genes

MDPI

Article

VarGenius-HZD Allows Accurate Detection of Rare Homozygous or Hemizygous Deletions in Targeted Sequencing Leveraging Breadth of Coverage

Francesco Musacchia [1,2,*], Marianthi Karali [1,3], Annalaura Torella [4], Steve Laurie [5], Valeria Policastro [6,7], Mariateresa Pizzo [1], Sergi Beltran [5,8,9], Giorgio Casari [1,10], Vincenzo Nigro [1,4] and Sandro Banfi [1,4]

1 Telethon Institute of Genetics and Medicine, 80078 Pozzuoli, Italy; karali@tigem.it (M.K.); pizzomar@tigem.it (M.P.); casari.giorgio@hsr.it (G.C.); vinnigro@gmail.com (V.N.); banfi@tigem.it (S.B.)
2 Center for Human Technologies, Istituto Italiano di Tecnologia, 16163 Genova, Italy
3 Eye Clinic, Multidisciplinary Department of Medical, Surgical and Dental Sciences, Università degli Studi della Campania 'Luigi Vanvitelli', 80138 Naples, Italy
4 Medical Genetics, Department of Precision Medicine, Università degli Studi della Campania 'Luigi Vanvitelli', 80138 Naples, Italy; annalaura.torella@gmail.com
5 CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain; steven.laurie@cnag.crg.eu (S.L.); sergi.beltran@cnag.crg.eu (S.B.)
6 Institute for Applied Mathematics "Mauro Picone" (IAC), National Research Council, 80131 Naples, Italy; valeria.policastro@gmail.com
7 Department of Environmental, Biological and Pharmaceutical Sciences and Technologies, Università degli Studi della Campania 'Luigi Vanvitelli', 81100 Caserta, Italy
8 Universitat Pompeu Fabra (UPF), 08017 Barcelona, Spain
9 Department of Genetics, Microbiology and Statistics, Universitat de Barcelona (UB), 08028 Barcelona, Spain
10 Neurogenomics Unit, Center for Genomics, Bioinformatics and Biostatistics, San Raffaele Scientific Institute, 20132 Milan, Italy
* Correspondence: francesco.musacchia@iit.it

Abstract: Homozygous deletions (HDs) may be the cause of rare diseases and cancer, and their discovery in targeted sequencing is a challenging task. Different tools have been developed to disentangle HD discovery but a sensitive caller is still lacking. We present VarGenius-HZD, a sensitive and scalable algorithm that leverages breadth-of-coverage for the detection of rare homozygous and hemizygous single-exon deletions (HDs). To assess its effectiveness, we detected both real and synthetic rare HDs in fifty exomes from the 1000 Genomes Project obtaining higher sensitivity in comparison with state-of-the-art algorithms that each missed at least one event. We then applied our tool on targeted sequencing data from patients with Inherited Retinal Dystrophies and solved five cases that still lacked a genetic diagnosis. We provide VarGenius-HZD either stand-alone or integrated within our recently developed software, enabling the automated selection of samples using the internal database. Hence, it could be extremely useful for both diagnostic and research purposes.

Keywords: copy-number variation; homozygous deletion; rare diseases

check for updates

Citation: Musacchia, F.; Karali, M.; Torella, A.; Laurie, S.; Policastro, V.; Pizzo, M.; Beltran, S.; Casari, G.; Nigro, V.; Banfi, S. VarGenius-HZD Allows Accurate Detection of Rare Homozygous or Hemizygous Deletions in Targeted Sequencing Leveraging Breadth of Coverage. *Genes* **2021**, *12*, 1979. https://doi.org/10.3390/genes12121979

Academic Editor: Yuval Itan

Received: 23 November 2021
Accepted: 8 December 2021
Published: 13 December 2021

1. Introduction

Next-generation sequencing (NGS) is commonly used to unveil genetic causes of diseases and whole-exome-sequencing (WES) has become one of the most commonly used diagnostic tools both in the clinic and in several programs investigating rare genetic diseases. Rare diseases collectively affect a significant fraction of the population (estimated to be about 4–5%) [1,2] with a resulting high impact on health-care costs and mortality rates. Currently, the standard protocol to investigate rare diseases includes multiple clinical diagnostics assays. Nonetheless, half of the cases still remain without a diagnosis [3–5]. One of the reasons for this is the limited knowledge of how to detect Copy Number Variation (CNV) from sequencing data. It is estimated that about 12% of the genome in the human

population is subject to copy number changes [6,7]. To detect CNVs, diagnostic laboratories often use multiplex ligation-dependent probe amplification (MLPA) and array comparative genomics hybridization analysis (ArrayCGH) prior to executing NGS-based analysis [8]. However, both methods have high ranges in resolution (from kilobases to megabases) and add complexity to the overall patient screening process. Whole-genome-sequencing (WGS) data are more even in coverage in comparison to WES because of the enrichment protocols used, making it more reliable for CNV calls. However, due to extensive use of WES in diagnostics, there is a need for reliable methods to infer CNVs from exome data as well [9–11]. Indeed, leveraging the sequencing outcome to detect CNVs offers potential advantages leading to increased diagnostic yield without increasing laboratory costs [10,12].

