Detection of copy number variation from low-coverage whole-genome sequencing data

Zuzana Klinovská[1], **Marcel Kucharík**[1,2], **Martina Sekelská**[4,5,6], **Michaela Hýblová**[4,5,6], **Jaroslav Budiš**[1,2,3], **Tomáš Szemes**[1,2,7]

[1]Geneton Ltd, Bratislava, Slovakia
[2]Comenius University Science Park, Bratislava, Slovakia
[3]Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
[4]Medirex Inc., Bratislava, Slovakia
[5]TrisomyTest Ltd., Bratislava, Slovakia
[6]**Medirex** Group Academy, Bratislava, Slovakia
[7]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava

Copy number variation (CNV) is a form of structural variant that leads to an abnormal number of copies of genomic regions in a cell. Next-generation-sequencing technologies enabled efficient whole-genome-sequencing, which made the detection of CNVs cheaper and faster. In this article we review four tools for the detection of CNVs on low-coverage data and compare their results and highlight their advantages and disadvantages.

Key words: structural variants, detection of the structural variants, tools for the detection of structural variants, copy number variations, detection of copy number variations.

Detekcia štrukturálnych variantov v genóme z dát s nízkym pokrytím.
Varianty počtu kópií (CNV) sú jednou z foriem štrukturálnych variantov, ktoré spôsobujú abnormálne kopírovanie niektorých regiónov genómu v bunke. Sekvenovanie druhej generácie umožnilo efektívne celogenómové sekvenovanie a tým sa umožnila rýchlejšia a lacnejšia detekcia týchto variantov. V našom článku posudzujeme štyri nástroje pre detekciu CNV variantov na dátach s nízkym pokrytím. Porovnávame ich výsledky a pre každý nástroj sa snažíme ustanoviť jeho výhody a prípadné nevýhody.

Kľúčové slová: štrukturálne varianty, detekcia štrukturálnych variantov, nástroje pre detekciu štrukturálnych variantov, varianty počtu kópií, detekcia variantov počtu kópií.

Newslab, 2021; roč. 12 (2): 53 – 57

Introduction

In recent years the sequencing costs were dropping rapidly and this trend continues each day. The analysis of various structural variants using shallow-depth sequencing data is becoming more and more popular mainly due to a fact that it is cheaper, faster, and yields relatively accurate results. This approach can be used in the detection of a specific type of structural variants called Copy number variations or CNVs.

Fetal CNVs in NIPT and tools for their detection

Copy number variation is a phenomenon in which sections of DNA are duplicated or they are deleted. This phenomenon represents a significant source of genetic diversity among different species including humans.

However, CNVs are linked to various syndromes and diseases as well. They are associated with schizophrenia, autism, or susceptibility to HIV infection(1). What is more, CNVs can cover part of a gene, whole gene, or even several genes and therefore they are likely to have a role in the alternation of human physiological functions, which are essential processes such as metabolism, movements, reproduction, and others(2).

Genetic centers have developed numerous tools for the detection of CNVs. Our goal is to compare four different CNV detection tools. We are particularly interested in fetal CNVs that were acquired via non-invasive prenatal testing (NIPT), which uses shallow whole-genome sequencing. In this study, we analyzed data with very small coverages from 0.05x to 0.5x, which is common in NIPT and similar tests. Chosen tools share some similarities in the detection approach. However, due to various factors that affect this process, their performance is greatly varied. Overall, detection of any microdeletion/microduplication variant is limited by four main factors: fetal fraction in NIPT samples (proportion of cell-free fetal DNA - fetal DNA that circulates freely in the maternal blood), size of the particular CNV, coverage, and biological and technical variability of the event region(2). From these factors only coverage can be directly changed to obtain more accurate results, but with a higher production cost. Naturally, the higher the fetal fraction and the bigger the size of CNV, the easier the detection for any of the tools. Biological and technical variability of the event region refers to the fact that some sectors can be more variable than others. It can be caused by various factors such as repetitive elements, mapping ability, and so on[2]. As a result, these regions are harder to detect and are usually filtered out from the analyses.

