Different Strategies for Counting the Depth of Coverage in Copy Number Variation Calling Tools

Wiktor Kuśmirek

Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland.

Bioinformatics and Biology Insights
Volume 16: 1–9

DOI: 10.1177/11779322221115534

ABSTRACT: There are many copy number variation (CNV) detection tools based on the depth of coverage. A characteristic feature of all tools based on the depth of coverage is the first stage of data processing—counting the depth of coverage in the investigated sequencing regions. However, each tool implements this stage in a slightly different way. Herein, we used data from the 1000 Genomes Project to present the impact of another depth of coverage counting strategies on the results of the CNVs detection process. In the study, we used 7 CNV calling tools: CODEX, CANOES, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind; from each of these applications, we separated the process of counting the depth of coverage into independent modules. Then, we counted the depth of coverage by mentioned modules, and finally, the obtained depth of coverage tables were used as the input data set to other CNV calling tools. The performed experiments showed that the best methods of counting the depth of coverage are the algorithms implemented in the CLAMMS and CNVkit applications. Both ways allow obtaining much better sets of detected CNVs compared to counting the depth of coverage implemented in other tools. What is more, some CNV detection tools are reasonably resistant to changing the input depth of coverage table. In this study, we proved that the exomeCopy application gives an approximately similar set of the resulting rare CNVs, regardless of the method of counting the depth of coverage table.

KEYWORDS: Depth of coverage, copy number variation, whole-exome sequencing, next-generation sequencing

RECEIVED: April 7, 2022. **ACCEPTED:** July 2, 2022.

TYPE: Original Research Article

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported by the Polish National Science Center grant Preludium 2019/35/N/ST6/01983. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Wiktor Kuśmirek, Institute of Computer Science, Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland. Email: wiktor.kusmirek@pw.edu.pl

Introduction

Copy number variation (CNV) is a disturbance in the number of copies of a given DNA fragment[1] and may be one of the following types: deletion or duplication.[2] Currently, the subject of the occurrence of CNV in the human genome is being researched more and more often[3] because CNVs have been identified as a significant cause of structural variation in the genome, involving both duplications and deletions of sequences.[4-6] For example, Li et al[7] described strong associations between rare CNVs and 4 major disease categories, including autoimmune, cardio-metabolic, oncologic, and neurological/psychiatric diseases. Despite the great importance of detecting rare CNVs in the human genome, the current CNV detection tools are characterized by insufficient performance.[8-11]

There are 5 main strategies for detecting CNVs in whole-exome sequencing (WES) data[12]: (1) paired-end mapping, (2) split read, (3) read depth, (4) *de novo* assembly of a genome, and (5) combination of the above approaches. The most popular method is read depth (26 tools reported by Zhao et al[12]), mainly due to the popularity of high-coverage next-generation sequencing (NGS) data.

A typical pipeline of detecting CNVs based on read depth consists of several steps: (1) mapping DNA reads to reference genome, (2) counting number of mapped reads in sequencing regions, (3) quality control, (4) selecting reference sample set, (5) normalizing read depth, and (6) segmentation and calling CNVs. Even though some of the applications do not have all of the steps listed above, eg, selecting a reference sample set, there is a typical stage for all CNV callers: counting the number of mapped DNA reads in sequencing regions. However, this process in other applications is slightly different, eg, algorithms, parameter values, considered flags, etc; some different counting methods are presented below.

Herein, we present a comparison of strategies for counting the depth of coverage in other CNV calling tools. This article compares the theoretical advantages and disadvantages of various methods of counting the depth of coverage and their influence on the number of CNVs detected by different CNV callers. What is more, theoretical considerations have been confirmed by experiments on a real data set; all scripts, test data, and evaluated applications are available online: https://github.com/wkusmirek/cnv-depth-of-coverage-comparison.

Materials and Methods

Compared applications

As a part of our research, we compared the strategies for calculating the depth of coverage from the CODEX,[13] CNVind,[14] CLAMMS,[15] CANOES,[16] CNVkit,[17] exomeCopy,[18] and ExomeDepth[19] applications. All the algorithms compared are briefly described below and summarized in Table 1.

The CODEX[13] and CNVind[14] tools implement the same strategy for counting the depth of coverage. First, the applications expand each of the sequencing regions by flanking 10 kbp to the left and right. The default CODEX and CNVind algorithms for counting the depth of coverage count reads that are mapped to the specified sequencing region and (1) the

Table 1. Comparison of depth of coverage calculation algorithms implemented in different CNVs detection tools.

	FL. SEQ.	MIN. QUAL.	NORM.	TECHN.
CANOES	–	20	–	bedtools multicov
CODEX	10 kbp	20	–	R
exomeCopy	–	1	–	R
ExomeDepth	–	20	–	R
CLAMMS	–	30	+	samtools bedcov
CNVkit	–	0	+	Python
CNVind	10 kbp	20	–	R

Abbreviations: CNV, copy number variation; Fl. seq., size of flanking sequences; Min. qual., minimum mapping quality; Norm., normalization to the length of the sequencing window; Techn., technology used to implement the depth of coverage calculation module.

mapping quality is greater or equal to 20, (2) the DNA read is not a duplicate, (3) the DNA read passed quality control filters, and (4) the read is the first DNA read or mate-pair reads. If the counting process with the mentioned parameter set does not return any result, then the last filter (the first read of mate-pair) is disabled; the depth of coverage for the investigated sample is recounted (without this filter) one more time.