Several CNV-detection algorithms for WES data have been developed, all of which rely on the use of depth-of-coverage (DoC) from multiple samples to infer copy numbers [13–16]. Unfortunately, the CNV search is hampered by biases due to differences in capture protocol efficiency, the presence of GC-rich regions, and different coverage resolutions that influence DoC, among others [17–19]. Such heterogeneity complicates the downstream analysis of the detected events, leading to false positives [18,20–22] while compromising the ability to reliably detect CNVs when these span less than three exons [10,19,20,23,24]. Even though CNV detection could represent a valuable complementary way to analyze NGS data, the low concordance of detected events suggests that the algorithms designed so far are yet to be optimized [19,22,24,25]. Moreover, comparative works have demonstrated that these results are often difficult to replicate despite the high specificity and sensitivity declared [26]. One method to overcome these issues could be to generate a consensus of variants called by different algorithms [24]. However, to use any of these approaches, the user needs to prepare BAM files for unrelated samples sequenced with the same target writing ad hoc scripts, making such analyses difficult for those laboratories that do not have bioinformatics expertise. Therefore, the implementation of a fully automated CNV workflow along with different methods to investigate CNVs in WES data beyond the DoC strategies is of high importance for the scientific community.

Single-exon homozygous/hemizygous deletion (HD) detection methods, which compare normalized coverage values among samples produced with the same kits, already exist (e.g., Atlas-CNV, CoNVaDING, DECoN, and HMZDelFinder) [27–30]. While Atlas-CNV and CoNVaDING, as suggested by the authors, can only be used with high-coverage sequencing data (e.g., small targeted gene panels), HMZDelFinder and DECoN are ad hoc tools for exonic CNV detection. However, these tools are based on the assumption that data have a defined distribution and hence require intra- and inter-samples homogeneity [26].

To overcome these challenges, we developed a new algorithm for the detection of rare single-exon HDs that exploit breadth-of-coverage (BoC), and we named it VarGenius-HZD (where HZD stands for homozygous/hemizygous deletion detection). Additionally, we automated its execution along with that of ExomeDepth and XHMM within our recently developed software that we devised for variant detection analysis and management of samples, i.e., VarGenius [31]. This software is now able to automatically pick selected samples generated with the same target and to perform CNV, calling separately on autosomes and sex chromosomes and in parallel across different cores of a High-Performance Computing (HPC) system managed with a Portable Batch System (PBS) scheduler. The VarGenius-HZD algorithm is either integrated within VarGenius software, where it scales across HPC nodes, or is available as a stand-alone version that takes as input a list of manually selected BAM files and allows scaling across CPU cores.

We have validated our algorithm using 50 samples from the 1000 Genomes Project (1KGP) (https://www.internationalgenome.org/, accessed on 1 February 2021) for which both WGS and WES was present and in which we detected both existing and artificially inserted HDs. For these test cases we compared VarGenius-HZD results with those of HMZDelFinder, DECoN, and ExomeDepth, and our algorithm obtained the highest sensitivity. Furthermore, we applied VarGenius-HZD on targeted sequencing data from a cohort

of 188 individuals with Inherited Retinal Dystrophies (IRDs), resolving 5 out of 64 undiagnosed cases by identifying pathogenic HDs, which were then experimentally validated.

2. Results

2.1. BoC Can Be Used Along with DoC to Detect Rare HDs

2.1.1. Results Comparison with 1KGP Data

To compare the performances of different algorithms in detecting rare HDs, we applied VarGenius-HZD, ExomeDepth, HMZDelFinder, and DECoN to 50 samples from the 1KGP selecting only rare HD (see Methods). The resulting calls are in Table S1. One of the five HDs found in the 1KGP VCF file appeared to be a false positive (sample NA19473—position: chr9:107366951) when inspected in IGV, as it was covered by ~60 reads, and none of the tools detected it (Table S1 and Figure S1). Therefore, we considered only four real HDs. ExomeDepth detected only one event and two were filtered out (Tables S2 and S3); HMZDelFinder found only one HD (Tables S1 and S4); and DECoN could detect only one HD and one was filtered out (Table S5). VarGenius-HZD was able to detect the highest number of true positive events—three out of four—and one was filtered out (Tables S1 and S2). The HD in sample NA11919 (position: chr5:140222138) was called by all tools. For the sake of curiosity, we inspected in IGV the regions near single-nucleotide homozygous variants present in total in four samples: NA20798, NA19137, NA18504, and NA18950 (Table S6). Intriguingly, in sample NA20798, we found that the genes *CFHR3* and *CFHR1* were deleted. We could infer the call after inspection of the coverage of the nearby *CFHR* and *CFHR4* genes coverage and through comparison with control samples (Figure S2). This event was correctly detected by all tools but was not included in the 1KGP results (Tables S1–S5). Furthermore, a putative HD of gene *UGT2B28* in sample NA18504 was detected by VarGenius-HZD, ExomeDepth, filtered out by DECoN, and visually confirmed by comparing samples coverage in IGV (Figure S3 and Tables S1–S5).

To assess our results, we computed precision, recall, and specificity scores for all tools (none of the newly discovered variants was included in such calculations). VarGenius-HZD obtained higher recall, specificity, and precision when compared with ExomeDepth and DECoN (recall: ~25% vs. ~0.4%; specificity: ~2% vs. ~0.3%; precision: ~5% vs. ~0.3%). However, VarGenius-HZD results are comparable with those of HMZDelFinder: we obtained higher recall with our tool (75% vs. 25%) but higher specificity (10% vs. 2%) and precision (10% vs. 6%) with HMZDelFinder (Table 1). HMZDelFinder appeared to be the most precise tool returning very few events to inspect, reducing the number of false positives (FP) but at the cost of missing true positive (TP) events and losing sensitivity. The highest number of true positive calls was instead obtained by VarGenius-HZD. All tools found one additional putative TP HD. However, this variant should be experimentally confirmed, which was out of the scope of this study.

