diagnostics

MDPI

Article

Improving CNV Detection Performance in Microarray Data Using a Machine Learning-Based Approach

Chul Jun Goh [1,†], Hyuk-Jung Kwon [1,2,†], Yoonhee Kim [1], Seunghee Jung [1], Jiwoo Park [1], Isaac Kise Lee [1,2,3], Bo-Ram Park [1], Myeong-Ji Kim [1], Min-Jeong Kim [4] and Min-Seob Lee [1,4,*]

[1] Eone-Diagnomics Genome Center, Inc., 143, Gaetbeol-ro, Yeonsu-gu, Incheon 21999, Republic of Korea; cj.ko@edgc.com (C.J.G.); hjkwon@edgc.com (H.-J.K.); yh.kim@edgc.com (Y.K.); sh.jung@edgc.com (S.J.); jw.park@edgc.com (J.P.); ks.lee@edgc.com (I.K.L.); br.park@edgc.com (B.-R.P.); myeongji.kim@edgc.com (M.-J.K.)
[2] Department of Computer Science and Engineering, Incheon National University (INU), Incheon 22012, Republic of Korea
[3] NGENI Foundation, San Diego, CA 92127, USA
[4] Diagnomics, Inc., 5795 Kearny Villa Rd., San Diego, CA 92123, USA; mjkim@diagnomics.com
* Correspondence: mlee@edgc.com; Tel.: +82-10-3080-1393
† These authors contributed equally to this work.

Abstract: Copy number variation (CNV) is a primary source of structural variation in the human genome, leading to several disorders. Therefore, analyzing neonatal CNVs is crucial for managing CNV-related chromosomal disabilities. However, genomic waves can hinder accurate CNV analysis. To mitigate the influences of the waves, we adopted a machine learning approach and developed a new method that uses a modified log R ratio instead of the commonly used log R ratio. Validation results using samples with known CNVs demonstrated the superior performance of our method. We analyzed a total of 16,046 Korean newborn samples using the new method and identified CNVs related to 39 genetic disorders were identified in 342 cases. The most frequently detected CNV-related disorder was Joubert syndrome 4. The accuracy of our method was further confirmed by analyzing a subset of the detected results using NGS and comparing them with our results. The utilization of a genome-wide single nucleotide polymorphism array with wave offset was shown to be a powerful method for identifying CNVs in neonatal cases. The accurate screening and the ability to identify various disease susceptibilities offered by our new method could facilitate the identification of CNV-associated chromosomal disease etiologies.

Keywords: CNV; genome-wide SNP array; Korean newborn; machine learning; genomic wave

Citation: Goh, C.J.; Kwon, H.-J.; Kim, Y.; Jung, S.; Park, J.; Lee, I.K.; Park, B.-R.; Kim, M.-J.; Kim, M.-J.; Lee, M.-S. Improving CNV Detection Performance in Microarray Data Using a Machine Learning-Based Approach. *Diagnostics* **2024**, *14*, 84. https://doi.org/10.3390/diagnostics14010084

Academic Editor: Dechang Chen

Received: 17 October 2023
Revised: 26 December 2023
Accepted: 28 December 2023
Published: 29 December 2023

1. Introduction

Developmental disabilities can impact a range of domains, including perception, cognition, movement, and language. The disabilities predominantly arise from chromosomal abnormalities, such as copy number variations (CNVs). CNVs refer to large deletions or duplications of genomic material that are greater than 1 kilobase (kb) in size [1]. Although most CNVs are functionally benign, they are a common source of genomic structural variation [2–4], and some of the variations are associated with various diseases, such as intellectual disability, autism, schizophrenia, and developmental disorders [5–9]. Therefore, early and accurate detection of CNVs is essential for providing appropriate interventions and support to individuals and families affected by the CNVs.

Numerous methods exist for detecting chromosomal abnormalities, encompassing conventional techniques such as karyotyping, fluorescence in situ hybridization (FISH) [10], and multiplex ligation-dependent probe amplification [11], as well as contemporary approaches like chromosomal microarray analysis (CMA) [12]. These tests can be used to diagnose genetic disorders, including Down syndrome, Turner syndrome, and some forms

of cancer. In particular, chromosomal tests are also applicable for carrier screening, prenatal testing, and newborn screening [13,14].

Similar to whole genome sequencing (WGS) analysis, CMA is a high-resolution technique to screen the entire genome and identify CNVs [15]. While WGS can identify CNVs, SNPs, and other genetic variations, providing all-encompassing information of the entire genome, the cost and time requirements of WGS surpass those of microarray analysis, making it less feasible for regular clinical testing purposes [16,17]. Using an array with probes designed to selectively bind with DNA extracted from a sample, CMA demonstrates the capability to detect CNVs as small as 50–100 kb in size. Notably, this ability enables CMA to detect difficult-to-identify diseases, including developmental disorders and multiple congenital anomalies, with a detection rate of 15–20% [18–20].

CMA allows for the simultaneous detection of both CNVs and rare mutations in a single run [21]. For example, the Illumina Infinium Global Screening Array (GSA) is able to scan approximately 750,000 SNPs across the entire human genome [22]. By utilizing the log R ratio (LRR), a normalized signal intensity value for individual SNP markers, and B allele frequency (BAF), a normalized allelic intensity value for two alleles, data from the microarray, CNVs, and rare mutations can be detected [23]. This approach facilitates comprehensive genetic screening and analysis across diverse populations.

During the analysis of CNVs using microarrays, the presence of wave-like patterns characterized by genome-wide spatial autocorrelation has been noted [24–26]. The patterns were observed at the chromosomal level rather than in narrow subregions. Moreover, those were evident even when copy numbers were normal, attributed to the high variability of LRR. This phenomenon, referred to as a genomic wave, is speculated to be caused by variations in both quantity and quality of DNA. The pattern, observed across all chromosomes and varying between samples, is known to have negative impacts on the accuracy of CNV detection [24].

Since the identification of genomic waves, various methods have been developed and utilized to improve the accuracy of CNV detection in the presence of these waves. These methods include the utilization of Loess [27] and the Genomic Imbalance Map algorithm [28], in addition to the correlation with the guanine-cytosine content of the genome sequence [24]. These strategies serve to alleviate the impact of genomic waves on CNV detection.

In this study, a new method using machine learning models was employed to mitigate the effect of the genomic waves on CNV analysis. Among the different machine learning methods available, k-means [29,30] and k-nearest neighbor (k-NN) [31,32] were selected due to their simplicity and strong performance. Using the approaches, we obtained a new LRR value called modified LRR (mLRR). The effectiveness of the new method on CNV analysis was validated by comparing the results of analyzing samples with known CNVs, and the results from next-generation sequencing (NGS) were utilized to confirm the accuracy of our method. As a result of the validation, the new method showed a greater performance than the original one.

2. Materials and Methods

2.1. Subjects and Sample Preparation

This study was performed in accordance with the 2021 Guidelines for using health data by the Ministry of Health and Welfare of Korea. We analyzed the DNA CNVs in 16,046 peripheral or cord blood samples collected from newborn Korean babies. Each blood sample (0.1 mL) was placed into a BD Microtainer tube with K2EDTA (BD, Franklin Lakes, NJ, USA) and analyzed at clinical centers for genetic analysis between February 2018 and May 2021. The blood samples were transported at room temperature to the laboratory, where genomic DNA was extracted from the blood using a Chemagic DNA Blood 200 Kit (Perkin Elmer, Waltham, MA, USA) according to the manufacturer's protocol. Before performing the microarray assay, the genomic DNA concentration and purity were measured using an EpochTM microplate spectrophotometer (BioTek, Winooski, VT, USA).

Genomic DNA (200 ng) was used to generate targets for the Illumina Infinium HTS assay protocol, and NGS was performed using 500 ng of genomic DNA.

2.2. Subjects and Sample Preparation

We custom-engineered the Illumina Infinium GSA BeadChip (version 2, Illumina, San Diego, CA, USA) to include 742,759 SNP markers capable of detecting 138 CNV-related chromosomal disorders (Figure 1). Procedures for DNA amplification, fragmentation, hybridization, and staining were performed according to the Illumina Infinium HTS assay protocol (Illumina). Image and data files were obtained using the iScan control and Genome Studio software packages (v2.0.4) from Illumina.

