0.9998	0.9995
RX29.2019	9	3983	10	0	1.0000	0.9975	0.9975
RX30.2019	9	3988	5	0	1.0000	0.9987	0.9988
RX31.2019	1	4001	0	0	1.0000	1.0000	1.0000
RX32.2019	3	3991	6	2	0.6000	0.9985	0.9980
R2325.2020	12	3968	22	0	1.0000	0.9945	0.9945
R2326.2020	5	3995	0	2	0.7143	1.0000	0.9995
R2327.2020	2	3997	2	1	0.6667	0.9995	0.9993
R2328.2020	7	3980	15	0	1.0000	0.9962	0.9963
RE1628.2020	5	3997	0	0	1.0000	1.0000	1.0000
			Mean		0.8317	0.9987	0.9986
			Standard deviation		0.3035	0.0018	0.0017

Figure 2. Validation results when parameters were calculated by the single-exon method on 12 samples.

on single exons and the other the ability of the method to find deletions and duplications on multiple exons.

Conclusions

The results showed good performance for multi-exon CNV detection, qualifying the method for a screening phase, which will in any case be confirmed by a complementary diagnostic method. Our method reduces the number of analyses necessary and therefore reduces costs while increasing test conclusiveness.

Conflict of Interest

The Authors E. Sorrentino, M. Daja, F. Cristofoli, M. Bertelli and G. Marceddu are employed at MAGI EUREGIO.

References

1) Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 2012; 28: 2711-2718.
2) Aouiche C, Shang X, Chen B. Copy number variation related disease genes. Quant Biol 2018; 6: 99-112.
3) Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, Rupar CA, Adams

	VP	VN	FP	FN	Sensitivity	Specificity	Accuracy
RX26.2019	1	289	0	0	1.0000	1.0000	1.0000
RX27.2019	1	288	1	0	1.0000	0.9965	0.9966
RX29.2019	1	287	2	0	1.0000	0.9931	0.9931
RX30.2019	1	286	3	0	1.0000	0.9896	0.9897
RX31.2019	1	289	0	0	1.0000	1.0000	1.0000
RX32.2019	1	283	6	0	1.0000	0.9792	0.9793
R2325.2020	1	288	1	0	1.0000	0.9965	0.9966
R2326.2020	1	289	0	0	1.0000	1.0000	1.0000
R2327.2020	1	287	2	0	1.0000	0.9931	0.9931
R2328.2020	1	282	7	0	1.0000	0.9758	0.9759
RE1628.2020	1	289	0	0	1.0000	1.0000	1.0000
			Mean		1.0000	0.9931	0.9931
			Standard deviation		0.0000	0.0085	0.0084

Figure 3. Validation results when parameters were calculated by calling entire duplications and deletions.

P, Hegele RA, Lin H, Rodenhiser D, Knoll J, Ainsworth PJ, Sadikovic B. Clinical validation of copy number variant detection from targeted next-generation sequencing panels. J Mol Diagn 2017; 19: 905-920.

4) Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998; 20: 207-211.
5) Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996; 6: 986-994.
6) Ellingford JM, Campbell C, Barton S, Bhaskar S, Gupta S, Taylor RL, Sergouniotis PI, Horn B, Lamb JA, Michaelides M, Webster AR, Newman WG, Panda B, Ramsden SC, Black GC. Validation of copy number variation analysis for next-generation sequencing diagnostics. Eur J Hum Genet 2017; 25: 719-724.
7) Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med 2020; 12: 91.
8) Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 2012; 40: e69.
9) Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. CONTRA: Copy number analysis for targeted resequencing. Bioinformatics 2012; 28: 1307-1313.

10) Amarasinghe KC, Li J, Halgamuge SK. CoNVEX: Copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 2013; 14: S2.
11) Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 2011; 27: 2648-2654.
12) Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, Hultman CM, Sklar P, Purcell SM. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 2012; 91: 597-607.
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 2013; 14: S1.
14) Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP; NHLBI Exome Sequencing Project, Quinlan AR, Nickerson DA, Eichler EE. Copy number variation detection and genotyping from exome sequence data. Genome Res 2012; 22: 1525-1532.
15) Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JD, van der Hout AH, Westers H, Sinke RJ, Swertz MA, Sijmons RH, Sikkema-Raddatz B. CoNVaDING: Single exon variation detection in targeted NGS data. Hum Mutat 2016; 37: 457-464.
16) Marceddu G, Dallavilla T, Guerri G, Manara E, Chiurazzi P, Bertelli M. PipeMAGI: An integrated and validated workflow for analysis of NGS data for clinical diagnostics. Eur Rev Med Pharmacol Sci 2019; 23: 6753-6765.
17) COVID-19 information. Available at: http://www.ncbi.nlm.nih.gov/projects/SNP.
18) NHLBI Exome Sequencing Project (ESP). Exome Variant Server. Available at: https://evs.gs.washington.edu/EVS/.
19) IGSR: The International Genome Sample Resource. Available at: http://www.1000genomes.org.
20) Lift Genome Annotations. Available at: https://genome.ucsc.edu/cgi-bin/hgLiftOver.
21) CoNVaDING User Guide. Avalable at: https://molgenis.gitbooks.io/convading/.
22) Perl. Available at: https://www.perl.org/.
23) Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-2079.
24) Van Rossum G, Drake Jr FL. Python tutorial. Amsterdam: Centrum voor Wiskunde en Informatica 1995; 620.
25) Mc Kinney W. Data structures for statistical computing in Python. Proc of the 9th Python in Science Conf 2010; 51-56.
26) Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007; 9: 90-95.
27) Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine learning in Python 2011; 12: 2825-2830.
28) NIGMS Human Genetic Cell Repository. Available at: https://www.nigms.nih.gov/Research/specificareas/hgcr/Pages/default.aspx.