CNV detection approach

Tools for CNV detection share some similarities in their approaches. Usually, there are four main steps. Firstly, the reads of the target sequence are separated into smaller sections called bins. Bin-size is the number of bases inside the bin. Although this parameter can be often adjusted by the user, some tools propose a method to determine the optimal bin size. Final detection resolution strongly depends on the bin size. Larger bin size results in worse resolution, but faster computational time in some tools.

The second step is normalization. This process normalizes or purifies the bin counts from various biological biases such as GC content. The process of normalization is very important since it reduces the noise commonly seen in samples and therefore can considerably improve final sensitivity and specificity.

Following normalization comes segmentation of the signal. Segmentation is a process of splitting the signal into parts with equal height (or equal normalized bin-count in our scenario). Circular binary segmentation (CBS) is a fast, recursive algorithm, which finds change points in sequential data, where CNVs could be found. It is a popular method for segmentation.

Finally, the segments are categorized as baseline segments (no CNV detected) or segments with duplication/deletion. Furthermore, in the NIPT case, here we can usually distinguish segments with fetal and maternal CNV based on the strength of the signal compared to the fetal fraction.

CNV detection tools

Comparison of CNV detection tools helps highlight the advantages and disadvantages of the particular tools. Although there are similarities within them, the results vary, especially when the detection is done on a special type of samples, such as NIPT samples with low coverage. Overall, the comparison was done over four different tools.

The first tool we wanted to include in our comparison is WisecondorX[2,3]. In contrast to its predecessor WISECONDOR[2-4], the newer tool is faster and has a wider usage. The authors used the CBS algorithm in the segmentation step and thus lowered computation time. In addition to this, they made WisecondorX usable for analyses beyond NIPT, since the original tool could not be used on other samples.

The next tool is CNV-caller, which similarly to WisecondorX uses the CBS algorithm. However, since this procedure partitions a chromosome excessively, which can create excessive noise, authors of these tools apply a rule to determine the significance of a segment. This could lead to more accurate results.

The following tool is called iCopyDav[5], the default bin size for this tool for low-coverage data is 500bp which is more suitable for data with higher coverage. Previous tools had default sizes in tens of kilobases. In our analyses, we use the same bin size for every tool (20 000bp) and as a result, we expect less accurate results. However, this tool uses some interesting approaches and ideas that we wanted to put to a test. In the segmentation step, CBS is used, but in contrast to the previous two tools, iCopyDav uses a different method for smaller aberrations. This is because a CBS is predominantly used for identifying larger CNVs, consequently, iCopyDav should have a wider range of CNV predictions.

The last tool we compared is called CNVkit[5,6]. In this case, CNVkit is said to calculate bin size specifically for off-target (non-coding) regions, which is an interesting approach and we wanted to see how this methodology would do in the comparison. In addition, this tool accepts an input with no control samples, thus it can work without a reference from normal samples, while other tools require this data. However, a reference genome such as hg19 should be provided. For the segmentation step, CBS is used as in the other tools.

Materials and methods

One of our goals was to discover the relation between detection accuracy and parameters such as fetal fraction and CNV size. Three sets of samples were used for this analysis: training samples, mixed data samples, and healthy NIPT samples. All obtained samples are from patients, who signed consent with the study.

Training samples for all tools, except for the CNVkit, use a constructed reference for CNV detection. For the reference creation process, 134 samples were used. Training data were collected from standard NIPT samples, originating from confirmed genetically healthy singleton pregnancies.