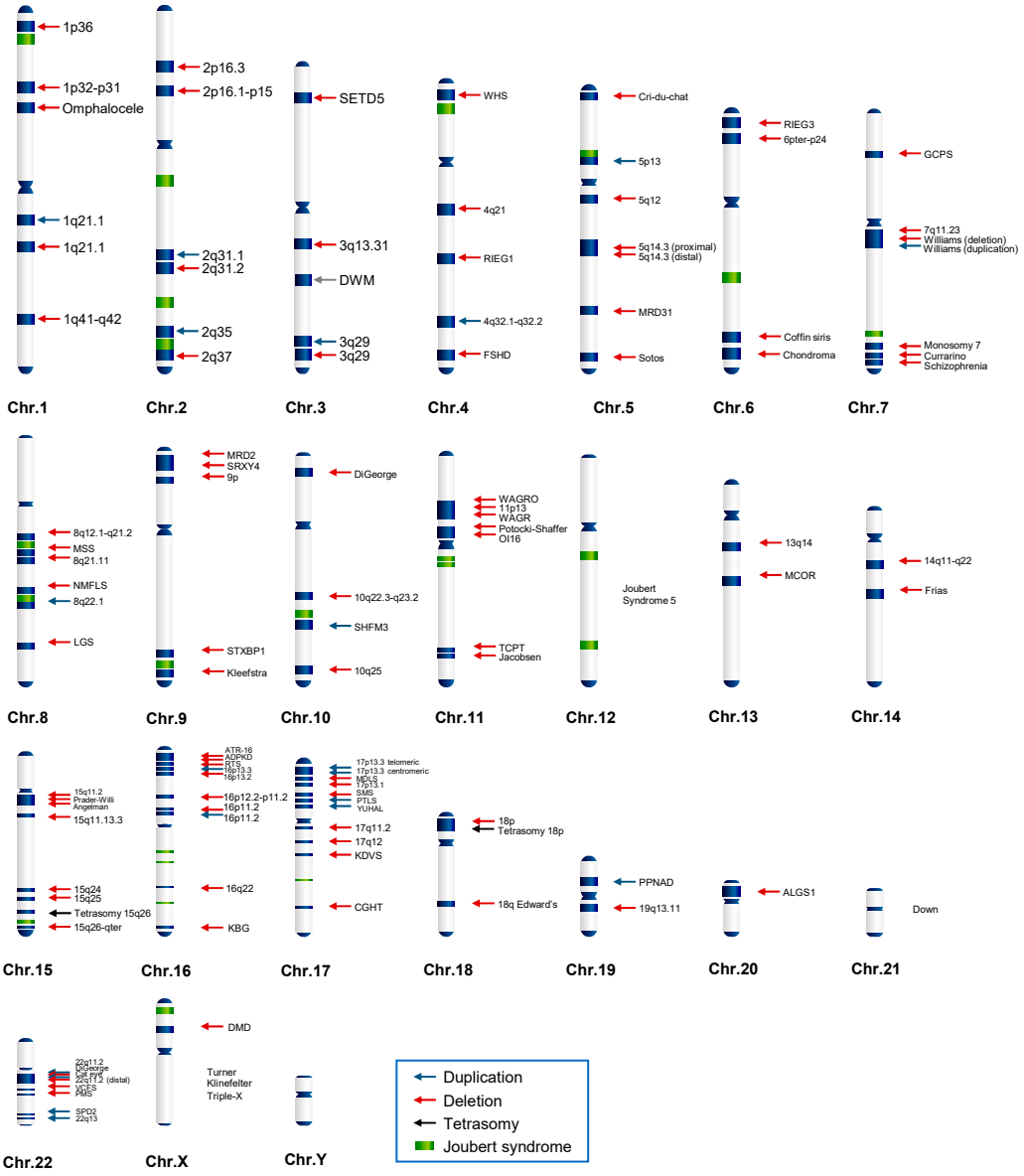

Figure 1. Schematic overview of 138 CNV-related chromosomal disorders. Red, blue, and black arrows indicate microdeletions, duplications, and tetrasomy, respectively. Green bars indicate Joubert syndrome types.

2.3. Preparation of Positive CNV Control Samples for Analytical Validation

The SNP array analysis was validated using 22 human cell line DNA samples provided by the Coriell Institute for Medical Research. Each analysis was repeated 2 or 3 times for reproducibility and accuracy. All experimental methods used in association with the human cell lines were performed in the same manner as those performed for newborn specimens.

2.4. Data Processing and CNV Analysis

Raw data from each sample were processed using in-house tools to generate the signal intensities (expressed as LRR) and allelic intensity ratios (expressed as BAF) of all SNPs. To ensure data quality, only markers with call rates of ≥0.98 and LRR SDs of ≤0.2 were selected.

The PennCNV [33] and QuantiSNP [34] were performed to identify copy number deletions and duplications using population frequencies of the B allele, which were calculated based on the BAF of each marker in 1100 samples. Subsequently, adjacent CNVs that were <200 kb apart were merged and filtered out based on the SNP number (>10), CNV length (>50 kb), and confidence score (>50), all of which were generated using PennCNV and QuaintiSNP. The CNVs detected by each program were compared against our custom database, which contains positional data related to 138 chromosomal disorders associated with CNVs. Following this comparison, only the CNVs that corresponded to each specific disease were selected for further analysis. The results obtained from the programs were merged to reduce false negatives. The final result of the analysis was either the detection of a CNV (if the condition was met) or normal status (if the condition was not met) (Figure 2).

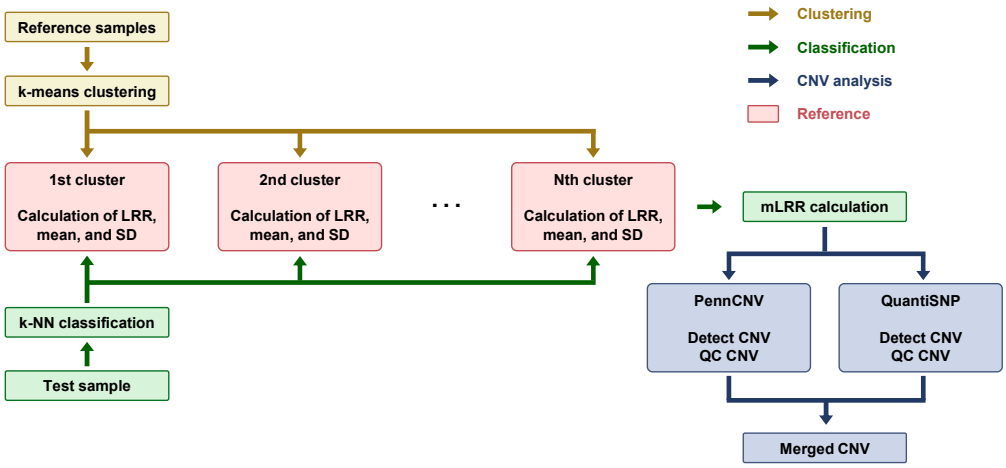

Figure 2. Key steps involved in Copy Number Variation (CNV) analysis using machine learning techniques. Reference samples indicate the clinical samples for clustering genomic waves. Golden brown arrows represent the clustering process, the green color indicates the classification process of the "Test sample", and the navy color indicates the CNV analysis process after "mLRR calculation".

2.5. Clustering Genomic Waves from 5399 Clinical Samples Using GSA

The genomic waves in the results of the GSA chip were identified using 5399 clinical samples. Autosomal chromosomes were divided into 1 Mb bins, and the LRR means, and SDs of markers within the region were calculated. Bins with no markers or LRR SDs of ≤0.05 were excluded from the analysis because small SDs result in even distribution and are not useful for analysis. As such, 238 domains were created, and the LRR mean was used as the feature for analysis.

To cluster the waves into patterns, it was necessary to determine the optimal number of clusters. This was achieved by calculating the sum of distances between the cluster center and its members while incrementally increasing the number of clusters from 2 to 20. The 'elbow point' of the value, as determined using the elbow method [35], indicated that

the decrease of k-means becomes smaller after 6 clusters. As a result, 6 was chosen as the optimal number of clusters.

In total, 5399 samples were clustered using the k-means method with the optimal value of 6. The 6 clusters represented distinct wave patterns and consisted of 768, 1202, 1241, 743, 788, and 657 samples, respectively. To determine the clustered LRR pattern, each sample was divided into 1 Mb portions, and the mean LRR of the included markers was calculated. For samples in the same cluster, the mean LRR mean was calculated between them. The resulting clustered data were subjected to dimension reduction analysis methods, specifically t-distributed stochastic neighbor embedding and principal component analysis. Subsequent plots and comparisons with the k-means clusters were conducted. In both analyses, the samples were effectively categorized into the 6 clusters (Figure 3).

