diagnostics

MDPI

Article

Copy Number Variant Detection with Low-Coverage Whole-Genome Sequencing Represents a Viable Alternative to the Conventional Array-CGH

Marcel Kucharík [1,2,*], Jaroslav Budiš [1,2,3], Michaela Hýblová [4], Gabriel Minárik [4] and Tomáš Szemes [1,2,5]

[1] Geneton s.r.o., 841 04 Bratislava, Slovakia; jaroslav.budis@geneton.sk (J.B.); tomas.szemes@geneton.sk (T.S.)
[2] Comenius University Science Park, 841 04 Bratislava, Slovakia
[3] Slovak Centre of Scientific and Technical Information, 811 04 Bratislava, Slovakia
[4] Trisomy Test s.r.o., 841 04 Bratislava, Slovakia; michaela.hyblova@medirex.sk (M.H.); gabriel.minarik@medirex.sk (G.M.)
[5] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, 842 15 Bratislava, Slovakia
* Correspondence: marcel.kucharik@geneton.sk; Tel.: +421-904-855-365

Abstract: Copy number variations (CNVs) represent a type of structural variant involving alterations in the number of copies of specific regions of DNA that can either be deleted or duplicated. CNVs contribute substantially to normal population variability, however, abnormal CNVs cause numerous genetic disorders. At present, several methods for CNV detection are applied, ranging from the conventional cytogenetic analysis, through microarray-based methods (aCGH), to next-generation sequencing (NGS). In this paper, we present GenomeScreen, an NGS-based CNV detection method for low-coverage, whole-genome sequencing. We determined the theoretical limits of its accuracy and obtained confirmation in an extensive in silico study and in real patient samples with known genotypes. In theory, at least 6 M uniquely mapped reads are required to detect a CNV with the length of 100 kilobases (kb) or more with high confidence (Z-score > 7). In practice, the in silico analysis required at least 8 M to obtain >99% accuracy (for 100 kb deviations). We compared GenomeScreen with one of the currently used aCGH methods in diagnostic laboratories, which has mean resolution of 200 kb. GenomeScreen and aCGH both detected 59 deviations, while GenomeScreen furthermore detected 134 other (usually) smaller variations. When compared to aCGH, overall performance of the proposed GenemoScreen tool is comparable or superior in terms of accuracy, turn-around time, and cost-effectiveness, thus providing reasonable benefits, particularly in a prenatal diagnosis setting.

Keywords: CNV detection; low-coverage WGS; CNV detection comparison; aCGH replacement

check for updates

Citation: Kucharík, M.; Budiš, J.; Hýblová, M.; Minárik, G.; Szemes, T. Copy Number Variant Detection with Low-Coverage Whole-Genome Sequencing Represents a Viable Alternative to the Conventional Array-CGH. *Diagnostics* **2021**, *11*, 708. https://doi.org/10.3390/diagnostics11040708

Academic Editor: Lucía Inglada-Pérez

Received: 20 March 2021
Accepted: 13 April 2021
Published: 15 April 2021

1. Introduction

Copy number variations (CNVs) represent a phenomenon in which sections of the genome are repeated while the number of repeats in the genome varies between individuals. CNVs contribute substantially to normal population variability. However, abnormal CNVs are known to cause numerous genetic disorders. Several methods for CNV analysis are used, from the conventional cytogenetic analysis, through microarray-based approaches, to next-generation sequencing (NGS) [1].

Array-based comparative genomic hybridization (aCGH) delivers genome-wide coverage at a great resolution, even on the scale of dozens of kilobases (10–25 kb) [2]. This fact resulted in aCGH having been the gold standard in CNVs detection for several years. Even though current microarrays offer flexibility in coverage across variable resolution formats, there are still some disadvantages to be considered. For example, in prenatal diagnosis from amniotic fluid, micrograms of genomic DNA are typically needed to hybridize to an array. This can be accomplished either by time-consuming culturing taking up to two weeks, or by whole-genome amplification, which can introduce bias into the analysis. On the contrary, NGS utilizes mere nanograms of DNA, thus not requiring additional amplification. There is lower likelihood of sample contamination due to less material required. The transition from the proven microarray platform to NGS often reveals some new and unexpected data; however, it seems to be a very slow event though the cost and time aspect is already quite unprecedented. Additionally, while aCGH equipment serves a single purpose only, commonly used NGS platforms are very versatile, enabling numerous applications, including exome, genome, targeted panels, transcriptome, or episome sequencing. The whole-exome and targeted sequencing aims to reduce the sequencing cost but is limited to certain regions (protein-coding or custom), where most known disease-causing mutations occur [3]. NGS provides a sensitive and accurate approach for the detection of the major types of genomic variations, including CNVs [4,5].

A handful of CNV detection tools have been introduced in recent years, specifically for targeted and exome sequencing [6–12]. However, these tools are not suitable for data from whole-genome, low-coverage sequencing. The notable whole-genome CNV detection tools include Wisecondor X [13] (successor of Wisecondor [14] tool), CNVkit [15], CNVnator [16], or iCopyDav [17]. Partial comparison of some of these tools is provided in the publication of Wisecondor X [13].

In this paper, we present GenomeScreen—a low-coverage, whole-genome NGS-based CNV detection method and estimate its accuracy in theoretical and in silico settings. This method is partially based on the previously published non-invasive prenatal testing (NIPT) CNV detection method [18,19]. The main differences are the parameters of the reported CNVs—in the NIPT setting, the CNVs corresponding to more than 5% fetal fraction and at least 3 Mb in size were reported. Here, on the other hand, we focus on full (non-mosaic) aberrations with much shorter length (100 kb and larger). Furthermore, we compare the sensitivity of GenomeScreen to the more conventional aCGH method on 106 laboratory-prepared clinical samples. The comparison of GenomeScreen and different CNV detection tools goes beyond the scope of this article due to focus on the comparison with the aCGH method itself.

