The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine

eJIFCC

In *silico* Copy Number Variation (CNVs) bioinformatics estimation: dream or nightmare?

Leandro Gutiérrez, Lara Parada-Fennen, Angela Rosaria Solano

Laboratorio de Genotipificación, Centro de Educación Médica e Investigaciones Clínicas "Norberto Quirno" (CEMIC), Ciudad Autónoma de Buenos Aires, Argentina

ARTICLE INFO

Corresponding author:
Leandro Gutiérrez, PhD
Laboratorio de Genotipificación
Centro de Educación Médica e Investigaciones Clínicas "Norberto Quirno" (CEMIC)
Ciudad Autónoma de Buenos Aires
Argentina
Phone: +5411 5299-0100 (ext 2914)
E-mail: lgutierrez@cemic.edu.ar

Key words:
next generation sequencing,
copy number variation, bioinformatics

LETTER TO THE EDITOR

Decades before the availability of next generation sequencing (NGS) technology, definition of copy number variations (CNVs) in human genetics were mainly rare changes in the quantity and structure of chromosomes. These included aneuploidies and rearrangements (1, 2, 3). Subsequently, with the advent of molecular technology, smaller and more abundant alterations were observed, including, various repetitive elements that involve short DNA sequences (micro and mini-satellites), insertions, deletions and duplications (4).

Targeted next-generation sequencing (NGS) is an established, but not the only, method for the detection of germline variants in cancer predisposition genes. While variants involving a few nucleotides, i.e., single-nucleotide variants (SNVs) and short insertion/deletion events (indels), can be detected accurately, the identification of larger genomic rearrangements (copy number variations (CNVs)) remains a challenge.

At present time, CNVs is considered a segment of DNA that is present at a variable copy number in comparison with a reference genome. They can derive from duplications, deletions, insertions and even translocations, and can vary in length, may be short or include thousands of bases (5, 6, 7, 8); for this research, it could be more adequate an average size of ~100 bp MLPA resolution level, as a parameter for defining CNV length. Several *in silico* tools have been developed to predict CNVs using targeted NGS data. However, several studies suggested that existing tools for CNV detection using targeted NGS data show limited accuracy and robustness (9, 10, 11, 12).

We investigated the performances of *in silico* CNV commercial prediction tool Celemics CNV Analysis Algorithm® in 13 cancer predisposition genes: *APC* (NM_000038.6), *ATM* (NM_000051.4), *BRCA1* (NM_007294.4), *BRCA2* (NM_000059.4), *CDH1* (NM_004360.5), *CHEK2* (NM_007194.4), *EPCAM* (NM_002354.3), *MLH1* (NM_000249.4), *MSH2* (NM_002878.4), *MSH6* (NM_000179.3), *MUTYH* (NM_001048174.2), *PALB2* (NM_024675.4), and *STK11* (NM_000546.5), evaluated in 80 patients with hereditary cancer syndrome, for those of who had the results by multiplex ligation-dependent probe amplification (MLPA) as the assay for the variation in copy number.

In this analysis, the algorithm predicted 8 CNVs, of which 1 (12.5%) it was a real CNV (exons 1 to 7 in *MSH2* gene). The remaining 7 (87.5%) were false positive (were not detected by MLPA). False positive predictions affected target genes: *APC* (figure 1)*, BRCA1, BRCA2, CDH1, MSH2* and *PALB2* without a clear predisposition for a gene o region.

The overall real CNV prevalence was 6.25% (5/80) (*MSH2* (n=3), *APC* (n=1) and *EPCAM* (n=1). Of these, 4 true positive CNVs were none predicted by CNV analysis algorithm.

As other *in silico* CNV prediction tools, the Celemics CNV algorithm uses read depth-based approaches. CNV is based on the hypothesis that a CNV determines the relative read depth per target region. Thus, low or high fluctuating read depths of a target region will likely affect accurate CNV prediction.