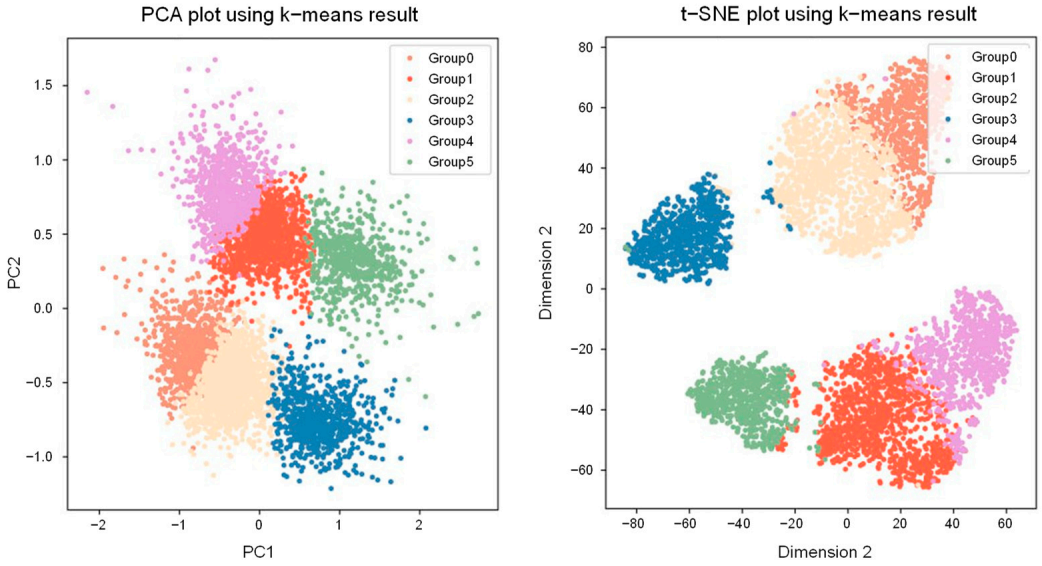

Figure 3. Principal component analysis and t-distributed stochastic neighbor embedding plots of 5399 clinical samples. Six colors represent each of the groups, which represent wave patterns.

2.6. Cluster Matching of Analytical Samples and Calculation of Modified LRR Values

The LRR mean of 238 regions used for k-means analysis was calculated and classified using k-NN. The LRR data for matched cluster samples were normalized into Z-scores using the following formula:

$$Zi = \frac{Xi - \overline{Xi}}{Si}$$

where X represents the LRR value in the sample, $\overline{X}$ represents the mean, and S represents the standard deviation (SD) calculated for the samples within the group.

Due to the differences in the range of normalized values compared to the original LRR, adjustment to the original range was required. This involved resizing the original LRR SD and Z-score SD to a similar value, resulting in the creation of a new LRR value referred to as the modified LRR (mLRR). The offset effect was verified at the chromosome level: a wave was observed in the results using the LRR, while no wave was observed in the results of the mLRR (Figure 4a). This phenomenon was particularly pronounced at each end of the chromosomes (Figure 4b).

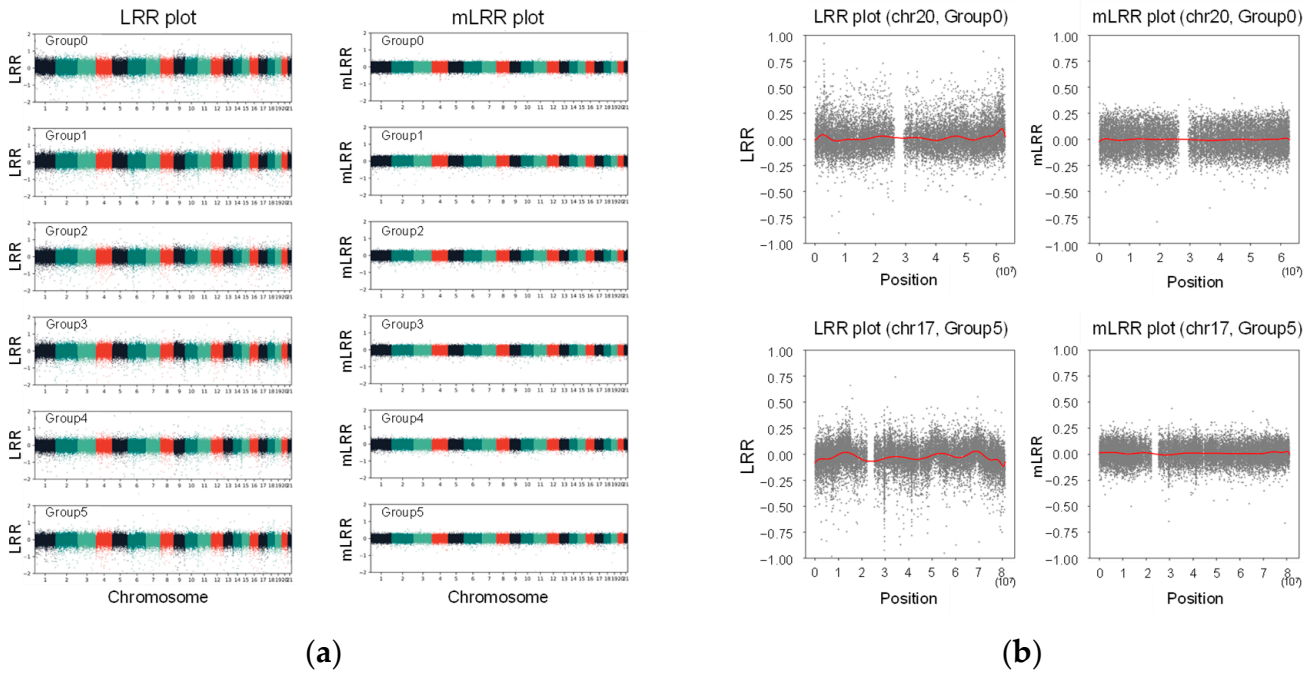

Figure 4. Comparison between Log R ratio (LRR) and modified LRR (mLRR) plots for each cluster. The horizontal axis of the plot represents the chromosome numbers, while the vertical axis represents the LRR and mLRR values: (**a**) Total results from the 21 chromosomes are represented. Each chromosome is identified by a different color; (**b**) Some examples of the end of the chromosomes are shown. The red lines represent the trend lines.

2.7. NGS Sequencing for Accuracy Validation

To confirm the accuracy of our method using NGS, all genomic DNA passing our QC criteria (OD260/OD280 $\geq$ 1.8; 1.9 $\leq$ OD260/OD230 $\leq$ 2.2) were prepared for library construction. Briefly, 30 ng of genomic DNA was sheared into small fragments (170–200 bp) using a focused M220 ultrasonicator (Covaris, Woburn, MA, USA). Following end repair, the addition of an A overhang, and adapter ligation, all ligated fragments were cleaned up using Hiaccubead magnetic beads (Accugene, Incheon, Korea). Libraries were prepared using the Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences, Ann Arbor, MI, USA) according to the manufacturer's protocols. The size distribution of each library was assessed using a 4200 Tapestation system (Agilent Technologies, Palo Alto, CA, USA). The libraries were sequenced using an Illumina NextSeq platform with paired-end sequencing (36 $\times$ 2) following the manufacturer's protocols.

The Ion Torrent Proton platform from Thermo Fisher Scientific was also used as follows. Libraries were prepared using an Ion AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific, Waltham, MA, USA). Adapter ligation, end repair, PCR amplification, and barcoding were performed using an Ion Xpress Adapter 1–96 Kit (Thermo Fisher Scientific). An Ion Chef system was used to complete emulsion PCR and enrichment steps according to the manufacturer's protocol. The resulting libraries were sequenced using an Ion Torrent Proton system with an Ion PI Chip Kit V3 (Thermo Fisher Scientific).

The sequencing data were aligned to the hg 19 human reference genome using Burrows-Wheeler Aligner (ver. 0.7.15) [36]. Using in-house software, duplicated reads were removed, and read depths and z-scores for each position were calculated.

3. Results

3.1. Enhancing CNV Analysis Accuracy through Customized Machine Learning Model

To address the issue of wave patterns that can impede accurate CNV analysis using an array, we developed a customized machine-learning analysis. This involved three

processes: (1) clustering wave patterns with k-means, (2) classifying samples into their nearest cluster using k-NN, and (3) utilizing the original LRR standard deviation (SD), mean values, and Z-score SD values. These approaches enabled us to normalize and offset waves, thereby improving the accuracy of the array analysis. Modified log R ratio (mLRR) values, which would serve as input parameters for the CNV analysis tool, were obtained.

3.2. Analytical Validation Using Known Positive CNV Control DNA Samples

A total of 22 samples from the Coriell cell line repository with defined chromosomal abnormalities were analyzed using LRR and mLRR values to assess performance for the detection of CNVs between the two values. To confirm reproducibility and evaluate the accurate performance, the analysis was repeated 2 or 3 times for each sample (Figure 5 and Table S1).