2. Materials and Methods

2.1. Sample Collection and Processing

All patient samples were analyzed as a part of commercially available testing in cooperation with gynecologists, clinical geneticists, and genetic centers. All patients signed informed consent regarding participation in the research project. Samples of chorionic villi, amniotic fluid, placenta, tissue, or peripheral blood were obtained from 106 patients in the clinical sample group and 789 patients in the training group. Peripheral blood was sampled in K2E (EDTA) vacuum tubes (BD Vacutainer, Plymouth, UK) or Cell-Free DNA BCT (STRECK) vacuum tubes (Streck, La Vista, NE, USA), inverted several times after collection, stored in chilled environment (4–10 °C) for EDTA and at room temperature for STRECK tubes, and transported to the laboratory within 36 h. DNA was extracted from 200 µL of whole blood or 700 µL of amniotic fluid using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol and stored at −20 °C until further analysis.

Genomic DNA from clinical samples was fragmented using 1 U/µL dsDNA Shearase™ Plus (Zymo Research, Irvine, CA, USA) and incubated for 23 min at 42 °C to generate 100–500 bp fragments. For adapter-ligated DNA library construction, the TruSeq Nano kit (Illumina, San Diego, CA, USA) with an in-house optimized protocol was used. Low-coverage sequencing (0.3×) was performed on the Illumina NextSeq 500/550 platform (Illumina, San Diego, CA, USA) with paired-end setting 2 × 35 using High-Output Sequencing Kit v2.5. Library quantity and quality were measured by fluorometric assay on Qubit 2.0 (dsDNA HS Assay Kit, Life Technologies, Eugene, OR, USA). Fragment analysis was performed on the 2100 Bioanalyzer (High Sensitivity DNA Kit, Agilent Technologies, Waldbronn, Germany). We targeted 5 M uniquely mapped reads per sample, while none of the analyses were excluded due to lower (or higher) read counts (more details in Supplementary material Table S1).

2.2. Theoretical Minimal Read Count Estimation

Let us suppose that we model sequencing as a random choice of reads from the whole (mappable) genome. Then, we can theoretically deduce the number of necessary uniquely mapped reads for a certain accuracy criterion. The random choice for a target region is described by the binomial distribution with the mean $\mu = np$ and the variance $\sigma^2 = np(1-p)$. Here, p is the probability of choosing a read from the target region, and n is the number of reads sequenced. The probability p can furthermore be expressed as the ratio of the region length l_c to the whole-genome length l_g ($p = l_c/l_g$). When predicting a CNV, we need to have a certain confidence traditionally determined by the Z-score (Z), defined as follows:

$$Z = \frac{\delta - \mu}{\sigma} \tag{1}$$

Here, δ represents the number of reads that we observe in the target region. We assume that the number of reads in the target region will be proportional to the number of present copies of gonosomes, i.e., either $\delta = n(p + p/2)$ for duplication or $\delta = n(p - p/2)$ for deletion of the region on a single chromosome. If we solve the equation for Z^2 and substitute

$$Z^2 = \frac{(\delta-\mu)^2}{\sigma^2} = \frac{(n(p+p/2)-np)^2}{np(1-p)} = \frac{n^2p^2}{4np(1-p)} = \frac{np}{4(1-p)} = \frac{nl_c}{4(l_g-l_c)} \tag{2}$$

then we can estimate the minimal number of reads (n) to be able to predict a variation with length l_c with the desired Z-score (Z):

$$n \geq \frac{4Z^2(l_g-l_c)}{l_c} \tag{3}$$

2.3. Variant Identification

To identify variations, we performed the following pipeline:

1. Mapping and binning
   a. Mapping reads using Bowtie 2 [20];
   b. Binning reads into same-size 20 kb bins;
   c. Normalizing bin counts.
2. Normalization (similar to the one published previously by [21])
   a. GC bias correction by LOESS smoothing method [22];
   b. Principal component analysis (PCA) normalization to remove higher-order population artifacts on autosomal chromosomes;
   c. Subtracting per-bin mean bin count to obtain data normalized around zero.
3. Filtration of unusable bins
   a. Unmappable or poorly mappable regions (zero or low mean of bin count);
   b. Repetitive regions or areas with certain systematically increased mappability (high mean of bin count);
   c. Highly variable regions (high variance of bin count).
4. Segment identification and reporting
   a. Circular binary segmentation algorithm [23] to identify consistent segments of similar coverage;
   b. Assigning significance to segments based on the proportion of reads;
   c. Visualization of findings (Figure 1).