Table 1. Precision/Recall/Specificity obtained by the tools with the 50 samples from 1KGP dataset. TP (True positives), TN (True Negatives), FN (False Negatives), FP (False Positives).

Algorithm	TotalPutativeHZDel	TP	TN	FN	FP	Recall	Specificity	Precision	NewTP
ExomeDepth	274	1	1	3	273	0.25	0.0036	0.0036	2
VarGenius-HZD	51	3	1	1	48	0.75	0.0204	0.0588	2
HMZDelFinder	10	1	1	3	9	0.25	0.10	0.10	1
DECoN	267	1	1	3	266	0.25	0.0037	0.0037	1

2.1.2. Detection of Synthetic HDs

To evaluate our results with synthetic data, we simulated five deletions in five distinct samples of 1KGP selected randomly from our cohort of fifty. After running the tools using the fifty samples, to ameliorate positive and negative identification, we considered only HD calls for the selected samples after filtering, as described in Material and Methods. ExomeDepth and DECoN could not find any of the simulated events, HMZDelFinder missed only one, and VarGenius-HZD found them all (Table 2). Since all the detected

events were true positives, HMZDelFinder obtained 100% precision and 80% recall. On the contrary, VarGenius-HZD detected all the true-positive synthetic deletions inserted as well as one additional putative false positive, reaching 100% recall and 83% precision. We speculate that downstream CNV filtering is always needed and performed in several ways (e.g., visual inspection in IGV, gene panel selection, clinical phenotype, etc.); hence, for clinical diagnostics, a higher recall at the cost of a bit of downstream manual work would be preferable, as it leads to a higher number of positive genetic diagnoses.

Table 2. Precision/recall/specificity obtained by the tools used with the synthetic HD test.

Algorithm	TotalCalls	TotalFiltered	TP	TN	FP	FN	Recall	Specificity	Precision
HMZDelFinder	NA	4	4	0	0	1	0.8	0	1
VarGenius-HZD	4201	6	5	0	1	0	1	0	0.83
DECoN	38234	45	0	0	45	5	0	0	0
ExomeDepth	3949	3	0	0	3	5	0	0	0

2.2. Automated SNV/Indel and CNV Calling

We had previously developed VarGenius to execute SNV/Indel calling and annotation while exploiting the GATK BestPractices pipeline. VarGenius is able to scale across nodes of an HPC cluster with the PBS scheduling system and to construct a PostgreSQL database to store sample information enabling several queries for general genetic investigation.

We have embedded within VarGenius the execution of ExomeDepth, XHMM, and VarGenius-HZD for CNV analysis. Validation of results from ExomeDepth demonstrated high specificity and sensitivity for the detection of rare variants [16,21,32], while further studies suggested Conifer and XHMM for the low occurrence of false-positives, but with the disadvantage of a low detection sensitivity [13,14,18,21,25]. Thus, XHMM, Conifer and ExomeDepth are the tools best adapted to detect rare variants [33]). However, Conifer has not been updated for years, and it relies on the installation of old versions of R, making its integration within an automated pipeline difficult.

State-of-the-art CNV detection tools need as input several BAM files sequenced with the same enrichment kit, and different analyses should be performed for autosomes and sex chromosomes to avoid ploidy biases. We automated this process in VarGenius by querying for their numeric identifiers within the PostgreSQL database (see Methods Section and Figure 1). Several software are available to execute scalable CNV analysis for targeted sequencing data such as bcbio (https://github.com/bcbio/bcbio-nextgen accessed on 1 Februay 2021) and nf-Sarek [34], yet they require the user to manually select the BAM files to use. Other open-source tools (e.g., Hpexome, HemoMIPs and Swift/T) allow automated and scalable detection of SNV/Indel using multiple samples, but not CNVs (Table 3). Since VarGenius automates the complete workflow needed to execute CNV analysis, it can be a valuable resource for laboratories lacking bioinformatic expertise.

Table 3. Availability of SNV/CNV analysis automation in existing open-source software.

Software	SNV/Indel Calling	CNV Calling	Scalability	Automated Dataset Creation
bcbio	yes	yes	yes	no
Nf-sarek	yes	yes	yes	no
Hpexome	yes	no	yes	no
HemoMIPs	yes	no	yes	no
Swift/T	yes	no	yes	no