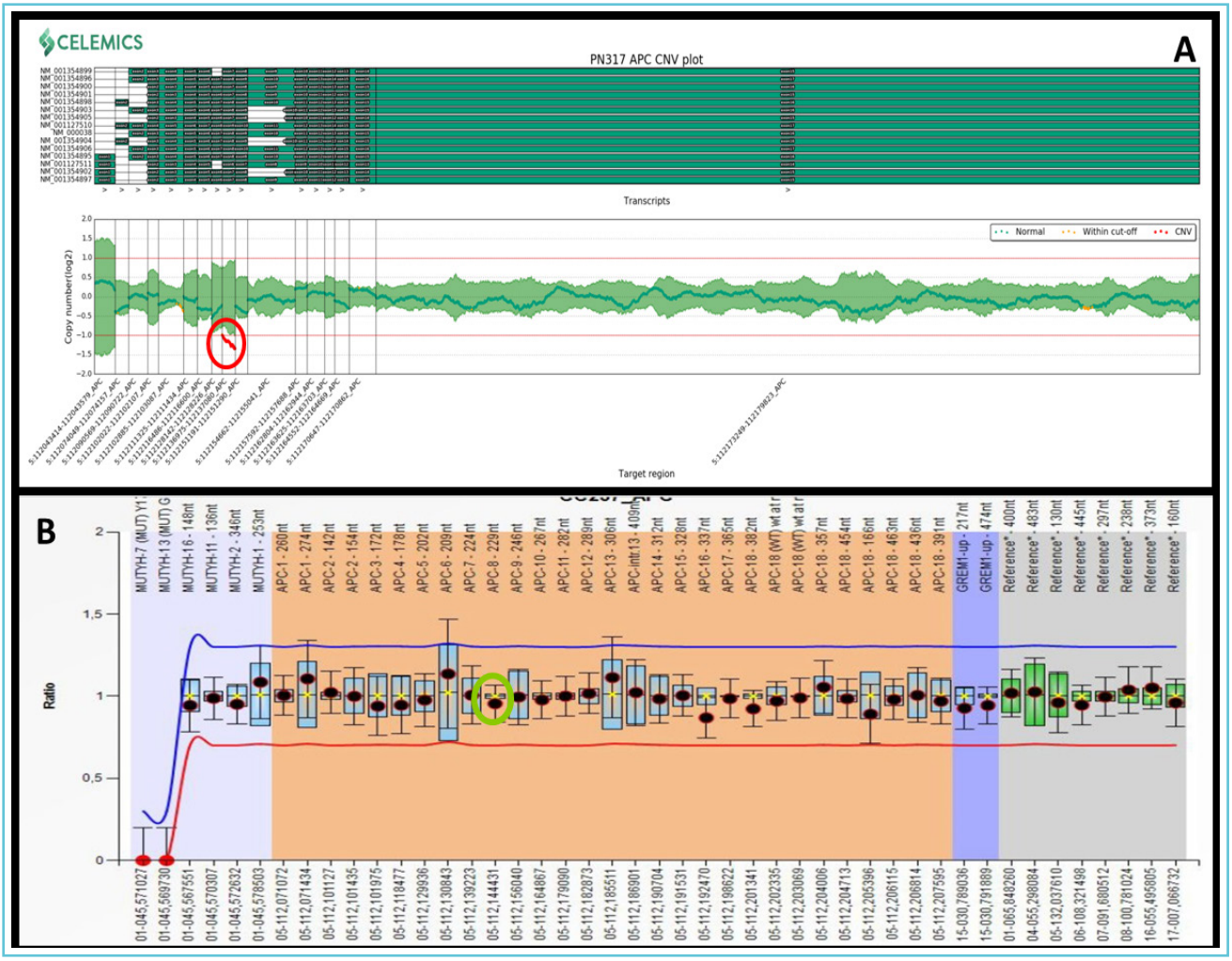
In this scenario, the probability of CNV analysis algorithm prediction representing a true positive CNV, its positive predictive value (PPV), was 12.5% (1/8). Although the series analysed is small, this value represents an important limitation to use the bioinformatics' estimation of CNV as the only analysis tool.