The CLAMMS[15] application does not implement its own strategy for counting the depth of coverage, but advises the use of external tools: *samtools bedcov* or *GATK DepthOfCoverage*. On the github page, the CLAMMS authors suggest setting the minimum mapping quality for a read to be counted to 30.

The CANOES[16] tool expresses the depth of coverage in a slightly different way—as the total number of DNA reads that intersect with the investigated sequencing region. To count the depth of coverage table, bedtools[20] multicov package with minimum mapping quality set to 20 is used. Other parameters are set to default values. The results include all DNA reads that intersect with the specified sequencing region (even if the start of the read does not belong to the investigated target).

The CNVkit[17] implements the depth of coverage as the number of reads whose origin is mapped in a given sequencing window. The application then normalizes the obtained number by the length of the sequencing window.

The exomeCopy[18] and ExomeDepth[19] tools use the same *countBamInGRanges* function to count the depth of coverage. However, both applications set slightly different parameters for the mentioned function, which significantly affects the depth of coverage. First of all, the minimum mapping quality for exomeCopy is equal to 1, while for ExomeDepth this value is set to 20. Second, exomeCopy counts the DNA read as correctly mapped, with the default settings, if the start of the investigated DNA read is mapped in the specified sequencing region. On the contrary, ExomeDepth counts the DNA reads as correctly mapped if any intersection of the DNA read with the specified sequencing region is present. Moreover, the critical difference between the exomeCopy and ExomeDepth applications is the approach to paired-end tags. The exomeCopy application handles each DNA read independently, without checking that the read is correctly paired and where the second DNA read in the pair is mapped on the reference genome. However, ExomeDepth only considers correctly paired reads; even if both DNA reads in a pair are mapped to the same sequencing region, the depth of coverage for that sequencing region is increased by 1 (not 2).

Workflow

The workflow of the research is presented in Figure 1. Briefly, the data processing began by counting the depth of coverage on each sequencing region by 7 applications: CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind. This process resulted in 7 raw depth of coverage tables, where consecutive samples are in columns, in rows—successive sequencing regions. Then, the quality control process was carried out on each 7 resulting raw depth of coverage tables. Next, each of the 7 depth of coverage tables was given as input of 7 CNVs callers, resulting in 49 CNV sets detected (7 tables of coverage depths × 7 CNVs detection applications). Finally, the resulting sets of CNVs were evaluated. All of the steps listed and the set of test data are characterized later in this section.

The first stage of our research was to count the coverage for each sequencing region in each sample using modules from different CNV callers. In our research, we used the CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind tools. This process resulted in 7 depth of coverage tables. Each of the tables had the same number of rows and columns, numerical values in the tables represent the depth of coverage of a given sample in a given sequencing region.

Then, the quality control process from the CODEX tool was applied on each of the resulting depth of coverage table. Briefly, all targets with median read depth across all samples below 20 or greater than 4000, targets shorter than 20 bp or longer than 2000 bp, with mappability factor below 0.9 and GC content below 20% or greater than 80% were removed. The process was applied to all coverage tables, from each coverage table the same (not passing quality control filters) sequencing regions were removed.

After quality control, we normalized the depth of coverage tables and called CNVs by CODEX, CANOES, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind tools. Each of the reported application was launched 7 times with different input depth of coverage table. What is more, for CODEX and exomeCopy tools, we did not divide samples into groups by