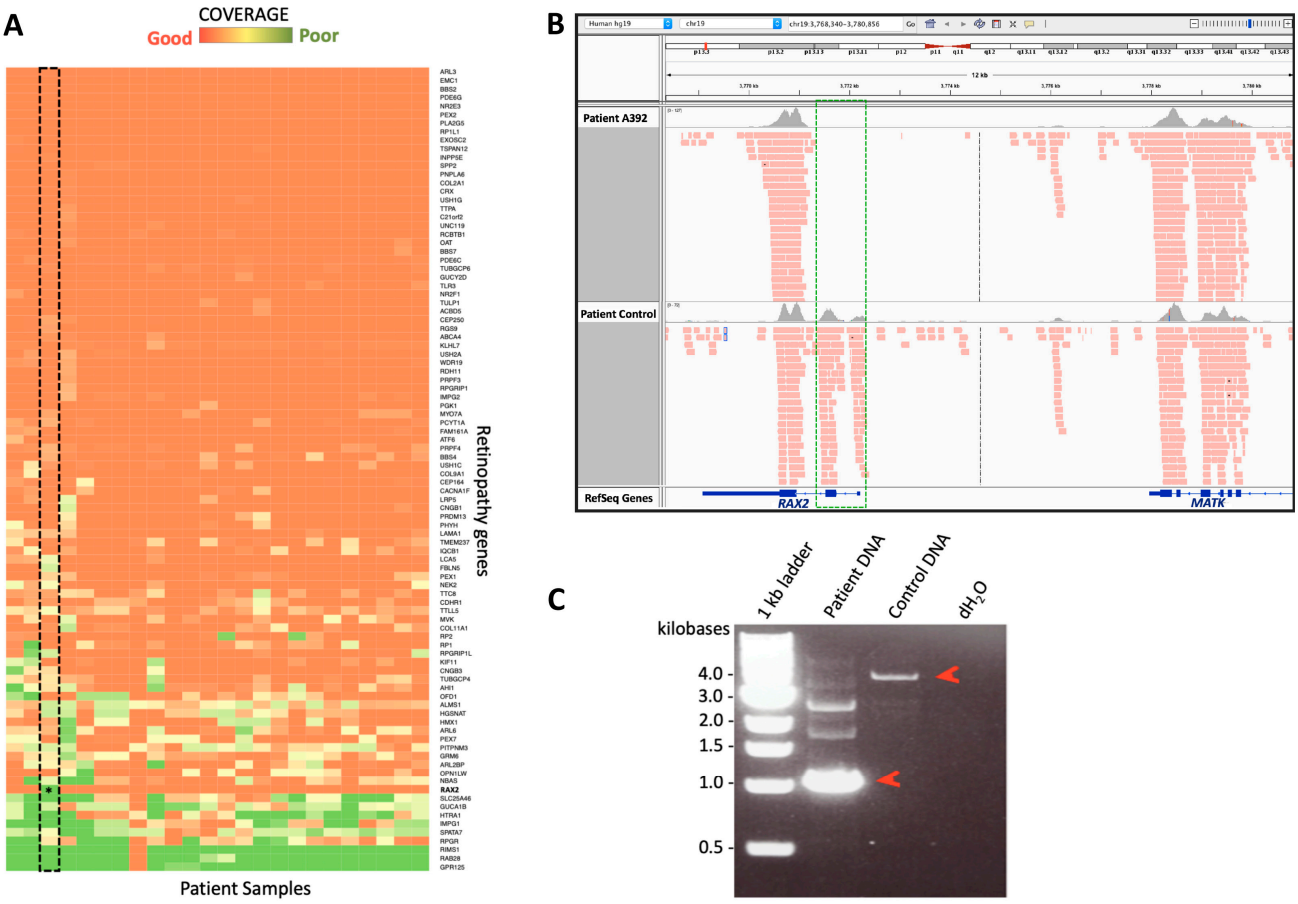


Figure 1. Experimental validation of the HD detected with our algorithm in the *RAX2* gene. (**A**) Coverage heatmap of retinopathy genes in the WES data of IRD patients. Patient samples are shown in the x axis and gene names on the y axis. The extent of coverage is plotted according to the reported color scale. The *RAX2* gene is well covered across all individuals but poorly covered in A392 (asterisk in the framed column). (**B**) IGV coverage tracks for the alignment file from patient A392 (upper track) and a control patient (lower track). The lack of reads spanning the exon 2 of *RAX2* in A392 (green box) suggested that the corresponding region was deleted in both alleles of the analyzed proband. (**C**) PCR amplification of the genomic region spanning the identified deletion in the proband's genomic DNA ('Patient DNA') and in a control DNA sample. The difference in size between the two amplicons (red arrowheads) indicates the presence of an extensive HD in the proband.

We have applied ExomeDepth, XHMM, VarGenius-HZD, and HMZDelFinder with default parameters to the 188 samples of the IRD cohort running different analyses for different enrichment kits. We first filtered only calls for the 64 unsolved cases using our panel of retinopathy genes and the thresholds suggested in the corresponding user guides (see Methods). ExomeDepth obtained the highest number of calls and, as a consequence, of false positives to filter followed by VarGenius-HZD and XHMM. After filtering, we have selected candidate HDs to inspect in IGV according to disease gene and its association with the patient phenotype (in Table S7, the sum of VisButNotFit and FitButNotVis) and obtained eight HDs from ExomeDepth, five from XHMM, ten from VarGenius-HZD, and from HMZDelFinder. Only six events in total passed all the evaluation filters and five of them were confirmed through PCR (Table 4). VarGenius-HZD identified all these events and HMZDelFinder found four out of five. We went back to unfiltered results from the other tools to see at which stage they were lost: XHMM did not find any of these HDs; ExomeDepth detected all of them, but they were initially filtered out because of their low BF value (<5). Indeed, reducing the BF threshold in ExomeDepth increases the number of calls to assess. The excellent performance of VarGenius-HZD was particularly striking as it obtained the highest number of true positives at a low cost of variants to manually inspect.