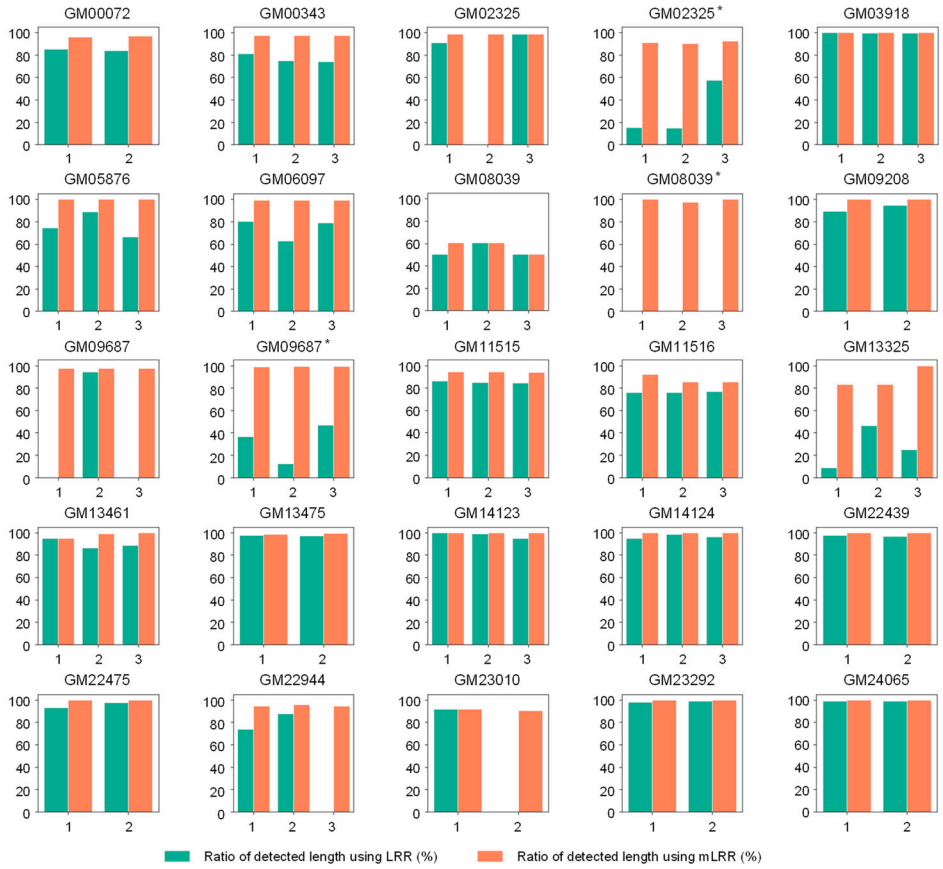

Figure 5. The results of CNV detection from Coriell cell line repository. The vertical axis represents the ratio of detected length to known CNV length. The horizontal axis indicates the number of repeats for each analysis. The green color indicates the results of the analysis using LRR, and the orange color represents the results of the analysis using mLRR. If the result is not detected, no bar is displayed. Asterisks (*) indicate cell lines that have two CNV regions. Please refer to Table S1 for more information.

As a result, the utilization of mLRR yielded a more powerful detection performance than the original one. All the known CNV regions were detected with our new method from the 67 analyses, whereas some regions were missed in the analysis based on the standard LRR values (Figure 5 and Table S1). Among the 67 repeats, 7 (10.45%) were not detected in the standard LRR analysis. The length of the detected CNV was higher or the same as that of mLRR in all analyses, except for cases where it was not detected. For example, in the case of GM08039 with a known CNV spanning 22,723,028 bp associated with Trisomy 16, the mLRR method detected 99.977%, 95.508%, and 99.977% of the CNV

region in three repeats, respectively. In contrast, when LRR was employed, CNV was not detected in 2 analyses out of 3 repeats, and only 0.404% was detected in one case. In the analysis of GM05876, which harbors a 1,435,491 bp CNV associated with DiGeorge syndrome, the CNV was detected in all analyses using LRR and mLRR. However, when LRR was employed, only 8.792%, 46.376%, and 24.800% were detected in three repeated analyses, respectively. In contrast, the method using mLRR exhibited high detection rates of 83.205%, 83.205%, and 99.843% (Figure 5 and Table S1).

3.3. CNV Analysis Using 16,046 Neonate Samples from South Korea

From February 2018 to May 2021, we collected 16,046 neonate samples from the clinical centers located in South Korea. We utilized the mLRR values, whose performance had been validated, to analyze the samples and attempted to detect 138 CNV-related chromosomal disorders (Figure 1) using a customized GSA BeadChip.

As a result of the screening, the genome-wide SNP array chip targeting 138 CNV-related chromosomal disorders identified 342 cases of 39 CNV-associated chromosomal disorders (Figure 6 and Table S2).

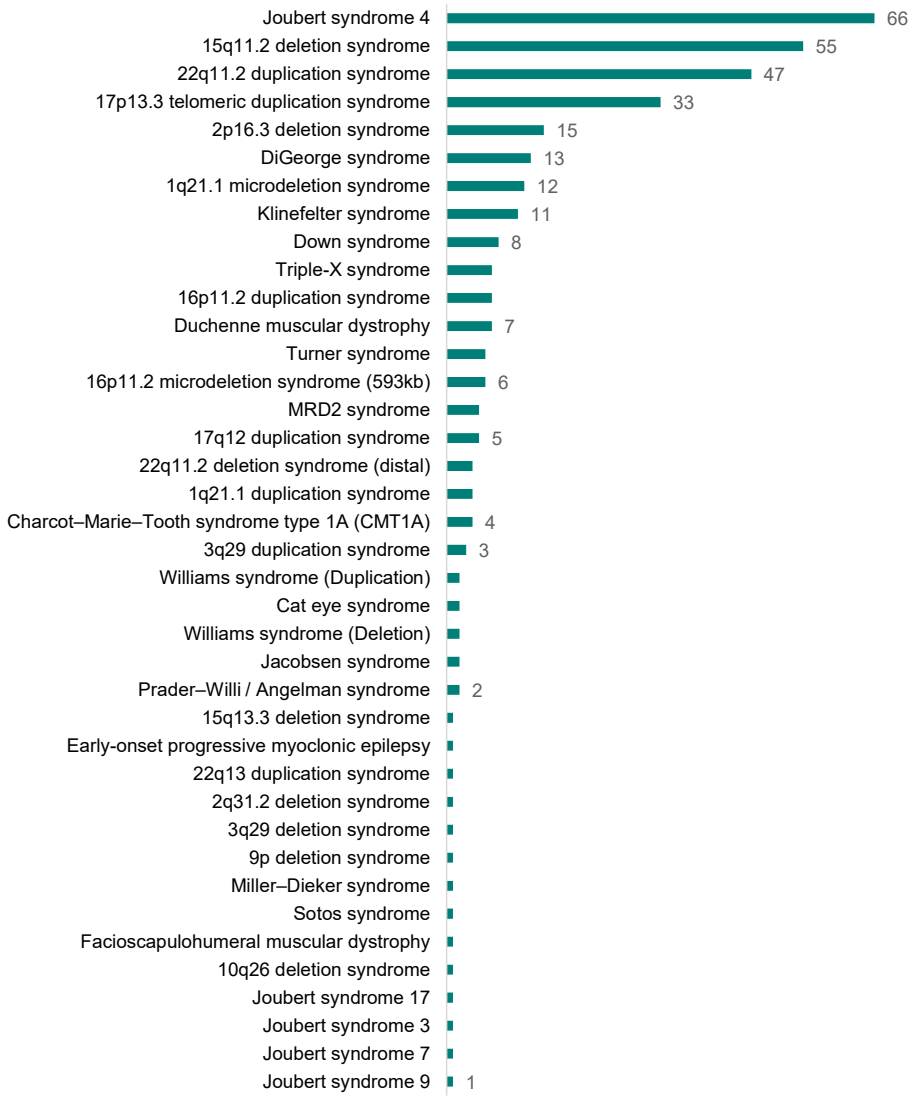

Figure 6. The number of identified chromosomal disorders from the screening of the 16,046 neonate samples. The numbers next to each bar represent the detected number, and bars of the same height represent the same number. Please refer to Table S2 for more information.

The most frequently detected disorder was Joubert syndrome 4 in the 2q13 region (66 of 342 cases). The 2q13 microdeletion encompasses genes encoding a MAL-like protein and nephrocystin 1 (NPHP1). A homozygous deletion of NPHP1 on chromosome region 2q13 is known to cause a rare genetic disorder, Joubert syndrome 4 [37,38]. The syndrome shows a condition in which parts of the brain do not develop properly. All the 66 cases of 2q13 deletions were identified as heterogeneous deletions [arr[hg19] 2q13 (110,852,875–110,983,320) × 1]. The detected cases were presumed to be carriers of the Joubert syndrome 4.

The second most commonly detected chromosomal abnormalities (55 cases) were located in the 15q11.2 region. The 15q11.2 microdeletion pertains to a characteristic 500 kb (0.5 Mb) deleted segment situated between breakpoint 1 (BP1) and breakpoint 2 (BP2). Approximately 8–10% of the individuals with 15q11.2 deletions exhibit characteristics such as developmental delays in motor and language skills [39]. Disruptions of genes within the 15q11.2 region result in an autosomal dominant form of disability with low penetrance. It might offer a plausible explanation for the higher-than-normal frequency in the population. The prevalence of this relatively high frequency is corroborated by a study carried out by the University of Kansas Medical Center [40], which highlights that CNVs at the 15q11.2 BP1-BP2 microdeletion region are estimated to be present in 0.5% to 1.0% of the population.