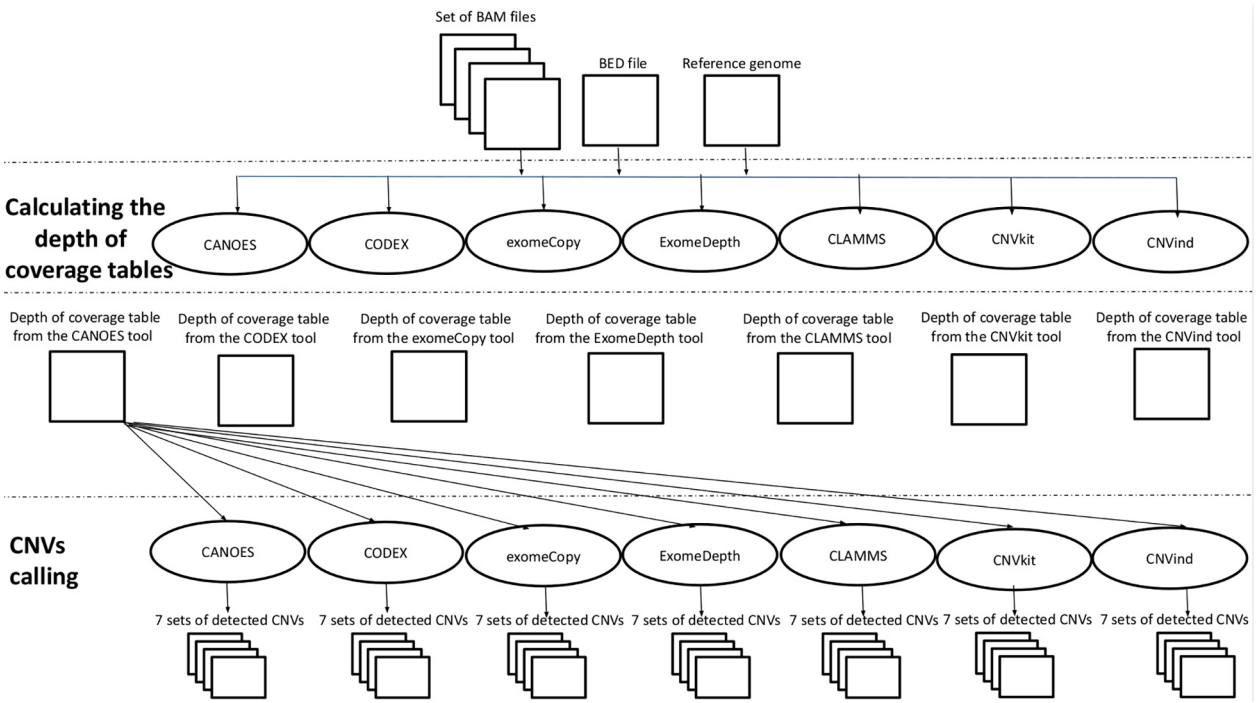

Figure 1. Research workflow. The input data for the study were a set of BAM files with the results of mapping DNA reads to the reference genome, a BED file containing the coordinates of the sequencing regions in WES, and the DNA sequence of the reference genome. The depth of coverage was counted using 7 strategies implemented in CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind. As a result of the depth of coverage counting, we obtained 7 other tables containing the depth of coverage in a given sequencing region for each sample. The resulting tables were used as input to the appropriately modified CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind tools to detect CNVs. As a result of the 7 CNVs callers' operation, we obtained 49 result sets of CNVs—7 CNVs calling tools were run 7 times, each time using a different table with the depth of coverage. CNV indicates copy number variation; WES, whole-exome sequencing.

k-means[21] algorithm as in Kuśmirek et al.[22] For CANOES and ExomeDepth, we used default reference sample set selection algorithms—CANOES with k nearest neighbors algorithm[23] (kNN), where k is equal to 30 (default value), ExomeDepth also with kNN where k value is determined automatically by ExomeDepth software.

Finally, we have evaluated the resultant CNVs set of each pair of (1) depth of coverage table and (2) CNVs calling tool, comparing the resultant CNVs of the pair and the CNVs call set golden record provided by 1000 Genomes Consortium[24] generated based on the whole-genome sequencing (WGS) data. To accurately evaluate the results, the variants were categorized into short (encompassing 1 or 2 sequencing regions) and long (encompassing more than 2 sequencing regions) CNVs, as well as rare (frequency $\leqslant$5%) and common (>5%) events.

Results

To compare different strategies for counting mapped DNA reads in sequencing regions and to present the impact on CNV calling results, we used 1000 Genomes project phase 3 WES data from 861 individuals (444 females and 417 males), including 313 samples from Asia, 276 samples from Africa, 205 samples from Europe, and 67 samples from America. To speed up calculations we limited our research to chromosomes 1 and 11 only. As a result of the quality control process, 2273 out of 20 106 sequencing regions for chromosome 1 and 966 out of 10 565 sequencing regions from chromosome 11 were removed.

Comparison of another depth of coverage tables

First, we compared the depth of coverage tables obtained from the other counting algorithms implemented in the CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind tools. The results of the comparison are presented in Figure 2. The diagram shows that the average depth of coverage counted by the CODEX and ExomeDepth tools is smaller than the average depth of coverage counted by the CANOES and ExomeCopy tools. The main reason for this observation is that the CODEX and ExomeDepth tools count pairs of reads where the first read is mapped to the sequencing region; the CANOES and exomeCopy tools treat each DNA read independently. It follows that if we are dealing with a large sequencing region in which both DNA reads from a pair of paired-end tags are mapped, the CODEX and ExomeDepth applications increase the depth of coverage value by 1, while the CANOES and exomeCopy applications by 2. Moreover, it is worth noting that the diagrams for the CODEX and CNVind applications are identical—both applications count the depth of coverage in the same way. In addition, the graphs