Table 4. VRCIRD cases resolved by detection of a HD.

Sample	Gene	Region	XHMM	ExomeDepth (BF)	VarGenius-HZD	HMZDelFinder
ID_A739	*RAX2*	19:3772155-3772224	NO	7.4	YES	YES
CREv1_A392	*RAX2*	19:3771519-3772224	NO	11	YES	YES
CREv1_A348	*RP2*	X:46719422-46719537	NO	7.8	YES	YES
CREv1_ARRP129	*RP2*	X:46719424-46719537	NO	9	YES	YES
ID_A860	*RPGR*	X:38186587-38186793	NO	9	YES	NO

Experimental Validation of Detected HDs

The five identified HDs were validated by PCR and/or Sanger sequencing in the following patients: first, an HD of the first two exons of the *RAX2* gene [35] was identified in two reportedly unrelated subjects (a female and a male), who had a clinical diagnosis of autosomal recessive retinitis pigmentosa (Figure 1). Both patients were born in a small village of ~1000 inhabitants in Campania (Italy). We believe that a founder effect within this small isolated community could account for the fact that they both carried the same deletion. Second, a hemizygous deletion in the *RP2* gene was identified in two young male subjects (a 17- and a 21-year-old) who were first-degree cousins and diagnosed with X-linked early-onset retinitis pigmentosa (OMIM #300757, http://www.omim.org/entry/, accessed on 1 October 2021). Finally, the fifth case was a male patient carrying a hemizygous deletion of exon 1 of the *RPGR* gene, and his clinical presentation was consistent with a diagnosis of an X-linked Retinitis Pigmentosa associated with mutations in this gene (OMIM #312610, http://www.omim.org/entry/, accessed on 1 October 2021).

3. Methods

3.1. NGS Procedures

The 188 subjects considered in this study were selected for targeted sequencing after being assessed at the Referral Centre for Inherited Retinal Dystrophies of the Eye Clinic at Università degli Studi della Campania 'Luigi Vanvitelli' (VRCIRD) (Table 5). Peripheral blood samples were collected upon written informed consent of the patient or their parents/legal guardians (for minors). All procedures adhered to the tenets of the Declaration of Helsinki and were approved by the Ethics Board of Fondazione Telethon and Università degli Studi della Campania 'Luigi Vanvitelli'.

Table 5. Summary of samples used from the VRCIRD cohort.

Platform	CREv1	CCP	ID	Total	Solved Cases	Examined
NextSeq500	14	51	123	188	124	64

DNA samples from peripheral blood were processed using the Illumina NextSeq500 (Illumina Inc., San Diego, CA, USA). Three different enrichment kits (Agilent Technologies) were used for library preparation. In particular, 123 samples were sequenced using the Agilent ClearSeq Inherited Disease (ID) and 51 samples using the SureSelect Clinical Constitutional Panel (CCP), and 14 samples were prepared using the SureSelect Clinical Research Exome version 1 (CREv1) (Table 5). BCL files were processed using Illumina bcl2fastq. Raw fastq files were processed using our previously developed software [31].

3.1.1. Calling Homozygous Deletions Leveraging BoC

The VarGenius-HZD algorithm was written in PERL and R programming languages and needs the execution of three steps: sample selection, pre-processing, and rare HD detection (Figure 2).