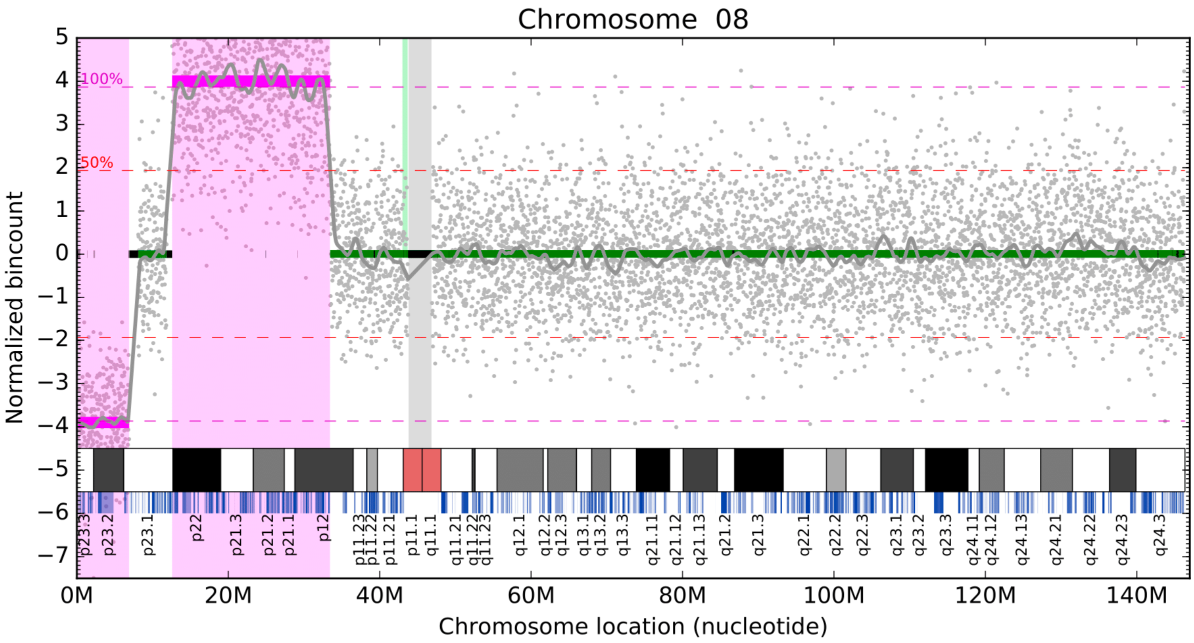

Figure 1. Visualization of the detected deviations on chromosome 8. Chromosome location is on the X-axis. Normalized bin count is on the Y-axis. Green lines represent normal bin count segments (normalized around zero), magenta lines visualize aberrations (one deletion at the start of the chromosome, one duplication on p22–p12). Filtered bins are depicted as black bars on the zero line on the Y-axis. The unmapped region around the centromere is visualized with the grey bar. Grey dots represent the normalized individual bin counts for each bin.

Scripts (Python 3.7) and data are available on the website https://github.com/marcelTBI/GenomeScreen (accessed on 14 April 2021).

2.3.1. Mapping and Binning

Firstly, the reads were mapped to a reference using Bowtie 2 [20] with –very-sensitive settings. We used the hg19/GRCh37 reference in all applications, but other references can be used without changes to the algorithm. The reads were then filtered for map quality of at least 40 and binned according to their starts to same-size 20 kb bins. All subsequent analyses were performed on the bin counts, while the algorithm did not use any other information about reads (for example, sequence). For training purposes, the bin counts corresponding to autosomal chromosomes for each sample were normalized to the identical number of reads (i.e., each bin was divided so the sum of all bins on autosomal chromosomes would be the same for each sample). Furthermore, the same was performed separately for chromosome X and chromosome Y. As a consequence of the separate normalization of sex chromosomes, the applied approach can only detect small sex chromosomal variations and not the whole sex chromosomal aneuploidies.

2.3.2. Normalization

Normalization consisted of three steps: firstly, a sample-wise LOESS-based GC correction was deployed on the bin counts [22]. Next, the principal component analysis (PCA) normalization was used to remove higher-order population artifacts on autosomal chromosomes [21]. For training of the PCA, LOESS-corrected bin counts of 789 NIPT samples with female fetuses were converted to principal component space and the first 15 principal components were stored. The bin count vector of a new sample was then transformed into principal component space defined by these first 15 components and transformed back to the bin space to obtain residuals that were then removed from the bin counts. The first principal components represent the noise commonly observed in euploid samples, and their removal facilitates data normalization. In the present case, the PCA normalization was performed only on autosomal chromosomes due to unavailability of a sufficient number of male samples for training. In the future, the training of PCA on both male and female samples is likely to increase the precision of prediction for sex chromosomes. Lastly, we subtracted per-bin mean bin counts to obtain data normalized around zero. This last step was trained already on the PCA normalized bin counts (where available) and helped compensate for the mapping inequality between various genomic regions.

2.3.3. Filtration of Unusable Bins

To further improve accuracy, we filtered bins that had an unusual signature—low mean (this signaled poor mappability of the region), high mean (repetitive regions or regions with a certain systematic bias), or high variance (highly variable regions). Furthermore, the filtered regions were manually curated to reduce their scatter, mainly around centromeres and in sex chromosomes. The filtration screened out around 15% of the genome, mainly due to the low mappability, especially in and around centromeres.

2.3.4. Segment Identification and Reporting

After normalization and filtering, we received a signal (grey dots in Figure 1) that required segmentation into identical -level parts to be evaluated. To this end, we used the circular binary segmentation (CBS) algorithm implemented in the R package DNAcopy [23]. After segmentation, each segment was assigned a significance level based on its length and difference from zero. Since we knew the mean bin counts, we could estimate the level for a complete deletion or duplication per single copy of a chromosome (magenta dashed lines in Figure 1). We then differed between five color-coded levels of significance: magenta—minimum 75%, minimum 200 kb, red—minimum 25%, minimum 200 kb, orange—minimum 25%, minimum 40 kb, yellow—minimum 12.5%, minimum 40 kb, and green—all others (very short segments or segments around zero). The findings were then reported as a text file for further machine processing, while each chromosome was visualized (Figure 1).