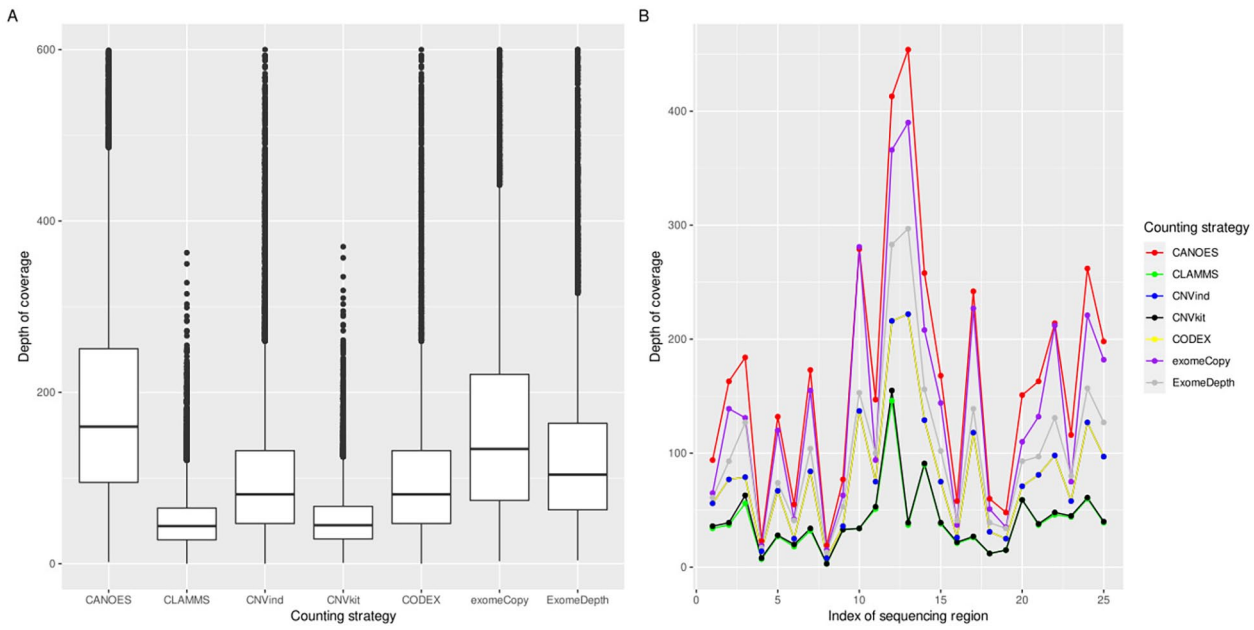

Figure 2. Comparison of the results of counting the depth of coverage with different tools. The figure presents the box plot (A) and course changes in the depth of coverage values (B) for chromosome 11 of the NA06984 sample. First, statistics of the counted depth of coverage values for each tool are different (A). Second, an interesting observation is that the other behavior of different counting depth of coverage strategies in the following sequencing regions. CNV indicates copy number variation.

for the CLAMMS and CNVkit applications are also very similar—they differ only in their outliers.

Moreover, the values of the depth of coverage in the NA06984 sample on the first 25 sequencing regions of chromosome 11 are presented on panel B. In the mentioned diagram, we can see that the 7 compared applications counted the depth of coverage from the same input BAM file very differently. In the diagram, we can see that, as expected, the greatest depths of coverage are counted by the algorithms implemented in the CANOES and exomeCopy applications. However, these values are not stable—there are sequencing regions for which CANOES and exomeCopy return nearly identical depth of coverage, while for other sequencing regions the depth can vary significantly. Moreover, a similar observation can also be seen for the CODEX, ExomeDepth, CLAMMS, CNVkit, and CNVind applications.

Comparison of another resultant CNVs data set

Second, we compared 49 result sets of CNVs detected using the CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind tools. Each of the CNVs callers was run 7 times with the different input depth of coverage table. The results for the chromosome 1 and chromosome 11 data sets are shown in Figures 3 and 4, respectively.

The results showed that for exomeCopy (panel C), the resulting sets of detected CNVs are very similar regardless of the input depth of coverage table selected. On the contrary, for rare CNVs detected by the CODEX tool, the depth of coverage tables from the CODEX, CLAMMS, CNVkit, CNVind, and ExomeDepth tools yield better results than the depth of coverage tables from the CANOES and exomeCopy tools. Especially for rare events, the best results were obtained from the CLAMMS and CNVkit depth of coverage table.

A similar observation can be seen for the ExomeDepth tool—the best rare CNV sets yielded depth of coverage tables from the CLAMMS and CNVkit tools. Mentioned sets of rare CNVs detected were much better than the resultant sets of rare CNVs obtained from the CANOES and exomeCopy depth of coverage tables.

Moreover, different CNV callers detect a different number of CNVs from different ranges. By far the least FP (false positive; the situation where the caller report CNV in the investigated sample, but in reality, there is no CNV in this position of this sample) CNVs are detected by the CANOES tool. Unfortunately, this is also associated with a very low number of TP (true positive; the situation where the caller report CNV in the investigated sample, and in reality, there is the CNV in this position of this sample) calls detected. On the contrary, some tools detect a much larger number of rare TP calls. Unfortunately, as the number of detected TP calls increases, the number of rare FP calls detected also increases.

Discussion

To sum up, in this article, we indicated the problem of the diversity of strategies for counting the depth of coverage in the WES data. There are many applications for detecting CNVs based on depth of coverage. One of the critical steps in these applications is the process of counting the depth of coverage table.

First, the presented research shows that the best methods of counting the depth of coverage are the algorithms implemented in the CLAMMS and CNVkit applications. Both ways allow obtaining much better sets of detected CNVs compared to counting the depth of coverage implemented in the CNVind, CODEX, ExomeDepth, CANOES, and exomeCopy tools. This observation concerns the rare CNVs and the CNV detection algorithms implemented in the CANOES, CLAMMS, CNVkit, CNVind, CODEX, and ExomeDepth tools. Moreover, this observation is reproducible for both independent data sets used in the study—chromosomes 1 and 11.

Second, some CNV detection tools are reasonably resistant to changing the input depth of coverage table. In this study, we proved that the exomeCopy applications give an approximately similar set of the resulting rare CNVs, regardless of the method of counting the depth of coverage table. This observation is quite surprising because, as presented in the study, the 7 investigated input depth of coverage tables are pretty different and represent another depth of coverage values, even for the same sample in the same sequencing region.