To analyze the effect of fetal fraction and CNV length, samples with different values of these factors were needed. We decided to use samples mixed in the laboratory from samples with confirmed selected microdeletion syndromes. We mixed these samples with healthy samples in different ratios imitating different fetal fractions. The selected microdeletion syndromes were: DGS-DiGeorge syndrome (chr 22), AS-Angelman syndrome (chr 15), PW-Prader Willy syndrome (chr15), CDC-Cri du chat syndrome (chr5), WHS-Wolf Hirchhorn syndrome (chr4), 1p36-1p36 (chr1). A total number of obtained mixed samples is 19.

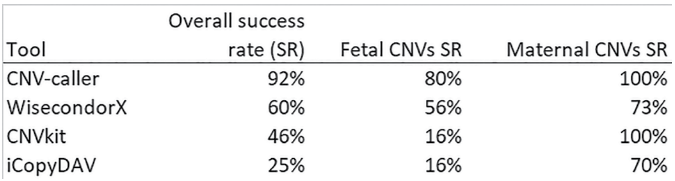
We gathered 35 NIPT samples from the production with confirmed 41 CNVs. Maternal CNVs were processed as well (11 CNVs). The average percentage of fetal fraction for these samples is 14%. For each CNV, the approximate start/end position on the chromosome was stated. All samples were sequenced by Illumina NextSeq. Subsequently, reads were aligned by Bowtie2[7] to a reference genome hg19.

Results

Comparison of the tools according to their success rate

The comparison of mentioned tools showed how tools react to various samples with different fetal fractions and CNV sizes. As expected, samples that contained little fetal fraction and smaller CNVs were harder to detect, yet some tools appear to be successful even with those cases. What we found interesting is that some tools seemed to detect maternal CNV with no problems while CNVs from NIPT were not detected. In the end, some tools are designed for NIPT samples, therefore we calculated the success rate individually for fetal and maternal aberrations. Gathered results for success rate are shown in **Table 1**.

Table1. Shows overall success rate for detection of CNVs.



Tool	Overall success rate (SR)	Fetal CNVs SR	Maternal CNVs SR
CNV-caller	92%	80%	100%
WisecondorX	60%	56%	73%
CNVkit	46%	16%	100%
iCopyDAV	25%	16%	70%

Naturally, we tested precision as well, meaning we observed whether the tool can rightly state the position of the aberration. This is because the tool might detect a CNV somewhere on the genome, but in order to correctly set a diagnosis, the tool should be as precise as possible.

Individual results for the tools

For mixed samples, we summarize the results of all four tools in **Table 2**.

Overall the best results yielded the CNV-caller tool, where the success rate reached approximately 92% for both mixed and normal NIPT samples as shown in **Table 1**. Furthermore, this tool was the most precise tool for normal NIPT samples as can be seen in **Table 1**. In the output, CNV-caller produces output, where CNVs are color-coded by the confidence or severity of the detection. Since this tool performed well with NIPT samples, we show in **Figure 1**. the importance of the fetal fraction and the size of the CNV for prediction accuracy.

Following CNV-caller comes WisecondorX with an overall success rate of 60% as is shown in **Table 1**. The results for mixed samples are also displayed in **Figure 2**.

For both tools, samples with less than 10% of fetal fraction remained challenging (as is shown in **Figure 1**. for CNV-caller and in **Figure 2**. for WisecodnorX) and proved the importance of this factor. Furthermore on **Figure 2**. we can see that even an aberration, with the size above average (17.7 Mb), was not detected due to low fetal fraction, which was less than 6%.