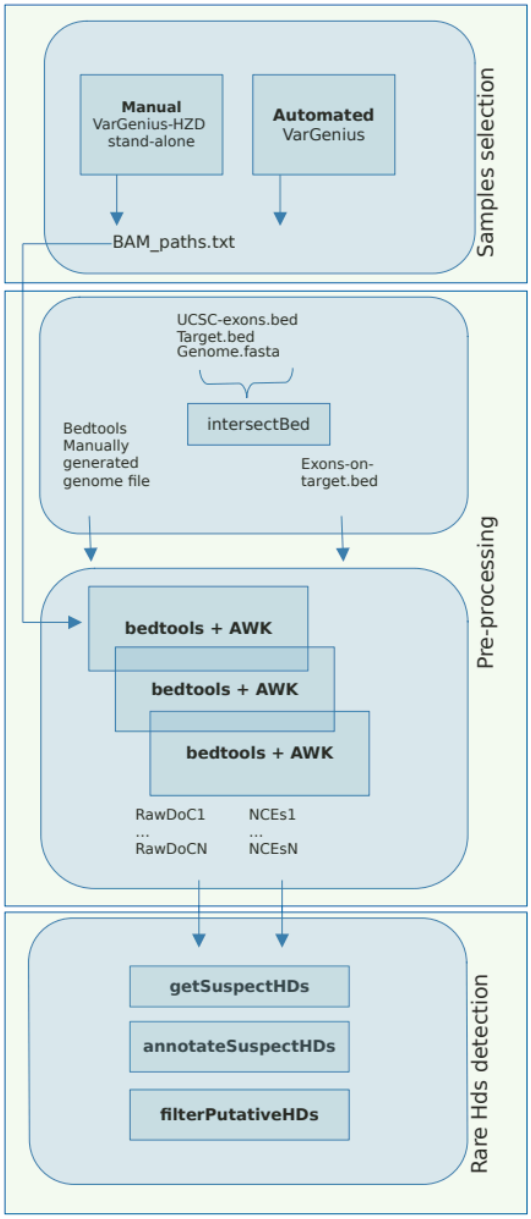


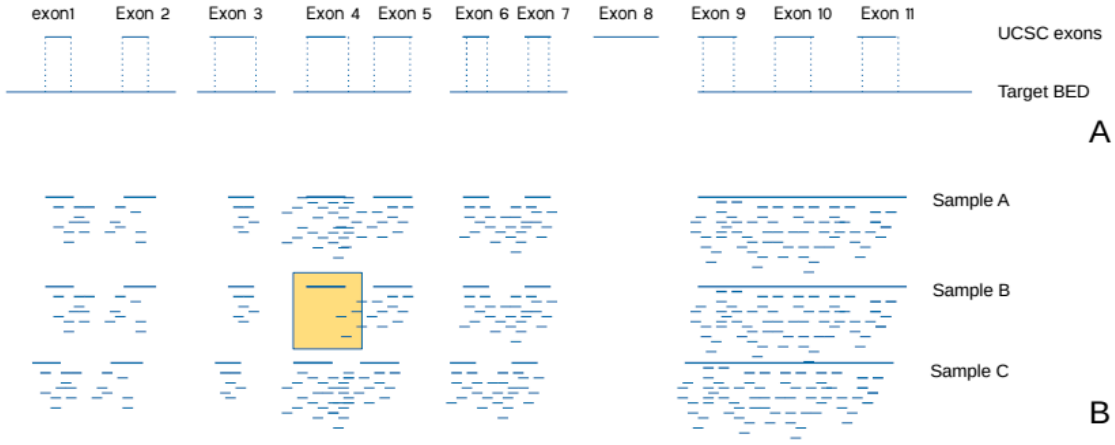
Figure 2. VarGenius-HZD workflow. The workflow of our algorithm consists of three steps: 1. sample selection, which is automated in VarGenius software and manual in the stand-alone version; 2. pre-processing, which includes the generation of NCEs files and raw DoC information; 3. rare-HD detection step, which involves the calculation of NCE frequencies, the detection of putative HDs, and the annotation of such regions for variant prioritization.

The samples selection step is automated within the VarGenius software by querying the PostgreSQL database for unrelated samples sequenced with the same target, while for the stand-alone version, this step is manual; i.e., the user must provide a file with the paths to the BAM files and the BED file for the target sequenced.

The pre-processing step aims to generate an exons-on-target-intervals file. To this end, BED files for genes and exons were downloaded from the University of California Santa Cruz (UCSC) platform (https://genome.ucsc.edu/cgi-bin/hgtables, accessed on 1 February 2021) and included within the package. We selected UCSC genes as the track, Hg19 as the genome assembly, start and end of exons/genes, and BED as output format. The BED file is then intersected with the target file using the bedtools intersect. This procedure is executed only once per target.

Furthermore, each BAM file undergoes bedtools coverage to compute the BoC and the DoC of the previously generated exonic intervals using this command: bedtools coverage -a input.bam -b exons_on_target.bed g ucsc.hg19.genomefile -sorted. As suggested in the bedtools guidelines, we used the -sorted parameter and a genome file in input to accelerate such computation. The *genome file* was generated with the following commands: samtools faidx ucsc.hg19.fa; cut -f1,2 ucsc.hg19.fa.fai > ucsc.hg19.genomefile (https://bedtools.readthedocs.io/en/latest/content/tools/coverage.html, accessed on 1 February 2021). The resulting output, containing the BoC, is filtered to select only exons with BoC < 0.2 (<20% of the exon covered) and annotated with the UCSC genes for downstream analyses. This procedure generated two tab separated files (TSV) for each sample: one containing putative non-covered exons (NCEs) and another containing raw DoC for the exons-on-target.

The third step aims at rare HDs detection using the two TSV files previously produced. First, all NCEs files are loaded within a unique array. Second, putative rare HDs are obtained by computing their frequency and selecting those where it is lower or equal to 2 (this parameter can be customized). Third, the exonic raw DoC of parents (whenever available), proband, and average across all samples are added. Once complete, the annotated table of putative rare HDs is provided, and manual filtering of relevant calls based on the difference in coverage between the proband and her/his parents and between the proband and the overall dataset can be performed downstream (Figure 3).



sample	chr	start	end	BoC	DoC	FDoC	MDoC	AvgDoC	Gene
SampleB	chr5	13578906	13579383	0.01	2	124	167	256	ALM
SampleC	chr12	7899030	7899568	0	0	12	16	15	HGA

C

Figure 3. VarGenius-HZD algorithm and results illustration. VarGenius-HZD leverages BoC along with DoC, which is used as follows: (**A**) the target BED used for the sequencing is intersected with UCSC exon intervals to obtain an exon-on-target file, which is used to compute the BoC and DoC exploiting bedtools coverage. (**B**) NCEs for each sample are counted, and only those with frequency lower or equal to 2 are retained as putative HDs (e.g., exon 4 in (**B**)). (**C**) The tabular output contains statistics for putative HDs: chromosome, start and end, the BoC for the subject sample, the DoC for the parents (FDoC and MDoC), and average exon DoC for the overall dataset.