One of the main directions of research in the future that opens the presented study is to analyze the process of mapping DNA reads to the reference genome. In the proposed research, the input data was the set of BAM files, ie, the set of binary files containing information about mapping DNA reads to the reference genome. The BAM files set was created with the

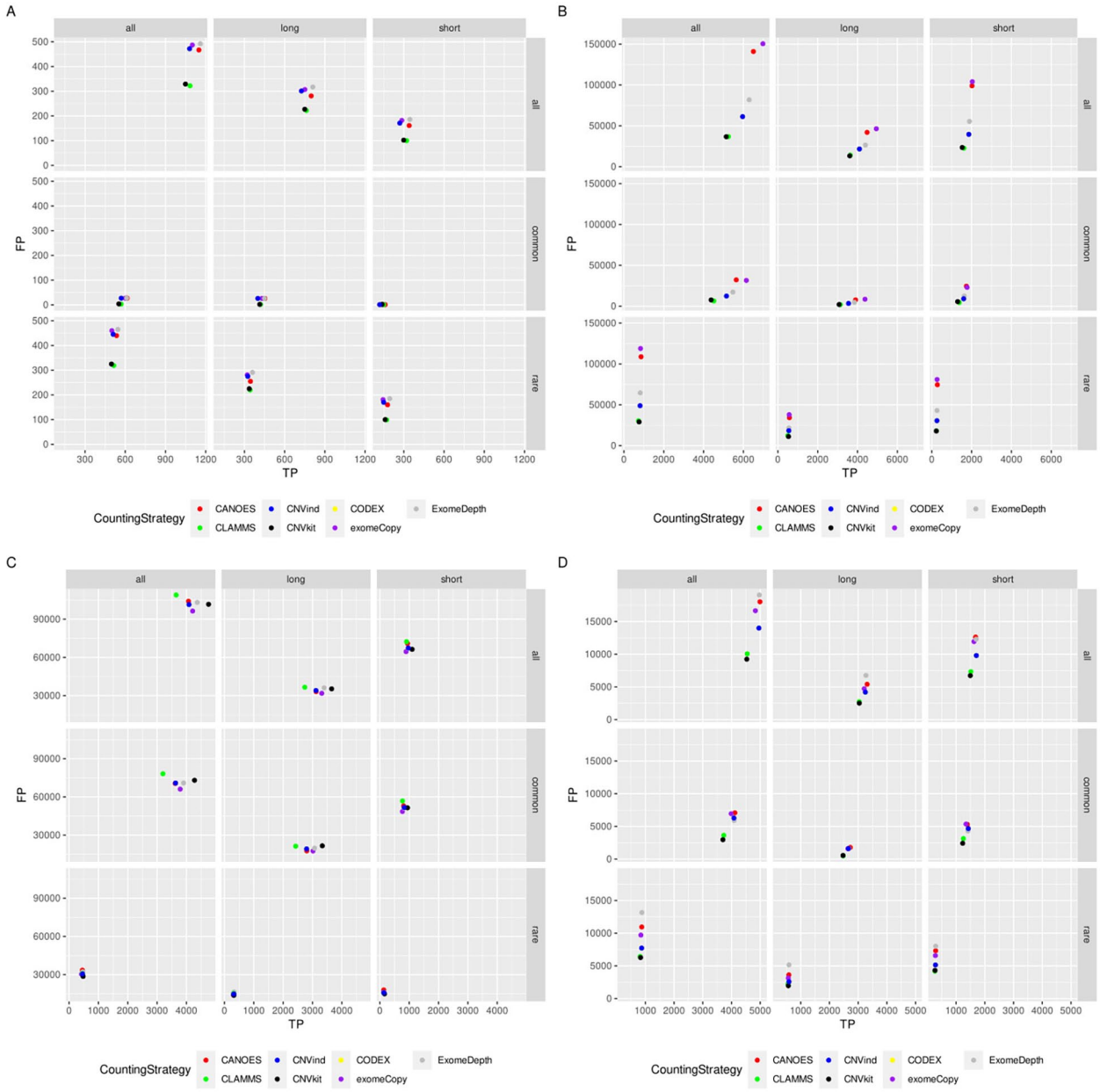

Figure 3. *(Continued)*

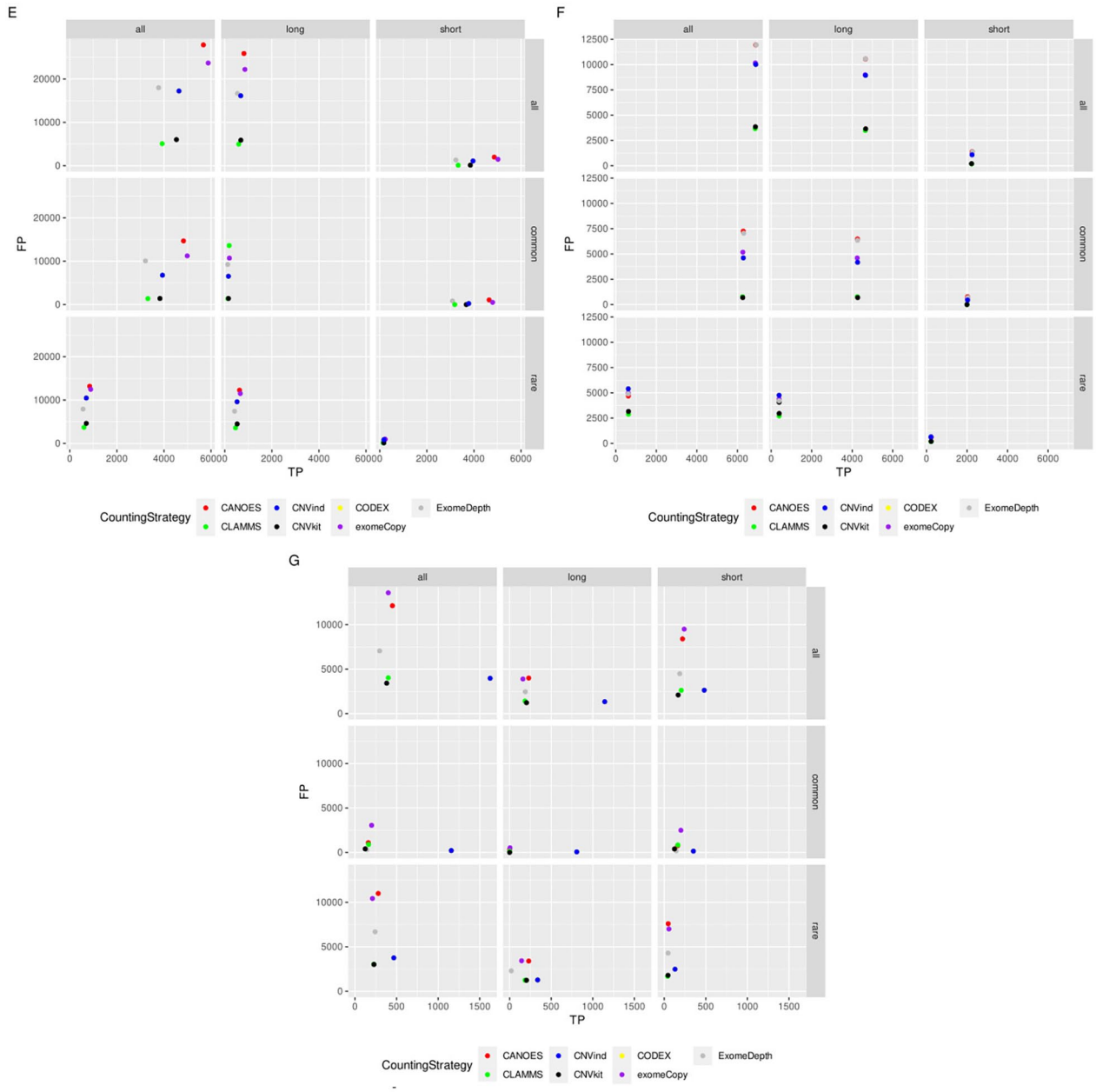

Figure 3. Comparison of the results obtained by the CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind applications for chromosome 1 data set. The diagram shows the evaluation of the resulting CNVs sets for the CANOES (A), CODEX (B), exomeCopy (C), ExomeDepth (D), CLAMMS (E), CNVkit (F), and CNVind (G) tools. Each of the tools was run 7 times with different input depth of coverage tables counted by other methods (the 7 strategies for counting the depth of coverage tested in the experiment were shown in different colors). Each of the result sets of detected CNVs was divided based on frequency (common and rare) and length (short and long). The number of FP calls is presented on the vertical axis, and the number of TP calls on the horizontal axis. It is worth noting that for all applications, the results obtained using the coverage table from the CODEX and CNVind tools are identical—the CODEX and CNVind tools calculate the coverage depth in the same way. CNV indicates copy number variation; FP, false positive; TP, true positive.

BWA[25] and mrsFAST[26] applications.[27] However, other applications and algorithms are also present for mapping DNA reads to the reference genome, such as Bowtie2,[28] HISAT2,[29] and MUMmer4.[30] In the future, we plan to check what impact the selection of different mapping algorithms has on the resulting BAM file and, as a result, on the resulting set of detected CNVs. In addition, other steps of CNV detection, which differ in other CNV callers, should be compared, eg, in the segmentation step, depending on the application, the Hidden Markov Model[31] or the circular binary segmentation[32] algorithm is used.