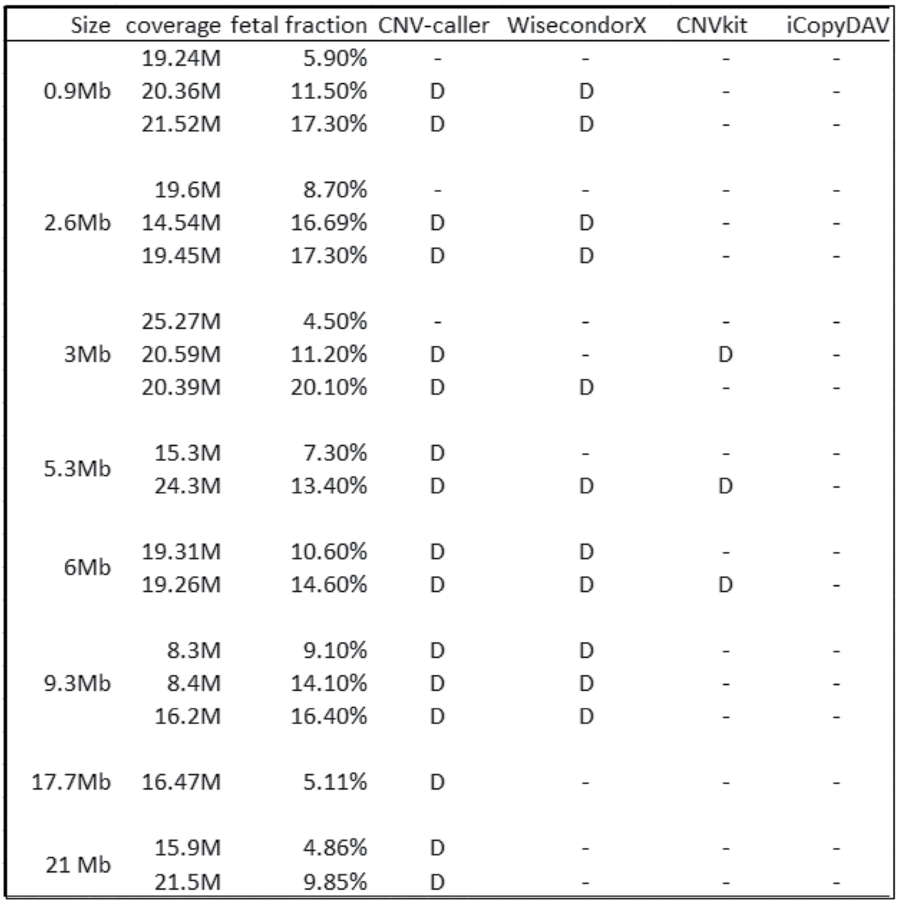
When it comes to practical use, both tools generate a.png file from which aberrations can be observed and text output with precise coordinates for machine processing.

Both remaining tools have an overall success rate below 50%, which is mainly due to low accuracy for fetal CNVs (both tools were not specifically created for the NIPT scenario). Coverage had a great impact on the detection as well. In the case of iCopyDav, we suspect that low coverage was the reason why no CNV was reported from this tool for mixed samples as shown in **Table 2**. Manuals for iCopyDav suggest using at least 1x coverage, while the average coverage of our data is significantly less (from 0.05x to 0.5x).

Discussion

copy number variations may lead to various diseases and by detecting these aberrations in the early state from NIPT samples we might be able to set a correct diagnosis

Table 2. Shows the size of the aberrations, coverage, and the percentage of the fetal fraction for mixed samples. Letter D means the CNV was detected by the particular tool, whereas symbol '-' represents CNVs that were not detected.



Size	coverage	fetal fraction	CNV-caller	WisecondorX	CNVkit	iCopyDAV
0.9Mb	19.24M	5.90%	-	-	-	-
	20.36M	11.50%	D	D	-	-
	21.52M	17.30%	D	D	-	-
2.6Mb	19.6M	8.70%	-	-	-	-
	14.54M	16.69%	D	D	-	-
	19.45M	17.30%	D	D	-	-
3Mb	25.27M	4.50%	-	-	-	-
	20.59M	11.20%	D	-	D	-
	20.39M	20.10%	D	D	-	-
5.3Mb	15.3M	7.30%	D	-	-	-
	24.3M	13.40%	D	D	D	-
6Mb	19.31M	10.60%	D	D	-	-
	19.26M	14.60%	D	D	D	-
9.3Mb	8.3M	9.10%	D	D	-	-
	8.4M	14.10%	D	D	-	-
	16.2M	16.40%	D	D	-	-
17.7Mb	16.47M	5.11%	D	-	-	-
21 Mb	15.9M	4.86%	D	-	-	-
	21.5M	9.85%	D	-	-	-

Figure 1. Describes success rate of CNV-caller for mixed samples. Circles displaying detected CNVs and crosses display CNVs that were not detected. The relation between fetal fraction and size of the CNV with the correct detection is shown through gradient with darker areas displaying regions that are harder to detect.