Comparing our data with those by Lepkes et al (N= 4208), we found that, the PPV values can vary greatly on the basis of different calculation algorithms. In their analysis, Lepkes et al. compared four bioinformatics calculation algorithms (the commercial tool incorporated in the CE-IVD-marked Sophia Genetics DDM pipeline®, and three publicly available tools, ExomeDepth, GATK gCNV and panelcn.MOPS) and established that the PPV values of these bioinformatics tools can vary between 7% to 68%, showing that there may be a great difference between the values of CNVs predicted by algorithms and their real existence (13).

The most relevant hypothesis at present explaining the great differences found between predicted and real CNVs strongly suggest that target region sequencing coverage along with target region characteristics, such as GC content, length, low sequencing coverage, determined the accumulation of false positive CNV predictions (13, 14, 15).

Future directions are strongly orientated to improve the use of CNVs NGS-derived information. However, verification of *in silico* predicted CNVs is required due to its high frequencies of false positive predictions.

Figure 2 **Graphic representations of the calculation of a deletion (*in silico*) and MLPA analysis (*in vitro*) for exon 8 in *APC* gene***



** A) Graphic representation of the calculation of a deletion (in silico predicted CNV, red circle) in exon 8 of the APC gene. B) Graphic representation of the exon analysis of the APC gene by MLPA (in vitro) showing the absence of a deletion in exon 8 (green circle).*

❖❖❖❖❖

Research funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests

Authors state no conflict of interest.

Ethical approval: Not applicable.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

❖❖❖❖❖

REFERENCES

1. Edwards, J. H., Harnden, D. G., Cameron, A. H., et al. A new trisomic syndrome. Lancet 1, 787–790 (1960).

2. Patau, K., Smith, D. W., Therman, E., et al. Multiple congenital anomaly caused by an extra autosome. Lancet 1, 790–793 (1960).

3. Bobrow, M., Joness, L. F. & Clarke, G. A complex chromosomal rearrangement with formation of a ring 4. J. Med. Genet. 8, 235–239 (1971).

4. Wright, A. F. in the Nature Encyclopedia of the Human Genome 2, 959–968 (Nature Publishing Group, London (2003).

5. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. Genetics, Nature Review 7, 85-97 (2006).

6. Zhang, F., Gu, W., Hurles, M.E., et al. Copy Number Variation in Human Health, Disease, and Evolution. Annu. Rev. Genomics Hum. Genet. 10, 451–481 (2009).

7. Gilchrist, D.A. https://www.genome.gov/genetics-glossary/Copy-Number-Variation (2023).

8. Carter, N., Church, D., Feuk L., et al. CNV Database of Genomic Variants. The Centre for Applied Genomics The Hospital for Sick Children Peter Gilgan Centre for Research and Learning, Canada. http://projects.tcag.ca/variation/ (2022).

9. Ruderfer, D.M., Hamamsy, T., Lek, M., et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat. Genet. 48, 1107–1111 (2016).

10. Hong, C.S., Singh, L.N., Mullikin, J.C., et al. Assessing the reproducibility of exome copy number variations predictions. Genome Med. 8, 82 (2016).

11. Roca, I., González-Castro, L., Fernández, H., et al. Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutat. Res. 779, 114–125 (2019).

12. Moreno-Cabrera, J.M., Del Valle, J., Castellanos, E., et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur. J. Hum. Genet. 28, 1645–1655 (2020).

13. Lepkes, L., Kayali, M., Blümcke, B., et al. Performance of In Silico Prediction Tools for the Detection of Germline Copy Number Variations in Cancer Predisposition Genes in 4208 Female Index Patients with Familial Breast and Ovarian Cancer. Cancers 13, 118 (2021).

14. Kuśmirek, W. and Nowak, R. CNVind: an open source cloud-based pipeline for rare CNVs detection in whole exome sequencing data based on the depth of coverage. BMC Bioinformatics. 23, 85 (2022).

15. Wan-Ping, L., Qihui, Z., Xiaofei, Y., Liu, S., Cerveira, E., Ryan, M., Mil-Homens, A., Bellfy, L., Ye, K., Lee, C. and Zhang, C. JAX-CNV: A Whole Genome Sequencing-based Algorithm for Copy Number Detection at Clinical Grade Level. Genomics Proteomics Bioinformatics. S1672-0229(22)00005-5 (2022).