2.4. In-Silico Analysis

For the in silico analysis, we chose 83 samples without any aberration and with a read count of at least 10 M. Firstly, the samples were down-sampled to the studied read count (3–10 M with the step of 1 M). Then, for each of the tested variation lengths (20–200 kb with the step of 20 kb), 100 random variations on autosomal chromosomes were generated that did not overlap with the filtered regions (see Section 2.3.3). To create a sample with an artificial aberration, the bins corresponding to the generated random variation were multiplied accordingly (thus, the most time-consuming mapping step was performed only once per sample). Next, variant identification was performed without changes.

In total, we gradually created 664,000 artificial samples (100 variations × 83 samples × 10 variation lengths × 8 read counts) and performed variant identification on them to analyze the impact of read count and variant length. Every detection that overlapped the simulated region (the exact match of the coordinates was not required) was reported as successful.

3. Results

3.1. Theoretical Minimal Read Count

The theoretical minimum of reads for predicting a variation with length l_c with the desired Z-score (Z) is estimated as (see Section 2.2)

$$n \geq \frac{4Z^2\left(l_g - l_c\right)}{l_c} \tag{4}$$

As a standard, the Z-score of 4 is used in the detection of whole chromosomal aneuploidies [24,25]; however, there are inherently more possible CNVs than whole chromosomal aneuploidies. Thus, the desired Z-score should be much higher in this instance to reduce the number of false positives. Moreover, in practice, the number of necessary reads would be even higher due to the uncertainty of sequencing and mapping, and inherent biological biases [26,27]. The theoretical minimal read count estimation for different Z-scores is displayed in Figure 2.

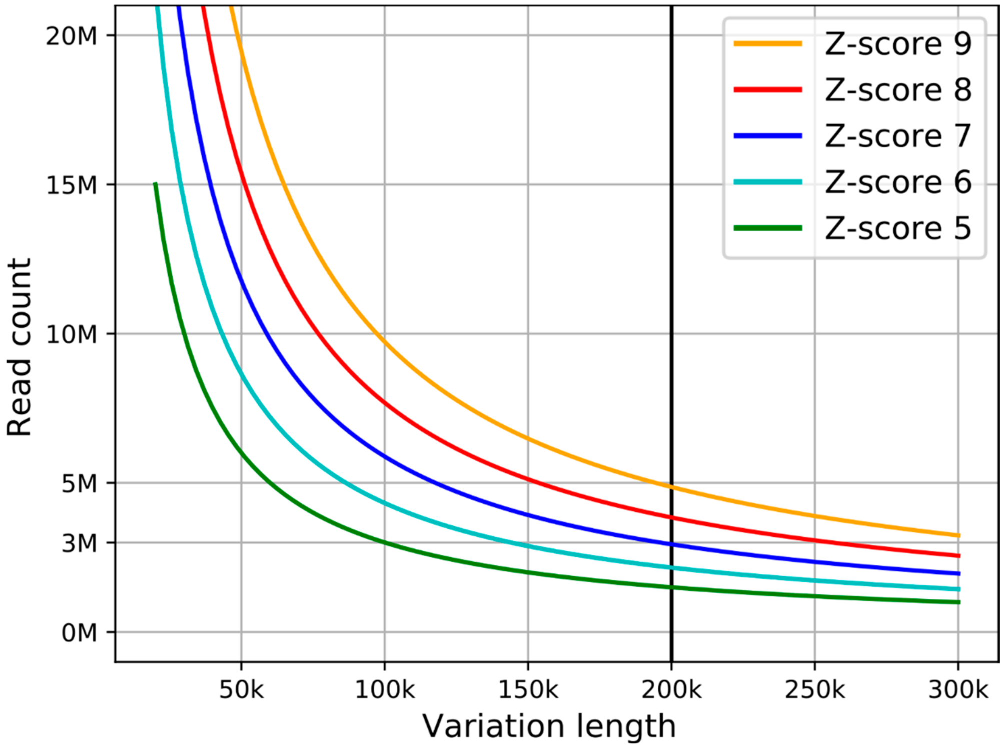

Figure 2. Theoretical minimal read count for successful estimation of copy number variation (CNV) with specified variation length. Different lines represent different Z-score confidence levels.

3.2. Detection Accuracy for Variable CNV Lengths and Read Count (In Silico)

To verify the theoretically estimated limitations, we first conducted a simulated in silico experiment. Artificial samples with simulated CNV were created from healthy samples by multiplication of bins corresponding to the simulated regions randomly selected on the genome. Only the regions that did not span into filtered positions were kept for further analysis (about 85% of the genome). The details can be found in Section 2.3.

The in silico analysis shows the influence of read count and CNV length on prediction accuracy (Figure 3). Based on the findings, we recommend using read counts of at least 8 M to achieve >99% prediction accuracy for variations with 100 kb and more. We therefore recommend following the line for the Z-score of 8 (red on Figure 2) to get an estimation for different CNV lengths.

Comparison of simulated and reported regions showed that the method can predict the exact simulated region coordinates in 88.2% or coordinates with one-bin difference in 97.7% of cases for 200 kb variation length and 10 M reads. These numbers slightly drop to 75.3% and 91.7% for 5 M reads. The imperfection in predicting coordinates is caused by low coverage and lower mappability of some genomic regions.