Moreover, the presented research could be the starting point of developing a new algorithm for counting the depth of coverage table. This algorithm will at its base use the algorithms from the CLAMMS and CNVkit tools, as these algorithms give the best results. One approach might be to

extract the average depth of coverage from both tools—each application computes its depth of coverage table. The average of both tables is taken, which is the input table for the CNV detection process. We should also consider whether the mentioned averaging process should use arithmetic mean, but a weighted mean with completely different weights for different sequencing regions. Another approach that should be checked in the future would be to extend the BAM file in the process of mapping the DNA reads to the reference genome, eg, characterizing the site on the reference genome to which the given DNA read is mapped (eg, segmental duplication). This approach will require the extension and modification of the appropriate DNA reads mapping application. The additional mapping information in the BAM file could be used in the depth of coverage counting process and have a crucial impact on the resulting set of detected CNVs.

Conclusions

The detection of rare CNVs is crucial in the diagnosis of many genetic diseases. Despite the vast role of rare CNVs in the human genome, detection methods based on the depth of coverage still do not obtain satisfactory results, mainly due to many FP calls. In this article, we compared different strategies for counting the depth of coverage in CNV detection applications. The results indicated that the best strategies for counting the depth of coverage are the methods implemented in the CLAMMS and CNVkit tools.