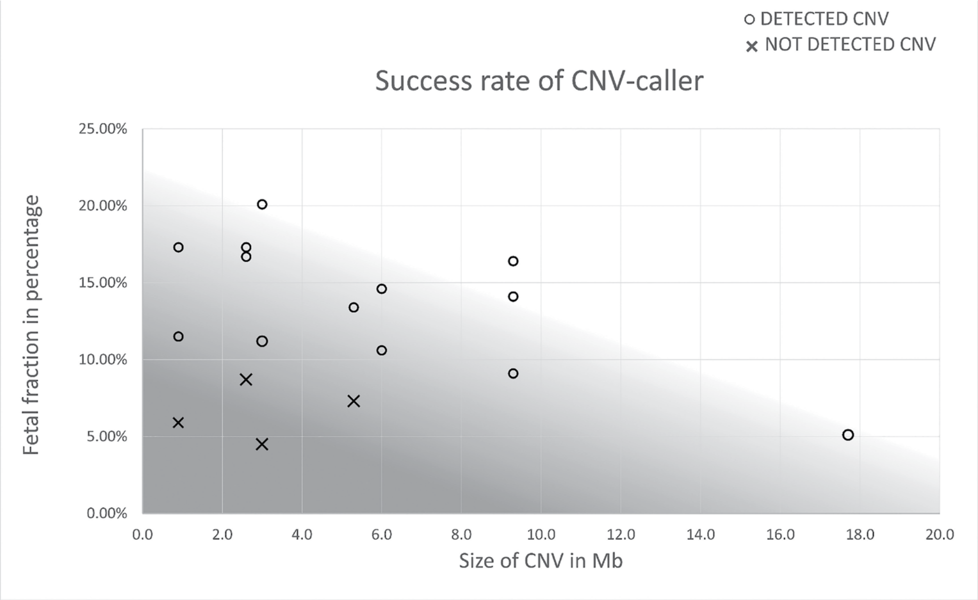
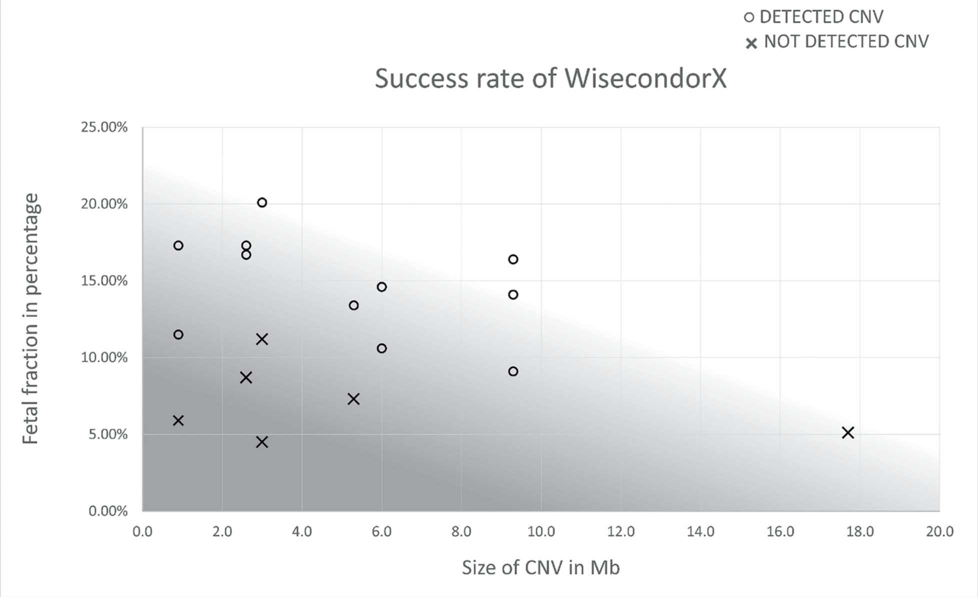