The third and fourth most frequently detected CNV-related disorders were the 22q11.2 duplication syndrome and the 17p13.3 telomeric duplication syndrome, which were detected in 47 and 33 cases, respectively. The features of the 22q11.2 duplication syndrome, which is caused by an extra copy of a piece of chromosome 22 containing about 30–40 genes, are known to be varied even among family members (i.e., intrafamilial variability exists) [41]. Some with the duplicated gene exhibit intellectual or learning disabilities in addition to developmental delay, slow growth, and weak muscle tone (hypotonia) [42]. Duplications involving one or more genes on chromosome 17p13.3 are associated with split-hand/foot malformation and long-bone deficiency-3 (SHFLD3), with the duplication of the basic helix-loop-helix transcription factor of the A9 (BHLHA9) gene especially associated with limb defects. SHFLD3 is a relatively rare autosomal dominant skeletal disease with a penetration rate of <50% and features a broad spectrum of intraindividual variability [43,44]. The following genetic disorders involving CNVs were found in 11–15 of 16,046 cases: 2p16.3 deletion, DiGeorge, 1q21.1 microdeletion, and Klinefelter syndromes (Table S2).

To verify the accuracy of our GSA array–based approach in a clinical setting, a comparison was performed with the results obtained using next-generation sequencing (NGS). From the results of the 16,046 samples, we selected CNV-associated chromosomal disorders that were frequently detected in this study, as well as those known to be rare, for comparison (Figure 7 and Figure S1).

The same DNA samples, analyzed with the custom-engineered chip, were analyzed using NGS. Read depths for each position were computed using reads aligned to hg19, the human reference genome. Subsequently, z-scores were calculated for all positions based on these read depths. Our analysis revealed a distinctive variation in z-score within the genomic region where the GSA-identified CNV was located, distinguishing it from adjacent positions.

Comparing the NGS results with the genome-wide SNP array analysis demonstrated complete consistency, achieving a 100% match. The CNVs identified through the array analysis were precisely mirrored in the NGS findings. This underlines the high consistency and robustness of our method in accurately detecting various CNVs, showcasing its strong performance and reliability.

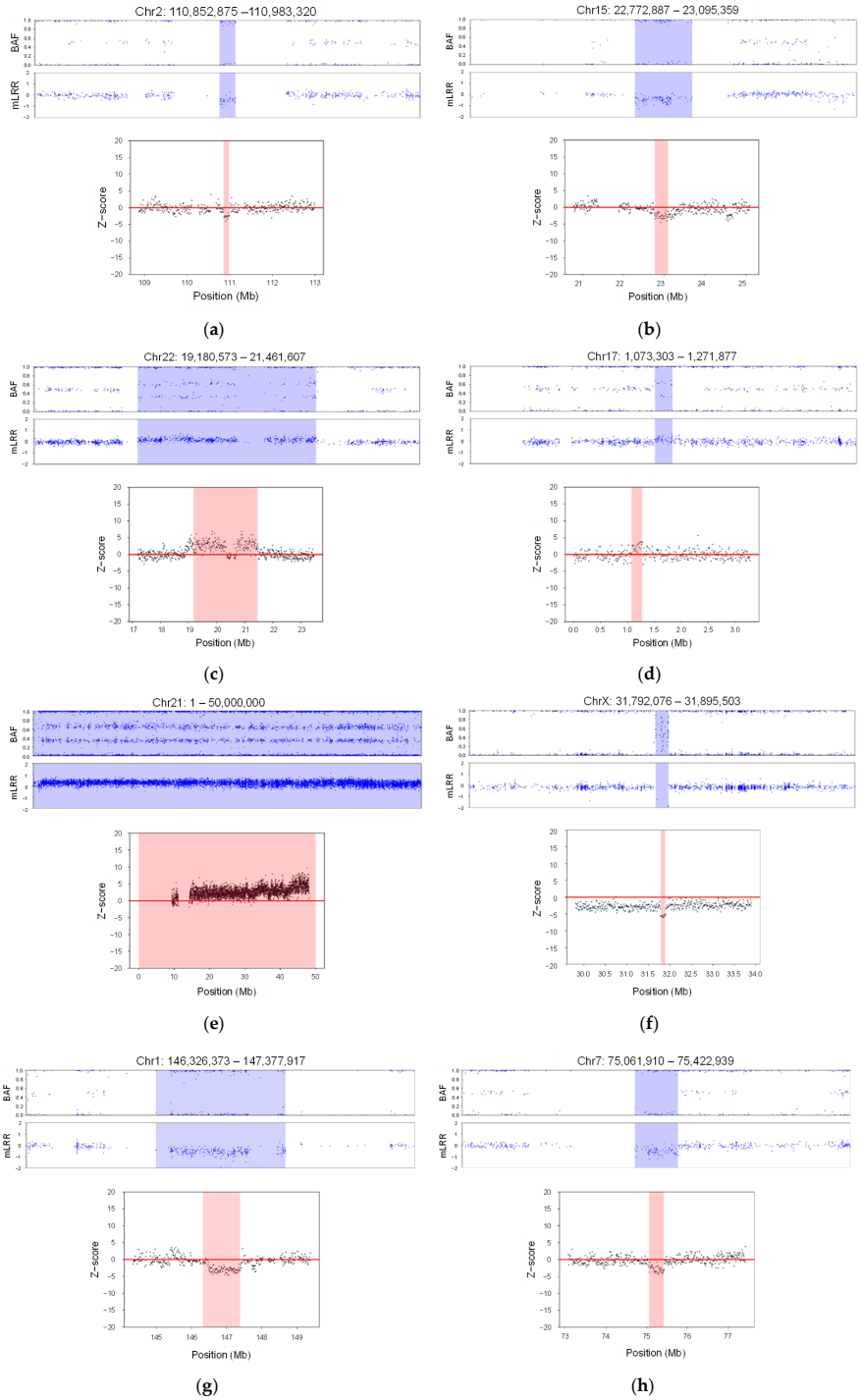

Figure 7. Validation results analyzing the same samples with a GSA array–based approach and NGS. A total of 8 chromosomal disorders are shown. The title displays the chromosome number along with the start and end positions of the detected region. The upper panel illustrates analysis results using the signal intensity patterns (B allele frequency, BAF) and modified log R ratio (mLRR). The vertical axis represents BAF and mLRR values, and each blue dot represents each value. The light blue color indicates the detected regions from the GSA. The lower panels represent the results from NGS analysis. The light pink regions represent the detected regions from the NGS. The vertical axis represents the z-score values, and the horizontal axis represents the positions: (**a**) Joubert syndrome 4; (**b**) 15q11.2 deletion syndrome; (**c**) 22q11.2 duplication syndrome; (**d**) 17q13.3 telomeric duplication syndrome; (**e**) Down syndrome; (**f**) Duchenne muscular dystrophy; (**g**) 1q21.1 deletion syndrome; (**h**) Williams syndrome (deletion).

4. Discussion

CNVs are associated with many neurodevelopmental disorders, such as autism spectrum disorders, schizophrenia, intellectual disability, attention deficit hyperactivity disorder, developmental delay, and epilepsy [45–47]. As the variations often span several mega-base pairs that encompass multiple genes [48,49], the rarity and dose-sensitive nature of individual CNV genes must be accurately determined. For instance, altering the copy number of a dose-sensitive gene like BHLHA9 can be detrimental to disease pathogenesis, whereas changing the copy number of dose-insensitive genes is unlikely to cause harm [50]. Even within the same chromosomal region, CNVs may be associated with different phenotypes, ranges of severity, and incomplete penetration [51]. In the present study, among 39 detected cases of chromosomal abnormalities, 8 CNV-related chromosomal disorders were known to have complete penetrance: DiGeorge, Klinefelter, Down, Triple-X, Turner, Williams (duplication and deletion), and Prader–Willi/Angelman syndromes. For example, Duchenne muscular dystrophy (OMIM#310200), an X-linked recessive myopathy caused by a mutation in the dystrophin gene located at Xp21, has a penetration rate of 100% in males [52]. Charcot–Marie–Tooth syndrome type 1A (CMT1A), which was detected in four cases in our study, exhibits varying penetration rates depending on the parent. Fathers with X-linked dominant CMT1A have a 100% risk of having an affected daughter, whereas their sons face no such risk; conversely, both sons and daughters of mothers with X-linked dominant CMT1A have a 50% chance of being affected by the syndrome [53]. Joubert syndrome 4, the most frequently detected CNV-related chromosomal abnormality in our study, is a rare autosomal recessive disorder involving a ~290 kb homozygous deletion containing NPHP1 in 2q13. All Joubert syndrome 4 cases identified in our screening were heterogeneous deletions at 2q13; thus, the individuals were assumed to be carriers of the disorder in all cases. Other chromosomal disorders (i.e., 15q11.2 deletion, 22q11.2 duplication, 17p13.3 telomeric duplication, and 2p15.2 deletion syndromes, among others) detected in this study represent syndromes with various penetration rates. Additionally, while disorders involving visual and hearing impairments are often detected early, e.g., before the age of three, invisible autism, as well as emotional and behavioral disorders, are more likely to occur after the age of three [54]. Thus, it is very important to detect chromosomal abnormalities early and accurately in order to minimize the symptoms of the disease and slow its progression.