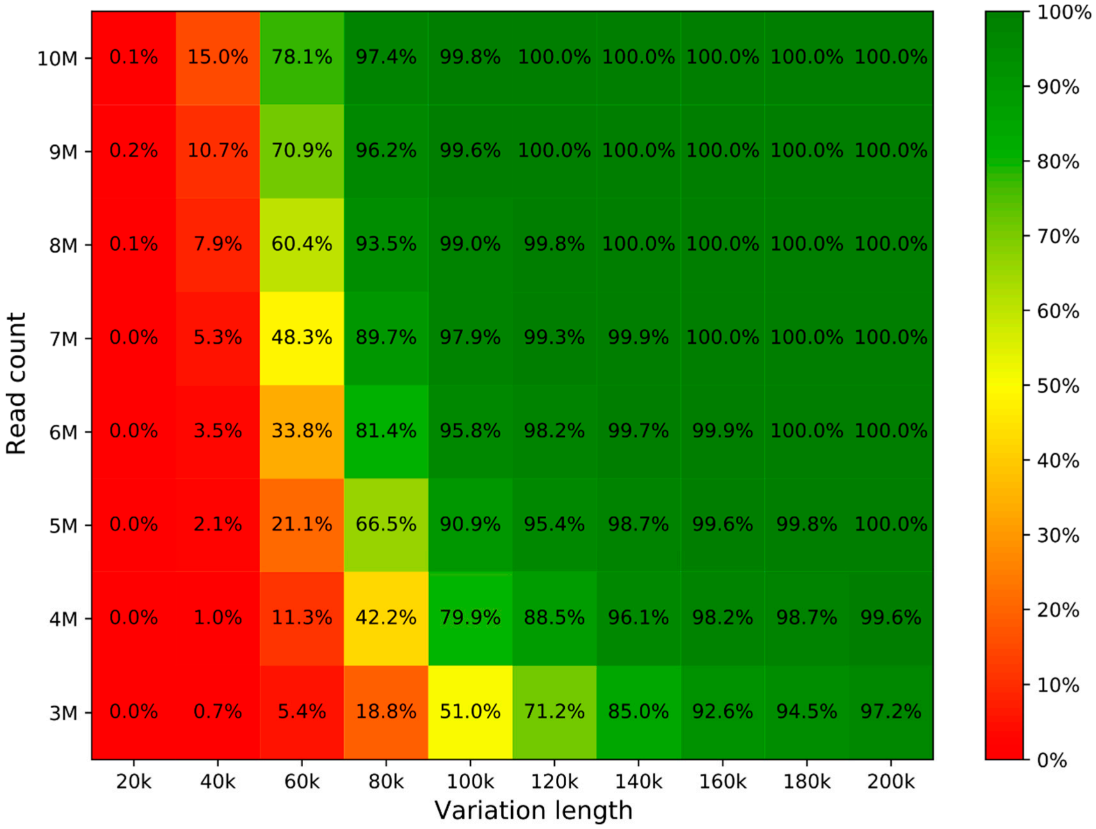

Figure 3. Prediction accuracy computed with in silico analysis based on the length of variation and read count. Each cell number is generated from 8300 simulations (100 randomly generated aberrations; 83 samples).

3.3. Validation of Clinical Samples

Finally, we ran an evaluation of samples analyzed previously in diagnostic settings using the aCGH method (Human Genome CGH Microarray 4 × 44 K Agilent [28]) and GenomeScreen. The selected aCGH method has 42,494 probes, which result in mean accuracy of detection of approximately 200 kb. However, the probes are focused mainly in gene regions and very sparsely in intergenomic regions; therefore, accuracy will be higher within the gene regions and lower outside the genes.

From the 106 tested samples, 58 did not show any detection on aCGH, and the rest contained 59 detections in total (lengths from 39 kb to 146 Mb), all of which were also detected by GenomeScreen. The detections on GenomeScreen and on aCGH show excellent concordance—median overlap of 94.37% (more data in Supplementary material Table S1). GenomeScreen furthermore detected 134 additional variations with ranges from 80 kb to 1.48 Mb, mainly in the regions with a low number of aCGH probes and protein-coding genes, where aCGH has low coverage (Figure 4 and Supplementary Material Table S1 and Figure S1).