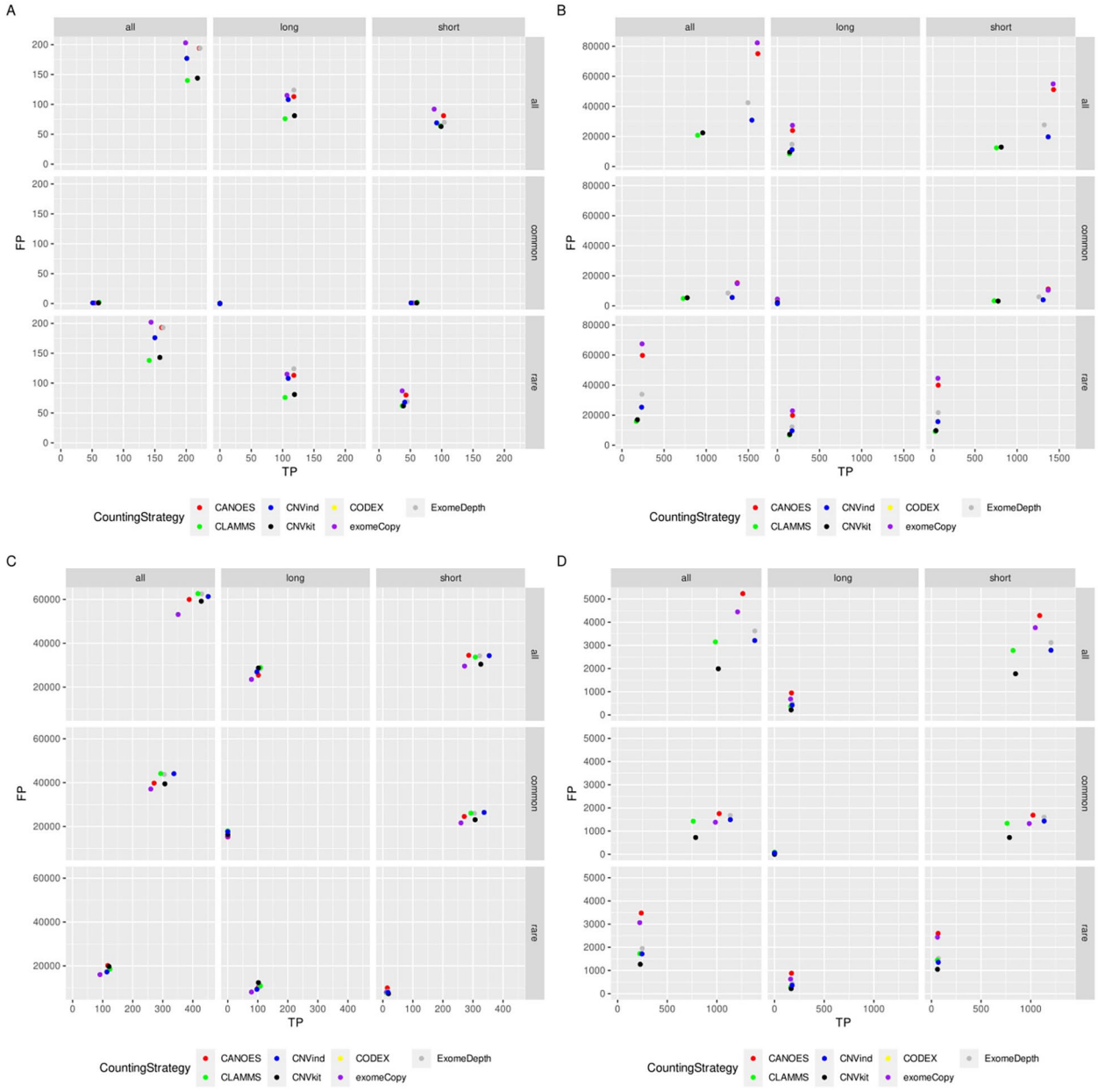

Figure 4. *(Continued)*

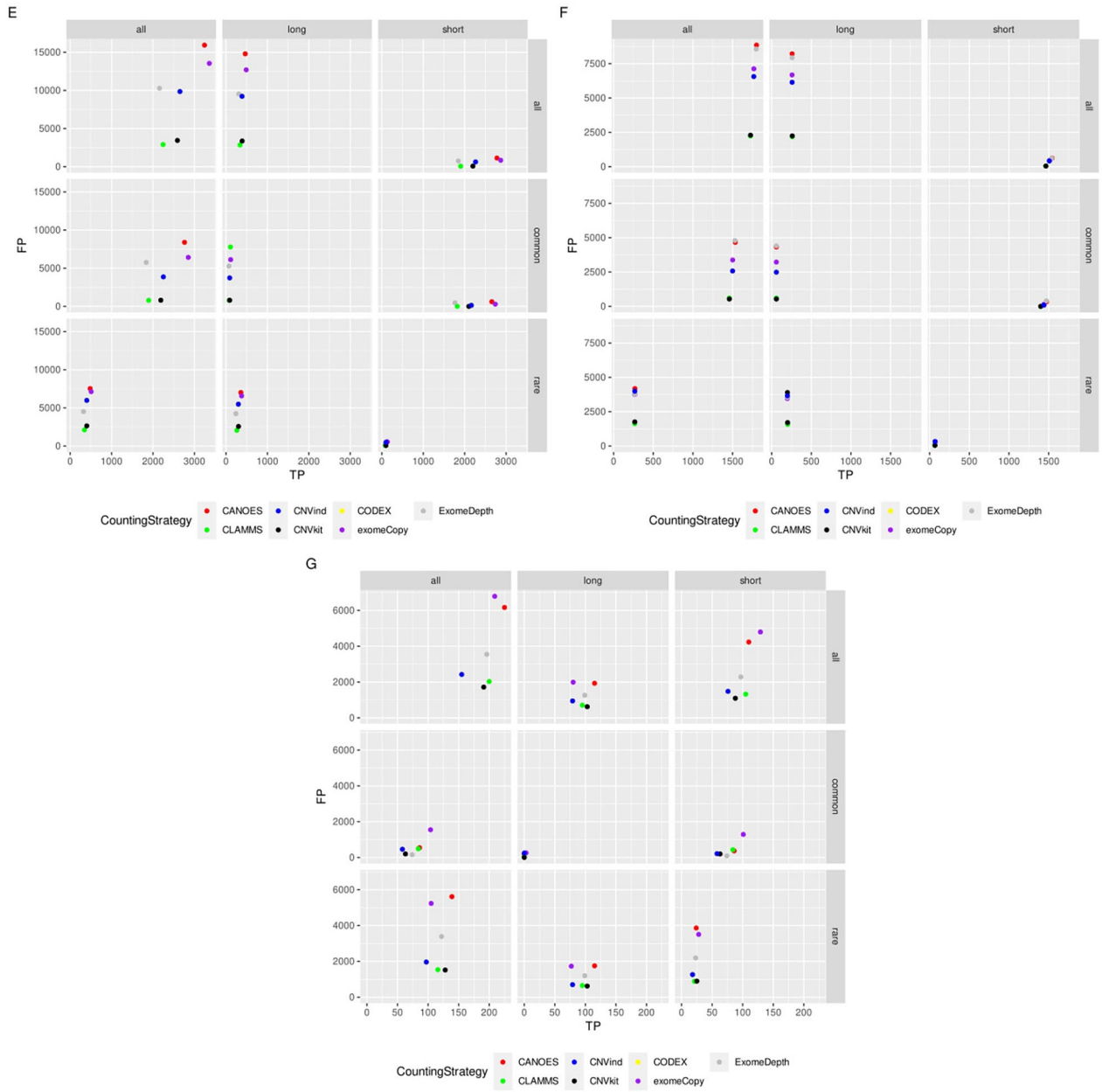

Figure 4. Comparison of the results obtained by the CANOES, CODEX, exomeCopy, ExomeDepth, CLAMMS, CNVkit, and CNVind applications for chromosome 11 data set. The diagram presents a comparison of the different results of CNVs detected by another tool and with different input depth of coverage tables. The following panels show the results for the CANOES (A), CODEX (B), exomeCopy (C), ExomeDepth (D), CLAMMS (E), CNVkit (F), and CNVind (G) tools, and different algorithms for counting the depth of coverage are marked with other colors. CNV indicates copy number variation; FP, false positive; TP, true positive.

Author Contributions

WK contributed to conceptualization, data curation, formal analysis, investigation, software, visualization, and writing. All authors approved the final manuscript.

Data Availability

All scripts that have been used are publicly accessible at GitHub repository https://github.com/wkusmirek/cnv-depth-of-coverage-comparison.

REFERENCES

1. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39:S37-S42.
2. Capalbo A, Rienzi L, Ubaldi FM. Diagnosis and clinical management of duplications and deletions. *Fertil Steril*. 2017;107:12-18.
3. Pös O, Radvanszky J, Buglyó G, et al. DNA copy number variation: characteristics, evolutionary and pathological aspects. *Biomed J*. 2021;44:548-559.
4. Eichler E. Copy number variation and human disease. *Nat Educ*. 2008;1:1.
5. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451-481.
6. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437-455.

7. Li YR, Glessner JT, Coe BP, et al. Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nature Communications*. 2020;11:1-9.
8. 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: 30-37.
9. Tan R, Wang Y, Kleinstein SE, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*. 2014;35: 899-907.