As a comprehensive and universal screening tool, the GSA method can detect various genetic abnormalities with high accuracy, making it a reliable option for large-population screening compared to other screening methods. The array offers a resolution level that is more than 10 times higher than conventional karyotyping or FISH analysis, allowing for the detection of micro-chromosomal abnormalities that are larger than 100 kb in size with higher confidence levels [55–57].

However, it is crucial to acknowledge the limitations of genotyping arrays; they are unable to detect translocations and inversions [58]. Recent studies have highlighted that low-pass genome sequencing technology surpasses microarray technology in terms of detection rate, resolution, and cost-effectiveness [59–62]. Nonetheless, the GSA method continues to be extensively utilized in diagnostic and research due to its relatively low cost and sample requirement.

Our study was focused on mitigating the disruptive influence of genomic waves—recurring wave-like patterns pervasive across the genome—which significantly impair the accuracy of detecting copy number variations (CNVs). By addressing these inherent challenges posed by genomic waves, our aim was to develop methodologies or techniques that improve the precision and reliability of CNV detection within genetic data analysis. To address this, we specifically employed k-means and k-NNs for clustering wave patterns and classifying samples, considering their simplicity and interpretability crucial when handling high-dimensional microarray data with hundreds of thousands of probes.

While these methods proved effective, the evolving realm of machine learning holds the potential for achieving even higher CNV detection accuracy in a neonatal setting. Recent

advances in machine learning, particularly in deep learning and ensemble methods such as convolutional neural networks (CNN) and random forest, have demonstrated exceptional classification performance in medical imaging and molecular diagnosis applications [63,64]. CNNs, known as powerful tools for image recognition, could be instrumental in identifying CNVs in abnormal chromosomes by extracting key features like edges and specific banding patterns, which are crucial for detecting small chromosomal deletions and duplications. Nevertheless, the utilization of intricate machine learning models alongside high-quality data brings about trade-offs, including escalated costs, risks of overfitting, and challenges in interpretability, necessitating careful consideration in future research endeavors.

The utilization of machine learning methods extends far beyond the scope of this study, encompassing widely employed techniques such as k-means and k-NNs, which have significant applications in the diagnosis and exploration of diseases such as Autism Spectrum Disorder (ASD) [65,66]. These methods play a crucial role in analyzing complex datasets and aiding in the understanding and identification of patterns associated with ASD and other medical conditions. Furthermore, a variety of other machine learning approaches are utilized to analyze microarray data, highlighting the diverse array of methods employed in medical research [67,68].

Our approach involved employing customized machine learning models alongside the newly obtained mLRR values. Through validation, we demonstrated the capability to detect CNVs that remained undetected using existing LRR values, especially in detecting chromosomal disorders associated with CNVs. Furthermore, its accuracy was also confirmed through comparison with NGS data.

In typical microarray analysis, the log ratio is generally computed as the logarithm of the ratio of expression levels between two distinct samples. The concept of log ratios is extended and modified in the context of genotyping arrays, particularly when assessing copy number variations (CNVs). Notably, the log-R ratio demonstrates a correlation with gene expression levels [69]. Recent research determined gene expression levels by analyzing microarray images using log ratio [70,71].

We anticipate that our newly introduced mLRR value harbors extensive potential for versatile applications, extending its utility beyond genotyping to include the assessment of expression levels. This innovative metric holds promise for yielding more precise results compared to existing methods, thereby offering prospects for enhanced precision and comprehensive analyses.

Our study lacked continuous clinical observation to evaluate the long-term accuracy in predicting developmental disabilities among tested newborns. However, considering our success in minimizing the impact of genomic waves and obtaining accurate detection results, our method utilizing whole-genome SNP arrays could be considered one of the most effective approaches for screening chromosomal abnormalities in newborns.

5. Conclusions

Comprehensive CNV screening using new methods has the potential to significantly improve the screening process for patients with developmental disabilities and congenital malformations due to rare mutations and CNV-related chromosomal disorders. Validation of CNV detection provides strong evidence of its effectiveness in identifying a wide range of genetic abnormalities inherited or newly acquired during pregnancy. Therefore, the newly developed genotyping analysis presented in this study shows promise as a routine clinical screening tool for newborns and individuals at high risk of genetic diseases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics14010084/s1, Figure S1: Examples of next-generation sequencing validation of CNV-related chromosomal disorders; Table S1: CNV analysis results with LRR and mLRR of Coriell cell lines.; Table S2: The number of detected chromosomal disorders from the 16,046 Korean neonate samples.

Author Contributions: Conceptualization, H.-J.K. and M.-S.L.; methodology, B.-R.P. and M.-J.K. (Myeong-Ji Kim); software, Y.K.; validation, S.J. and Y.K.; formal analysis, H.-J.K. and Y.K.; data curation, H.-J.K. and M.-S.L.; writing—original draft preparation, C.J.G., H.-J.K., I.K.L., and M.-S.L.; writing—review and editing, M.-J.K. (Min-Jeong Kim) and M.-S.L.; visualization, J.P. and S.J.; supervision, M.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Eone-Diagnomics Genome Center Inc., Incheon, Republic of Korea (no grant number).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Diagnomics Inc. (IRB No. DR_CPLX_001, 3 December 2018).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: We would like to express our gratitude to the patients and physicians who participated in this genetic study, without whom this research would not have been possible. We also extend our thanks to the medical and technical staff, who facilitated the collection and processing of the samples used in this study. We are also grateful for the financial support from the EDGC that enabled us to carry out this study.

Conflicts of Interest: Author B.-R.P., C.J.K., H.J.K., I.K.L., J.W.P., M.-J.K. (Myeong-Ji Kim), S.J., and Y.K. are employees of the Eone-Diagnomics Genome Center (EDGC). Author M.-J.K. (Min-Jeong Kim) is an employee of the Diagnomics. Author M.-S.L. is a founder of Diagnomics and EDGC and has stock in both. Thus, the authors have no relevant financial or non-financial interests to disclose.

References

1. Feuk, L.; Carson, A.R.; Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **2006**, *7*, 85–97. [CrossRef] [PubMed]
2. Zarrei, M.; MacDonald, J.R.; Merico, D.; Scherer, S.W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **2015**, *16*, 172–183. [CrossRef] [PubMed]
3. Zaninović, L.; Bašković, M.; Ježek, D.; Bojanac, A.K. Validity and utility of non-invasive prenatal testing for copy number variations and microdeletions: A systematic review. *J. Clin. Med.* **2022**, *11*, 3350. [CrossRef] [PubMed]
4. Pös, O.; Radvanszky, J.; Styk, J.; Pös, Z.; Buglyó, G.; Kajsik, M.; Budis, J.; Nagy, B.; Szemes, T. Copy number nariation: Methods and clinical applications. *Appl. Sci.* **2021**, *11*, 819. [CrossRef]
5. Weiss, L.A.; Shen, Y.; Korn, J.M.; Arking, D.E.; Miller, D.T.; Fossdal, R.; Saemundsen, E.; Stefansson, H.; Ferreira, M.A.; Green, T.; et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **2008**, *358*, 667–675. [CrossRef]
6. Wang, L.; Wang, B.; Wu, C.; Wang, J.; Sun, M. Autism spectrum disorder: Neurodevelopmental risk factors, biological mechanism, and precision therapy. *Int. J. Mol. Sci.* **2023**, *24*, 1819. [CrossRef]
7. Pinto, D.; Pagnamenta, A.T.; Klei, L.; Anney, R.; Merico, D.; Regan, R.; Conroy, J.; Magalhaes, T.R.; Correia, C.; Abrahams, B.S.; et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **2010**, *466*, 368–372. [CrossRef]
8. Malhotra, D.; Sebat, J. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell* **2012**, *148*, 1223–1241. [CrossRef]
9. Sharp, A.J.; Mefford, H.C.; Li, K.; Baker, C.; Skinner, C.; E Stevenson, R.; Schroer, R.J.; Novara, F.; De Gregori, M.; Ciccone, R.; et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **2008**, *40*, 322–328. [CrossRef]
10. Gozzetti, A.; Le Beau, M.M. Fluorescence in situ hybridization: Uses and limitations. *Semin. Hematol.* **2000**, *37*, 320–333. [CrossRef]
11. Kozlowski, P.; Jasinska, A.J.; Kwiatkowski, D.J. New applications and developments in the use of multiplex ligation-dependent probe amplification. *Electrophoresis* **2008**, *29*, 4627–4636. [CrossRef] [PubMed]
12. Levy, B.; Wapner, R. Prenatal diagnosis by chromosomal microarray analysis. *Fertil. Steril.* **2018**, *109*, 201–212. [CrossRef] [PubMed]
13. Bravo-Valenzuela, N.J.; Peixoto, A.B.; Júnior, E.A. Prenatal diagnosis of congenital heart disease: A review of current knowledge. *Indian Heart J.* **2018**, *70*, 150–164. [CrossRef] [PubMed]
14. Dorsey, M.J.; Puck, J.M. Newborn screening for severe combined immunodeficiency in the United States: Lessons learned. *Immunol. Allergy Clin. N. Am.* **2019**, *39*, 1–11. [CrossRef]
15. 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 Bioinform.* **2013**, *14* (Suppl. S1), S1. [CrossRef]