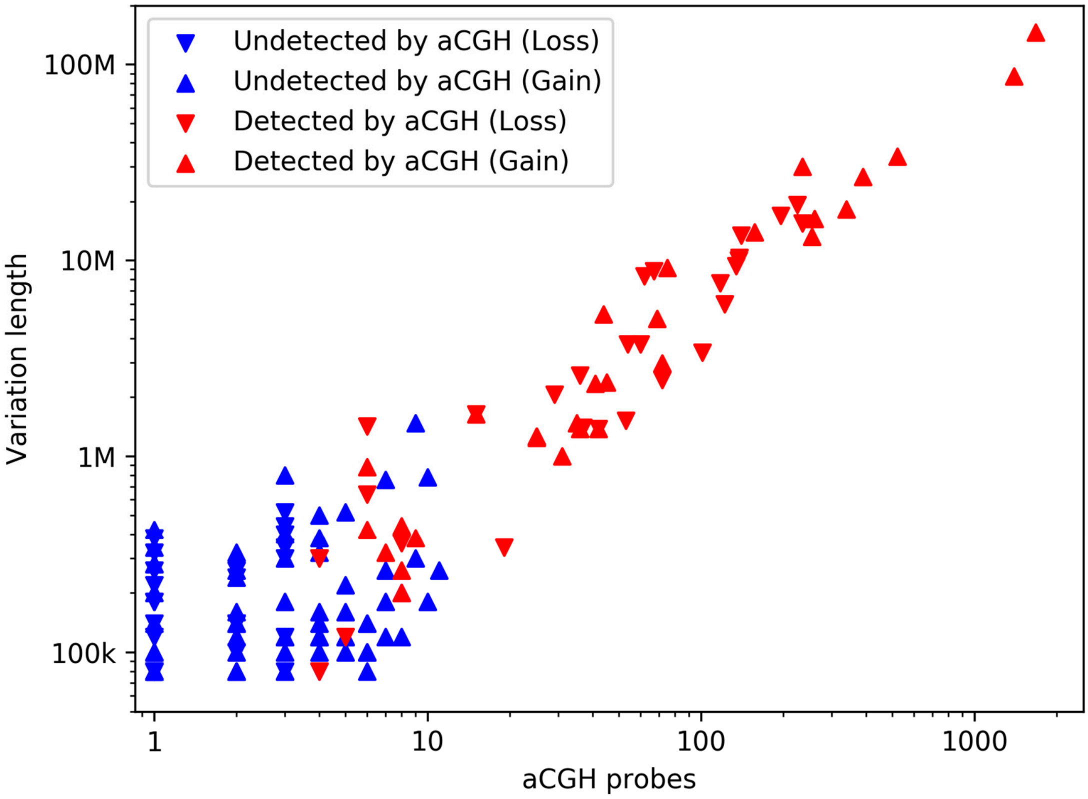

Figure 4. Detection of GenomeScreen (all) and array-based comparative genomic hybridization (aCGH) (red) based on the variation length and number of aCGH probes in the detected interval (by GenomeScreen). Deletions and duplications are visualized by downward and upward triangles, respectively.

4. Discussion

GenomeScreen test is a result of evolving laboratory methods and bioinformatic tools validated in our laboratory and is currently available commercially. The assay originated from a basic NIPT test focused on noninvasive prenatal screening for the three most common trisomies. Later, the development continued by adding the detection of sex chromosome aneuploidies and five selected microdeletions, and most recently it has been advanced to a whole-genome scan for chromosomal microaberrations [18,24,25]. The common link between all these tests is the method based on low-coverage, whole-genome sequencing. Because all the versions of the above-mentioned NIPT tests are intended only for screening, we wanted to validate the method also for diagnostic purposes with much broader applicability in prenatal and postnatal diagnostics. One of the key applications is the replacement of aCGH as the confirmatory method in noninvasive prenatal diagnostics. Therefore, in the pilot phase, the method was validated on plasma and amniotic fluid samples, while the analysis was later extended to chorionic villi, placental tissue, blood, buffy coat, and fetal tissue.

GenomeScreen uses a binning approach, and the genomic coordinates of detected variations are reported as a multiplier of the bin size (20 kb). Nevertheless, the prediction of exact coordinates of the variation is not perfect (see Section 3.2), and it is therefore not suitable for precise CNV detection at the level of exons. On the other hand, the aCGH method uses probes, which can be seen as variable-size bins, where the resolution is equal to the probe distance (which is sometimes larger than the 20 kb bin size). The precision of both GenomeScreen and aCGH can be easily increased (by decreasing the bin size and deeper sequencing in the case of GenomeScreen, or by introducing new probes in case of aCGH), but these adjustments inevitably bring higher production cost.

The overall accuracy strongly depends on the depth of sequencing (see Figure 3). If we set the GenomeScreen sequencing depth to achieve a slightly higher accuracy compared with aCGH, the cost per sample is 2–3 times lower for GenomeScreen. Furthermore, the turnaround time from submission of a sample to completion of the whole process including the analysis takes less time in the case of GenomeScreen (typically 2 to 5 days), whereas the aCGH process may take up to 2 weeks when culturing is required. The culturing or DNA amplification is usually required in NIPT setting, since the amount of retrieved DNA is not sufficient for direct application of aCGH. However, even without these prior preparations, the hybridization process itself takes at least 3 days to deliver the result.

The disadvantage of GenomeScreen is the necessity to train the used normalization on at least 100 nonaberrated samples (training on fewer samples results in filtration of an unnecessary large number of bins due to high variability), but we recommend using as many samples as possible for training. The training should be performed separately for each sample type (and/or different laboratory protocol), however, the trained parameters are quite close across the different sample types that we studied. The parameters can therefore be reused with only a slight decrease in accuracy and noise in CNV profiles. We did not experiment with different laboratory protocols; thus, we cannot assess how it may affect the training parameters. The need for re-training for different laboratory processing of the samples and/or sample types makes this approach difficult to test on datasets other than our own since the datasets available usually do not contain enough samples and information to train and test GenomeScreen. The study is based on analyses of 789 training and 106 control samples with both groups of plasma type.

The false-positive rate of GenomeScreen has not been studied in this paper and should be adequately addressed in the future. However, the loss or gain of the (nonmosaic) deviation with a length of at least 100 kb is so substantial that we do not expect to observe any false-positive detections.

One substantial, albeit only technological advantage of the GenomeScreen method, is the involvement of the same laboratories, protocols, chemistries, instruments, and laboratory technicians for both the screening NIPT test and the confirmatory GenomeScreen test. This was not possible in the case of the confirmatory aCGH test due to entirely different protocols, corresponding infrastructure, and chemistry. The ability to use a method and its modifications with the same technical specification for screening as well as diagnostics (subsequent and/or confirmatory) is rarely encountered in laboratory medicine. Therefore, the presented study results fit into the trend of unification of processes on the part of laboratory work as well as bioinformatics and its utilization in different fields of clinical testing.

5. Conclusions

In this article, we presented a new method for CNV detection based on low-coverage, whole-genome sequencing—GenomeScreen. We estimated its theoretical sensitivity and conducted a series of in silico tests to estimate it in a semi-real setting. Next, we compared this method directly with a commonly used aCGH method on 48 control samples with known aberrations. The new method detected all of the known aberrations and found even more aberrations mainly in intergenic regions where the studied aCGH delivers poor coverage.