Figure 2. Describes the success rate of WisecondorX for mixed samples. Circles displaying detected CNVs and crosses display CNVs that were not detected. The relation between fetal fraction and size of the CNV with the correct detection is shown through gradient with darker areas displaying regions that are harder to detect.



in time and assure a better treatment. In combination with a low-coverage method, we can predict not only the health risks but calculate the effectiveness (in both production time and costs). By comparing different tools we get a better perspective on limitations that are challenging for the tools and we get a better understanding of what causes inaccuracy or inability to detect a given CNV.

There is no doubt that the best performing tool was the CNV-caller. With a 92% success rate and 100% success rate for maternal CNVs, this tool can be used for the detection of CNVs in both NIPT and normal samples. In addition, it generates a.png file for clarity and with each found CNV there is information about the confidence of the said detection. This particular tool also detected some CNVs in samples with smaller fetal fractions even lower than 5%, but the CNVs were a bit longer to compensate for the lower fetal fraction.

To conclude, by comparing CNV detection tools and highlighting which samples are the most suitable for the particular tool we could design a decision tree that would pick the right tool to use for the CNV based on its fetal fraction level, size of the CNV, and coverage. In our study CNV-caller alongside WisecondorX proved to be the most reliable and accurate tools for CNV detection.

Acknowledgment

This article was created with the support of the OP Integrated Infrastructure for the project: Center for biomedical research – BIOMEDIRES – II. phase, ITMS: 313011W428, and for the project: Introduction of an innovative test for screening and monitoring of cancer patients – GenoScan LBquant, ITMS 313012Q927, both co-financed by the European Regional Development Fund. This work was supported by the Slovak Research and Development Agency under the Contract no. PP--COVID-20-0051.

Poďakovanie

Táto publikácia vznikla vďaka podpore v rámci Operačného programu Integrovaná infraštruktúra pre projekt Centrum pre biomedicínsky výskum – BIOMEDIRES – II. etapa , kód ITMS: 313011W428 a pre projekt Zavedenie inovatívneho testu pre screening a monitoring onkologických pacientov – GenoScan LBquant, ITMS: 313012Q927, ktoré sú spolufinancované zo zdrojov Európskeho fondu regionálneho rozvoja. Zároveň bola táto práca podporená Agentúrou na podporu výskumu a vývoja na základe Zmluvy č. PP-COVID-20-0051.

REFERENCES

1. Liu S, Yao L, Ding D, Zhu H. CCL3L1 copy number variation and susceptibility to HIV-1 infection: a meta-analysis. PLoS One. 2010;5: e15778.

2. Zhang F, Gu W, Hurles ME, Lupski JR. Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics. 2009. pp. 451–481. doi:10.1146/annurev.genom.9.081307.164217

3. Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. Wisecondor X: Improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res. 2019;47: 1605–1614.

4. Straver R, Sistermans EA, Reinders MJT. Introducing WISECONDOR for noninvasive prenatal diagnostics. Expert Rev Mol Diagn. 2014;14: 513–515.

5. Dharanipragada P, Vogeti S, Parekh N. iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. PLoS One. 2018;13: e0195334.

6. 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.

7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012. pp. 357–359. doi:10.1038/nmeth.1923

Zuzana Klinovská
Geneton s.r.o.
Ilkovičova 8, 841 04 Bratislava
e-mail: zuzana.klinovska@geneton.sk