3.1.2. KGP WES Dataset

We selected 50 samples for which genome wide CNV calls were available and from the 1KGP data as in [20] from (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/, accessed on 1 October 2021) and their consensus target BED file from (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/exome_pull_down/20120518.analysis_exome_targets.consensus.annotation.bed, accessed on 1 October 2021). Whole genome Variant Calling Format (VCF) files with genotypes were downloaded from

(http://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/, accessed on 1 October 2021). First the VCF file was intersected with the consensus BED files using bedtools intersect. Then, we filtered the VCF for rare HDs selecting those calls with sv_type = del within the VCF info field and where only one sample had gt = 1|1. BAM files for the 50 samples were used with ExomeDepth, HMZDelFinder, DECoN, and VarGenius-HZD. Results from the four tools are in Tables S3–S5. HDs called were further filtered with the following approaches: for ExomeDepth and DECoN, we retained only the calls with reads.ratio < 0.001; HMZDelFinder provided a filtered output, and hence we did not apply any filter; VarGenius-HZD produced a filtered output as well, and we picked those calls where the average DoC was greater than 50.

3.1.3. Synthetic Homozygous Deletion Detection

The 50 samples from 1KGP were also used to conduct a test using simulated deletions generated with bedtools. We inserted 5 HDs in 5 distinct samples (NA06989, NA07347, NA12058, NA12748, NA12830), choosing regions that we have visually inspected and had sufficient coverage surrounding the chosen HDs (>20x) across the overall dataset (Table 6). The commands used to generate such deletions were: bedtools intersect -a sample.bam -b deletion_i.bed -v > sample_deleted.bam; samtools sort sample_deleted.bam > sample_deleted_sort.bam; samtools index sample_deleted_sort.bam.

Table 6. The 5 simulated HDs inserted in samples of 1KGP dataset.

Sample	Chr	Start	End
NA06989	21	48063447	48063551
NA07347	21	27326904	27327003
NA12058	21	35091133	35091161
NA12748	21	10906904	10907040
NA12830	21	40188932	40189015

HDs were filtered with the following methods only for the 5 samples: DECoN and ExomeDepth: reads.ratio <= 0.001. VarGenius-HZD: raw average DoC >50; HMZDelFinder: no filtering.

3.1.4. Recall, Precision, and Specificity Scores

To compare results from different tools, we calculated recall, precision, and specificity scores with the following formula: Recall = TP/(TP + FP); Specificity = TN/(TN + FP); Precision: TP/(TP + FP).

In this formula, true positives (TP) are HDs called and present in the 1KGP VCF; true negatives (TN) are HDs not called that are not present; false positives (FP) are HDs called that are not present; false negatives (FN) are HDs not called that are instead in the 1KGP VCF.

3.2. Analysis of the VRCIRD Cohort

3.2.1. Automated CNV Detection Workflow

All samples of the VRCIRD cohort were subject to SNVs/Indels and CNV calling. We used the GATK3.8 BestPractices with default parameters [36]. Alignment used BWA [37], PCR duplicates were marked with PICARD MarkDuplicates (http://broadinstitute.github.io/picard/ accessed on 1 May 2021), and further BAM pre-processing was performed using BaseRecalibrator prior to variant calling with HaplotypeCaller. We performed GATK hard filtering with VariantFiltration, and Annovar was used for the annotation [38]. Further parsing of the VCF file and annotation table was performed to provide an XLS tabular output to the physician. Sample information (such as sample and analysis name, gender,

kinship, and target used) provided within the sample sheet were parsed and stored within the PostgreSQL database.

CNV detection was performed with XHMM, ExomeDepth, and VarGenius-HZD, as follows: the software executed a query to the PostgreSQL database to obtain samples which are sequenced with the same target. VarGenius automatically picked the BAM files to provide as input for the tools from its results folders using sample identifiers, kinship, gender, and the target used. XHMM and ExomeDepth were executed following the author's guidelines and with default parameters. Autosomes and sex chromosomes were analyzed separately to avoid gender biases. Once finished, all the CNVs called (herein "calls") were annotated using AnnotSV [39] (Figure 4). Causative SNVs/Indels for all subjects were investigated, and a subset of 124 cases received a diagnosis. The remaining 64 cases were subjected to CNV prioritization. However, in this work, we specifically discuss only HDs.