According to the presented results, GenomeScreen is currently able to detect almost all variations longer than 100 kb in mappable regions of the human genome. Moreover, it is cheaper and offers shorter turnaround times in comparison with the studied aCGH method. Thus, in the presented laboratory settings, it represents a favorable replacement for the more conventional aCGH method to detect CNVs longer than 100 kb.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/diagnostics11040708/s1, Figure S1: Detection of GenomeScreen (all) and array-based comparative genomic hybridization (aCGH) (red) based on the variation length and number of protein-coding genes in the detected interval (by GenomeScreen). Deletions and duplications are visualized by downward and upward triangles, respectively, Table S1: The details of CNV in used clinical samples.

Author Contributions: Conceptualization, T.S., G.M. and J.B.; methodology, M.K., M.H. and G.M.; software, M.K.; validation, M.K., M.H. and G.M.; investigation, M.K.; resources, M.H.; data curation, M.K. and M.H.; writing—original draft preparation, M.K.; writing—review and editing, M.K., M.H., G.M. and J.B.; visualization, M.K.; supervision, J.B.; funding acquisition, G.M. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This publication was supported by the project "Long-term strategic research and development focused on the occurrence of Lynch syndrome in the Slovak population and possibilities of prevention of tumors associated with this syndrome" (ITMS 313011V578) co-financed by the European Regional Development Fund (ERDF). The article was also created with the support of the OP Integrated Infrastructure for the project: Introduction of an innovative test for screening and monitoring of cancer patients—GenoScan LBquant, ITMS: NFP313010Q927, co-financed by the ERDF. The data infrastructure was built with the support of the Operational Program Integrated Infrastructure within the following project: "Horizontal ICT support and centralized infrastructure for research and development institutions", ITMS code 313011F988, co-financed by the ERDF.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved on 30 June 2015 by the Ethics Committee of the Bratislava Self-Governing Region (03899/2015/HF).

Informed Consent Statement: All participants gave their informed consent for inclusion before participation.

Data Availability Statement: Scripts (Python 3.7) and data are available on the website https://github.com/marcelTBI/GenomeScreen (accessed on 14 April 2021).

Acknowledgments: We thank Ondrej Pös and Zuzana Kubiritová for their valuable comments and help with editing the manuscript text.

Conflicts of Interest: We declare a potential competing financial interest in the form of employee contracts (see affiliations for each author) with Geneton Ltd. and TrisomyTest Ltd. Geneton Ltd. participated in the development of a commercial NIPT test in Slovakia; however, it is not a provider of this commercial test, but continues to do basic and applied research in the field of NIPT. On the other hand, TrisomyTest Ltd. is the commercial provider of NIPT testing in Slovakia. Its participation in the study was limited to the routine NIPT testing that generated the genomic results reused in our research. Related to this work, there are no patents, products in development, or marketed products to declare. The authors declare no other conflict of interest.

Abbreviations

aCGH	array-based comparative genomic hybridization
CBS	circular binary segmentation
CNV	copy number variant
NGS	next-generation sequencing
NIPT	non-invasive prenatal testing
WGS	whole-genome sequencing

References

1. Pös, O.; Budis, J.; Kubiritova, Z.; Kucharik, M.; Duris, F.; Radvanszky, J.; Szemes, T. Identification of Structural Variation from NGS-Based Non-Invasive Prenatal Testing. *Int. J. Mol. Sci.* **2019**, *20*, 4403. [CrossRef] [PubMed]
2. Yoon, S.; Xuan, Z.; Makarov, V.; Ye, K.; Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **2009**, *19*, 1586–1592. [CrossRef] [PubMed]
3. Bartha, Á.; Győrffy, B. Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers* **2019**, *11*, 1725. [CrossRef] [PubMed]
4. Russo, C.D.; Di Giacomo, G.; Cignini, P.; Padula, F.; Mangiafico, L.; Mesoraca, A.; D'Emidio, L.; McCluskey, M.R.; Paganelli, A.; Giorlandino, C. Comparative study of aCGH and Next Generation Sequencing (NGS) for chromosomal microdeletion and microduplication screening. *J. Prenat Med.* **2014**, *8*, 57–69.
5. Wang, H.; Nettleton, D.; Ying, K. Copy number variation detection using next generation sequencing read counts. *BMC Bioinform.* **2014**, *15*, 109. [CrossRef]
6. Fromer, M.; Moran, J.L.; Chambert, K.; Banks, E.; Bergen, S.E.; Ruderfer, D.M.; Handsaker, R.E.; McCarroll, S.A.; O'Donovan, M.C.; Owen, M.J.; et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **2012**, *91*, 597–607. [CrossRef]