16. Bharadwaj, S.; Dwivedi, V.D.; Kirtipal, N. Application of whole genome sequencing (WGS) approach against identification of foodborne bacteria. In *Microbial Genomics in Sustainable Agroecosystems*; Tripathi, V., Kumar, P., Tripathi, P., Kishore, A., Eds.; Springer: Singapore, 2019; Volume 1, pp. 131–148. [CrossRef]
17. Henderson, L.B.; Applegate, C.D.; Wohler, E.; Sheridan, M.B.; Hoover-Fong, J.; Batista, D.A. The impact of chromosomal microarray on clinical management: A retrospective analysis. *Anesth. Analg.* **2014**, *16*, 657–664. [CrossRef]
18. Miller, D.T.; Adam, M.P.; Aradhya, S.; Biesecker, L.G.; Brothman, A.R.; Carter, N.P.; Church, D.M.; Crolla, J.A.; Eichler, E.E.; Epstein, C.J.; et al. Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **2010**, *86*, 749–764. [CrossRef]
19. Werling, A.M.; Grünblatt, E.; Oneda, B.; Bobrowski, E.; Gundelfinger, R.; Taurines, R.; Romanos, M.; Rauch, A.; Walitza, S. High-resolution chromosomal microarray analysis for copy-number variations in high-functioning autism reveals large aberration typical for intellectual disability. *J. Neural Transm.* **2020**, *127*, 81–94. [CrossRef]
20. Hu, T.; Zhang, Z.; Wang, J.; Li, Q.; Zhu, H.; Lai, Y.; Wang, H.; Liu, S. Chromosomal aberrations in pediatric patients with developmental delay/intellectual disability: A single-center clinical investigation. *BioMed Res. Int.* **2019**, *2019*, 9352581. [CrossRef]
21. Wu, X.-L.; Li, R.; Fu, F.; Pan, M.; Han, J.; Yang, X.; Zhang, Y.-L.; Li, F.-T.; Liao, C. Chromosome microarray analysis in the investigation of children with congenital heart disease. *BMC Pediatr.* **2017**, *17*, 117. [CrossRef]
22. Tozzi, V.; Rosenberger, A.; Kube, D.; Bickeböller, H. Global, pathway and gene coverage of three Illumina arrays with respect to inflammatory and immune-related pathways. *Eur. J. Hum. Genet.* **2019**, *27*, 1716–1723. [CrossRef] [PubMed]
23. Wang, K.; Bucan, M. Copy number variation detection via high-density SNP genotyping. *Cold Spring Harb. Protoc.* **2008**, *2008*, pdb.top46. [CrossRef] [PubMed]
24. Diskin, S.J.; Li, M.; Hou, C.; Yang, S.; Glessner, J.; Hakonarson, H.; Bucan, M.; Maris, J.M.; Wang, K. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **2008**, *36*, e126. [CrossRef] [PubMed]
25. Ginsbach, P.; Chen, B.; Jiang, Y.; Engelter, S.T.; Grond-Ginsbach, C. Copy number studies in noisy samples. *BioTech* **2013**, *2*, 284–303. [CrossRef] [PubMed]
26. Marioni, J.C.; Thorne, N.P.; Valsesia, A.; Fitzgerald, T.; Redon, R.; Fiegler, H.; Andrews, T.D.; Stranger, B.E.; Lynch, A.G.; Dermitzakis, E.T.; et al. Breaking the waves: Improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.* **2007**, *8*, R228. [CrossRef]
27. Aboukhalil, A.; Bulyk, M.L. LOESS correction for length variation in gene set-based genomic sequence analysis. *Bioinformatics* **2012**, *28*, 1446–1454. [CrossRef]
28. Komura, D.; Shen, F.; Ishikawa, S.; Fitch, K.R.; Chen, W.; Zhang, J.; Liu, G.; Ihara, S.; Nakamura, H.; Hurles, M.E.; et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **2006**, *16*, 1575–1584. [CrossRef]
29. Hamerly, G.; Elkan, C. Learning the k in k-means. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 281–288.
30. Krishna, K.; Murty, M.N. Genetic k-means algorithm. *IEEE Trans. Syst. Man Cybern. Part B (Cybernetics)* **1999**, *29*, 433–439. [CrossRef]
31. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
32. Suguna, N.; Thanushkodi, K. An improved k-nearest neighbor classification using genetic algorithm. *Int. J. Comput. Sci. Issues* **2010**, *7*, 18–21.
33. Wang, K.; Li, M.; Hadley, D.; Liu, R.; Glessner, J.; Grant, S.F.; Hakonarson, H.; Bucan, M. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **2007**, *17*, 1665–1674. [CrossRef] [PubMed]
34. Colella, S.; Yau, C.; Taylor, J.M.; Mirza, G.; Butler, H.; Clouston, P.; Bassett, A.S.; Seller, A.; Holmes, C.C.; Ragoussis, J. QuantiSNP: An objective bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **2007**, *35*, 2013–2025. [CrossRef] [PubMed]
35. Toyama, M.; Vargas, L.; Ticlihuanca, S.; Quispe, A.M. Regional clustering and waves patterns due to COVID-19 by the index virus and the lambda/gamma, and delta/omicron SARS-CoV-2 variants in Peru. *Ann. Epidemiol.* **2022**, *6*, 74. [CrossRef]
36. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows—Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
37. Brancati, F.; Dallapiccola, B.; Valente, E.M. Joubert syndrome and related disorders. *Orphanet J. Rare Dis.* **2010**, *5*, 20. [CrossRef] [PubMed]
38. Riley, K.N.; Catalano, L.M.; Bernat, J.A.; Adams, S.D.; Martin, D.M.; Lalani, S.R.; Patel, A.; Burnside, R.D.; Innis, J.W.; Rudd, M.K. Recurrent deletions and duplications of chromosome 2q11.2 and 2q13 are associated with variable outcomes. *Am. J. Med. Genet. Part A* **2015**, *167*, 2664–2673. [CrossRef]
39. Cox, D.M.; Butler, M.G. The 15q11.2 BP1–BP2 microdeletion syndrome: A review. *Int. J. Mol. Sci.* **2015**, *16*, 4068–4082. [CrossRef]
40. Rafi, S.K.; Butler, M.G. The 15q11.2 BP1-BP2 microdeletion (*Burnside–Butler*) syndrome: In silico analyses of the four coding genes reveal functional associations with neurodevelopmental disorders. *Int. J. Mol. Sci.* **2020**, *21*, 3296. [CrossRef]
41. Fischer, M.; Klopocki, E. Atypical 22q11.2 microduplication with "typical" signs and overgrowth. *Cytogenet. Genome Res.* **2020**, *160*, 659–663. [CrossRef]