10. Moreno-Cabrera JM, Del Valle J, Castellanos E, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet*. 2020;28:1645-1655.
11. Zhao L, Liu H, Yuan X, Gao K, Duan J. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*. 2020;21:1-10.
12. 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.
13. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Research*. 2015;43:e39.
14. Kuśmirek W, 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*. 2022;23:1-16.
15. Packer JS, Maxwell EK, O'dushlaine C, et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*. 2015;32:133-135.
16. Backenroth D, Homsy J, Murillo LR, et al. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res*. 2014;42:e97.
17. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol*. 2016;12:e1004873.
18. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol*. 2011;10:52.
19. Plagnol V, Curtis J, Epstein MY, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics (Oxford, England)*. 2012;28:2747-2754.
20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841-842.
21. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. Oakland, CA, USA*. Berkeley, CA: University of California Press; 1967:281-297.
22. Kuśmirek W, Szmurlo A, Wiewiórka M, Nowak R, Gambin T. Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. *BMC Bioinformatics*. 2019;20:266.
23. Zhang Z. Introduction to machine learning: K-nearest neighbors. *Ann Transl Med*. 2016;4:218.
24. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
25. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2010;26:589-595.
26. Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res*. 2014;42:W494-W500.
27. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526:75-81.
28. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9:357-359.
29. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12:357-360.
30. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:e1005944.
31. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann Math Stat*. 1966;37:1554-1563.
32. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5: 557-572.