7. Krumm, N.; Sudmant, P.H.; Ko, A.; O'Roak, B.J.; Malig, M.; Coe, B.P.; Quinlan, A.R.; Nickerson, D.A.; Eichler, E.E.; Project, N.E.S. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **2012**, *22*, 1525–1532. [CrossRef]
8. Jiang, Y.; Oldridge, D.A.; Diskin, S.J.; Zhang, N.R. CODEX: A normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* **2015**, *43*, e39. [CrossRef] [PubMed]
9. Packer, J.S.; Maxwell, E.K.; O'Dushlaine, C.; Lopez, A.E.; Dewey, F.E.; Chernomorsky, R.; Baras, A.; Overton, J.D.; Habegger, L.; Reid, J.G. CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics* **2016**, *32*, 133–135. [CrossRef] [PubMed]
10. Fowler, A.; Mahamdallie, S.; Ruark, E.; Seal, S.; Ramsay, E.; Clarke, M.; Uddin, I.; Wylie, H.; Strydom, A.; Lunter, G.; et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res.* **2016**, *1*, 20. [CrossRef] [PubMed]
11. Johansson, L.F.; van Dijk, F.; de Boer, E.N.; van Dijk-Bos, K.K.; Jongbloed, J.D.H.; van der Hout, A.H.; Westers, H.; Sinke, R.J.; Swertz, M.A.; Sijimons, R.H.; et al. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum. Mutat.* **2016**, *37*, 457–464. [CrossRef] [PubMed]
12. Rajagopalan, R.; Murrell, J.R.; Luo, M.; Conlin, L.K. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med.* **2020**, *12*, 14. [CrossRef]
13. Raman, L.; Dheedene, A.; De Smet, M.; Van Dorpe, J.; Menten, B. WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res.* **2019**, *47*, 1605–1614. [CrossRef]
14. Straver, R.; Sistermans, E.A.; Holstege, H.; Visser, A.; Oudejans, C.B.M.; Reinders, M.J.T. WISECONDOR: Detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. *Nucleic Acids Res.* **2014**, *42*, e31. [CrossRef] [PubMed]
15. Talevich, E.; Shain, A.H.; Botton, T.; Bastian, B.C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **2016**, *12*, e1004873. [CrossRef] [PubMed]
16. Abyzov, A.; Urban, A.E.; Snyder, M.; Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **2011**, *21*, 974–984. [CrossRef] [PubMed]
17. Dharanipragada, P.; Vogeti, S.; Parekh, N. iCopyDAV: Integrated platform for copy number variations—Detection, annotation and visualization. *PLoS ONE* **2018**, *13*, e0195334. [CrossRef]
18. Hyblova, M.; Harsanyova, M.; Nikulenkov-Grochova, D.; Kadlecova, J.; Kucharik, M.; Budis, J.; Minarik, G. Validation of Copy Number Variants Detection from Pregnant Plasma Using Low-Pass Whole-Genome Sequencing in Noninvasive Prenatal Testing-Like Settings. *Diagnostics* **2020**, *10*, 569. [CrossRef]
19. Kucharik, M.; Gnip, A.; Hyblova, M.; Budis, J.; Strieskova, L.; Harsanyova, M.; Pös, O.; Kubiritova, Z.; Radvanszky, J.; Minarik, G.; et al. Non-invasive prenatal testing (NIPT) by low coverage genomic sequencing: Detection limits of screened chromosomal microdeletions. *PLoS ONE* **2020**, *15*, e0238245. [CrossRef]
20. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
21. Zhao, C.; Tynan, J.; Ehrich, M.; Hannum, G.; McCullough, R.; Saldivar, J.-S.; Oeth, P.; Boom, D.V.D.; Deciu, C. Detection of fetal subchromosomal abnormalities by sequencing circulating cell-free DNA from maternal plasma. *Clin. Chem.* **2015**, *61*, 608–616. [CrossRef]
22. Alkan, C.; Kidd, J.M.; Marques-Bonet, T.; Aksay, G.; Antonacci, F.; Hormozdiari, F.; Kitzman, J.O.; Baker, C.; Malig, M.; Mutlu, O.; et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **2009**, *41*, 1061–1067. [CrossRef]
23. Seshan, V.E.; Olshen, A. DNAcopy: DNA Copy Number Data Analysis. R Package Version 1.36.0. 2013. Available online: https://www.researchgate.net/publication/241191458_DNAcopy_A_Package_for_analyzing_DNA_copy_data (accessed on 14 April 2021).
24. Minarik, G.; Repiska, G.; Hyblova, M.; Nagyova, E.; Soltys, K.; Budis, J.; Ďuriš, F.; Sysak, R.; Bujalkova, M.G.; Vlkova-Izrael, B.; et al. Utilization of Benchtop Next Generation Sequencing Platforms Ion Torrent PGM and MiSeq in Noninvasive Prenatal Testing for Chromosome 21 Trisomy and Testing of Impact of In Silico and Physical Size Selection on Its Analytical Performance. *PLoS ONE* **2015**, *10*, e0144811. [CrossRef]
25. Sekelska, M.; Izsakova, A.; Kubosova, K.; Tilandyova, P.; Csekes, E.; Kuchova, Z.; Hyblova, M.; Harsanyova, M.; Kucharik, M.; Budis, J.; et al. Result of Prospective Validation of the Trisomy Test for the Detection of Chromosomal Trisomies. *Diagnostics* **2019**, *9*, 138. [CrossRef] [PubMed]
26. Chandrananda, D.; Thorne, N.P.; Ganesamoorthy, D.; Bruno, D.L.; Benjamini, Y.; Speed, T.P.; Slater, H.R.; Bahlo, M. Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. *PLoS ONE* **2014**, *9*, e86993. [CrossRef] [PubMed]
27. Gazdarica, J.; Budis, J.; Duris, F.; Turna, J.; Szemes, T. Adaptable Model Parameters in Non-Invasive Prenatal Testing Lead to More Stable Predictions. *Int. J. Mol. Sci.* **2019**, *20*, 3414. [CrossRef] [PubMed]
28. Agilent Technologies, Inc. Human Genome Cgh Microarray Kit, 4 × 44 k. Available online: https://www.agilent.com/en/product/cgh-cgh-snp-microarray-platform/cgh-cgh-snp-microarrays/human-microarrays/human-genome-cgh-microarray-kit-4x44k-228410 (accessed on 5 August 2020).