42. Wenger, T.L.; Miller, J.S.; DePolo, L.M.; de Marchena, A.B.; Clements, C.C.; Emanuel, B.S.; Zackai, E.H.; McDonald-McGinn, D.M.; Schultz, R.T. 22q11.2 duplication syndrome: Elevated rate of autism spectrum disorder and need for medical screening. *Mol. Autism* **2016**, *7*, 27. [CrossRef] [PubMed]
43. Armour, C.M.; E Bulman, D.; Jarinova, O.; Rogers, R.C.; Clarkson, K.B.; DuPont, B.R.; Dwivedi, A.; O Bartel, F.; McDonell, L.; Schwartz, C.E.; et al. 17p13.3 microduplications are associated with split-hand/foot malformation and long-bone deficiency (SHFLD). *Eur. J. Hum. Genet.* **2011**, *19*, 1144–1151. [CrossRef] [PubMed]
44. Petit, F.; Jourdain, A.; Andrieux, J.; Baujat, G.; Baumann, C.; Beneteau, C.; David, A.; Faivre, L.; Gaillard, D.; Gilbert-Dussardier, B.; et al. Split hand/foot malformation with long-bone deficiency and *BHLHA9* duplication: Report of 13 new families. *Clin. Genet.* **2013**, *85*, 464–469. [CrossRef] [PubMed]
45. Merikangas, A.K.; Corvin, A.P.; Gallagher, L. Copy-number variants in neurodevelopmental disorders: Promises and challenges. *Trends Genet.* **2009**, *25*, 536–544. [CrossRef] [PubMed]
46. Birnbaum, R.; Mahjani, B.; Loos, R.J.F.; Sharp, A.J. Clinical characterization of copy number variants associated with neurodevelopmental disorders in a large-scale multiancestry biobank. *JAMA Psychiatry* **2022**, *79*, 250–259. [CrossRef]
47. Lionel, A.C.; Crosbie, J.; Barbosa, N.; Goodale, T.; Thiruvahindrapuram, B.; Rickaby, J.; Gazzellone, M.; Carson, A.R.; Howe, J.L.; Wang, Z.; et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. *Sci. Transl. Med.* **2011**, *3*, 95ra75. [CrossRef]
48. Glessner, J.T.; Wang, K.; Cai, G.; Korvatska, O.; Kim, C.E.; Wood, S.; Zhang, H.; Estes, A.; Brune, C.W.; Bradfield, J.P.; et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **2009**, *459*, 569–573. [CrossRef]
49. Cooper, G.M.; Coe, B.P.; Girirajan, S.; A Rosenfeld, J.; Vu, T.H.; Baker, C.; Williams, C.; Stalker, H.; Hamid, R.; Hannig, V.; et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **2011**, *43*, 838–846. [CrossRef]
50. Yamasaki, M.; Makino, T.; Khor, S.-S.; Toyoda, H.; Miyagawa, T.; Liu, X.; Kuwabara, H.; Kano, Y.; Shimada, T.; Sugiyama, T.; et al. Sensitivity to gene dosage and gene expression affects genes with copy number variants observed among neuropsychiatric diseases. *BMC Med. Genom.* **2020**, *13*, 55. [CrossRef]
51. Kashevarova, A.A.; Drozdov, G.V.; Fedotov, D.A.; Lebedev, I.N. Pleiotropy of copy number variation in human genome. *Russ. J. Genet.* **2022**, *58*, 1180–1192. [CrossRef]
52. Park, J.; Jang, W.; Han, J.Y. Differing disease phenotypes of Duchenne muscular dystrophy and Moyamoya disease in female siblings of a Korean family. *Mol. Genet. Genom. Med.* **2019**, *7*, e862. [CrossRef] [PubMed]
53. Szigeti, K.; Lupski, J.R. Charcot-Marie-Tooth disease. *Eur. J. Hum. Genet.* **2009**, *17*, 703–710. [CrossRef] [PubMed]
54. Helland, W.A.; Lundervold, A.J.; Heimann, M.; Posserud, M.-B. Stable associations between behavioral problems and language impairments across childhood—The importance of pragmatic language problems. *Res. Dev. Disabil.* **2014**, *35*, 943–951. [CrossRef] [PubMed]
55. Mitrakos, A.; Kattamis, A.; Katsibardi, K.; Papadhimitriou, S.; Kitsiou-Tzeli, S.; Kanavakis, E.; Tzetis, M. High resolution Chromosomal Microarray Analysis (CMA) enhances the genetic profile of pediatric B-cell acute lymphoblastic leukemia patients. *Leuk. Res.* **2019**, *83*, 106177. [CrossRef] [PubMed]
56. Ronaghy, A.; Yang, R.K.; Khoury, J.D.; Kanagal-Shamanna, R. Clinical applications of chromosomal microarray testing in myeloid malignancies. *Curr. Hematol. Malign. Rep.* **2020**, *15*, 194–202. [CrossRef]
57. Ganesamoorthy, D.; Bruno, D.; McGillivray, G.; Norris, F.; White, S.; Adroub, S.; Amor, D.; Yeung, A.; Oertel, R.; Pertile, M.D.; et al. Meeting the challenge of interpreting high-resolution single nucleotide polymorphism array data in prenatal diagnosis: Does increased diagnostic power outweigh the dilemma of rare variants? *BJOG Int. J. Obstet. Gynaecol.* **2013**, *120*, 594–606. [CrossRef] [PubMed]
58. Zhao, S.; Jing, W.; Samuels, D.C.; Sheng, Q.; Shyr, Y.; Guo, Y. Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* **2018**, *19*, 765–775. [CrossRef] [PubMed]
59. Lü, Y.; Jiang, Y.; Zhou, X.; Hao, N.; Xu, C.; Guo, R.; Chang, J.; Li, M.; Zhang, H.; Zhou, J.; et al. Detection of mosaic absence of heterozygosity (AOH) using low-pass whole genome sequencing in prenatal diagnosis: A preliminary report. *Diagnostics* **2023**, *13*, 2895. [CrossRef]
60. Wang, H.; Dong, Z.; Zhang, R.; Chau, M.H.K.; Yang, Z.; Tsang, K.Y.C.; Wong, H.K.; Gui, B.; Meng, Z.; Xiao, K.; et al. Low-pass genome sequencing versus chromosomal microarray analysis: Implementation in prenatal diagnosis. *Anesth. Analg.* **2020**, *22*, 500–510. [CrossRef]
61. Chau, M.H.K.; Wang, H.; Lai, Y.; Zhang, Y.; Xu, F.; Tang, Y.; Wang, Y.; Chen, Z.; Leung, T.Y.; Chung, J.P.W.; et al. Low-pass genome sequencing: A validated method in clinical cytogenetics. *Hum. Genet.* **2020**, *139*, 1403–1415. [CrossRef]
62. Chaubey, A.; Shenoy, S.; Mathur, A.; Ma, Z.; Valencia, C.A.; Nallamilli, B.R.R.; Szekeres, E.; Stansberry, L.; Liu, R.; Hegde, M.R. Low-pass genome sequencing: Validation and diagnostic utility from 409 clinical cases of low-pass genome sequencing for the detection of copy number variants to replace constitutional microarray. *J. Mol. Diagn.* **2020**, *22*, 823–840. [CrossRef]
63. Singh, M.; Pujar, G.V.; Kumar, S.A.; Bhagyalalitha, M.; Akshatha, H.S.; Abuhaija, B.; Alsoud, A.R.; Abualigah, L.; Beeraka, N.M.; Gandomi, A.H. Evolution of machine learning in tuberculosis diagnosis: A review of deep learning-based medical applications. *Electronics* **2022**, *11*, 2634. [CrossRef]
64. Senescau, A.; Kempowsky, T.; Bernard, E.; Messier, S.; Besse, P.; Fabre, R.; François, J.M. Innovative DendrisChips® technology for a syndromic approach of in vitro diagnosis: Application to the respiratory infectious diseases. *Diagnostics* **2018**, *8*, 77. [CrossRef]

65. Kong, S.W.; Collins, C.D.; Shimizu-Motohashi, Y.; Holm, I.A.; Campbell, M.G.; Lee, I.-H.; Brewster, S.J.; Hanson, E.; Harris, H.K.; Lowe, K.R.; et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS ONE* **2012**, *7*, e49475. [CrossRef]
66. Krishnan, A.; Zhang, R.; Yao, V.; Theesfeld, C.L.; Wong, A.K.; Tadych, A.; Volfovsky, N.; Packer, A.; Lash, A.; Troyanskaya, O.G. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **2016**, *19*, 1454–1462. [CrossRef]
67. Cheng, L.; Wang, P.; Yang, S.; Yang, Y.; Zhang, Q.; Zhang, W.; Xiao, H.; Gao, H.; Zhang, Q. Identification of genes with a correlation between copy number and expression in gastric cancer. *BMC Med. Genom.* **2012**, *5*, 14. [CrossRef]
68. Nogueira, A.; Ferreira, A.; Figueiredo, M. A Machine learning pipeline for cancer detection on microarray data: The role of feature discretization and feature selection. *BioMedInformatics* **2023**, *3*, 585–604. [CrossRef]
69. Parisi, F.; Micsinai, M.; Strino, F.; Ariyan, S.; Narayan, D.; Bacchiocchi, A.; Cheng, E.; Xu, F.; Li, P.; Kluger, H.; et al. Integrated analysis of tumor samples sheds light on tumor heterogeneity. *Yale J. Biol. Med.* **2012**, *85*, 347–361.
70. Joseph, S.M.; Sathidevi, P.S. An automated cDNA microarray image analysis for the determination of gene expression ratios. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *20*, 136–150. [CrossRef]
71. Belean, B.; Gutt, R.; Costea, C.; Balacescu, O. Microarray image analysis: From image processing methods to gene expression levels estimation. *IEEE Access* **2020**, *8*, 159196–159205. [CrossRef]