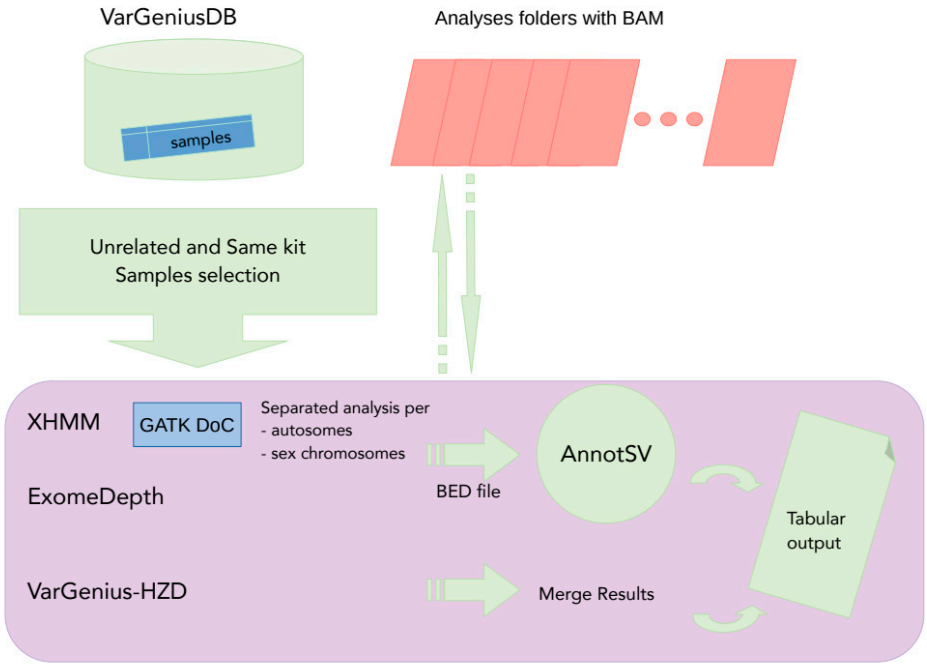


Figure 4. Flowchart of CNV detection and annotation pipeline in VarGenius. This is performed using XHMM, ExomeDepth, and the VarGenius-HZD algorithm. Several unrelated samples must be used for such analyses; thus VarGenius collects sample identifiers from the database querying for samples sequenced with the same target and considering the kinship. XHMM requires the use of GATK DepthOfCoverage with specific parameters. This is called for all samples parallelizing the execution within the cluster. Once all tools produced their calls, results are merged within a unique tabular output and are annotated using AnnotSV.

3.2.2. Homozygous Deletions Filtering for Patients

Manual inspection in Integrative Genomics Viewer (IGV) of detected HDs was performed for 64 undiagnosed cases, which could not be solved with a causative SNV/Indel (Table 5). To increase the probability of diagnosis through known disease genes, we filtered the resulting calls using our internal panel of known retinopathy genes (data not published). Selection of resulting events to inspect in IGV as a first-tier validation was manually performed keeping into account AnnotSV annotation (OMIM, Decipher) and different scores depending on the tool. For ExomeDepth, we considered the Bayes factor (BF > 10); for XHMM, we used the mean normalized DoC (MEAN_RD); for VarGenius-HZD, we selected calls with average raw DoC > 30. BAM files for the proband and her/his parents (whenever available) or for different probands were loaded as controls.

3.2.3. Polymerase Chain Reaction (PCR) and Deletion Breakpoint Analysis

For the amplification of deleted coding exons, PCR on genomic DNA was performed using Taq polymerase according to standard protocols. Breakpoint analysis was done only for the *RAX2* deletion. To this aim, long-range PCR was performed using the High-Fidelity LA Taq DNA Polymerase (Takara) according to the manufacturer's recommendations and the oligonucleotide primers RAX2_del1_F: 5′-TGTTACCCACACCATTCTCTGC-3′ and RAX2_del1_R: 5′-CCCTCTCCTTTCCATCTCTAG-3′. Amplicons spanning the junction of the deletion extremities were Sanger sequenced and aligned to the reference genome (hg19) using the UCSC Genome Browser (http://genome.ucsc.edu/, accessed on 1 October 2021) to determine the breakpoints at the nucleotide level.

4. Discussion

HDs often lead to loss of function with pathogenic roles both in Mendelian diseases and cancer [40–43]. Indeed, a significant percentage of human Mendelian diseases is reported to be caused by molecular disruption within exons [6,7]. NGS-based approaches became cheap during the last decade, allowing diagnostic laboratories to use targeted sequencing [44]. Nonetheless, the investigation of CNVs in WES is still challenging for several reasons mostly due to uneven coverage and due to enrichment kits and regions of the genome difficult to sequence [18,45,46]. State-of-the-art tools require as input several samples for such comparison that should be unrelated and sequenced with the same target [13,16,20]. Yet, comparative works have demonstrated a high number of false positives and hence alternative CNV detection strategies and filtering methods are needed [18,20,22].

The goal of this work was to explore different solutions for HD discovery in targeted sequencing and to automate the overall workflow. We developed VarGenius-HZD, which searches for HDs within the single sample and leverages multi-sample information to corroborate such calls, and we integrated it within our recently developed VarGenius. CNV detection is still a challenging task, and we think that currently only highly trained bioinformaticians might disentangle the intrinsic difficulty of detection of such types of variation to understand the underlying complexities and cavities, especially for clinical practice. However, being able to automate CNV analysis and to reduce false positives for HD detection and, as a consequence, the number of events to manually inspect out of the tool could increase the availability of human-readable results and, hopefully, of genetic diagnoses for those laboratories lacking bioinformatics expertise. To make VarGenius-HZD useful for researchers exploiting other software for variant calling, we also developed a stand-alone VarGenius-HZD; in this version the user provides the list of full paths to the BAM files and the target file. One limitation of the stand-alone tool (compared to the complete VarGenius software) is that it cannot provide parents' coverage as annotation but only on average across all samples used.

To compare our algorithm with state-of-the-art methods, we applied VarGenius-HZD, ExomeDepth, HMZDelFinder, and DECoN to 50 samples from 1KGP. The highest number of TPs was achieved only with our algorithm; hence, it is more sensitive than state-of-the-art tools, demonstrating that BoC can be effectively used to detect such variants. Furthermore, our tool was able to correctly detect all the synthetic HDs that we inserted within randomly chosen samples in the same dataset, achieving a sensitivity of 100%, while the only comparable results were obtained with HMZDelFinder with a sensitivity of 80%. ExomeDepth and DECoN were not able to detect any of the simulated HDs. Our results are in agreement with other comparative studies, which describe ExomeDepth's ability to discover long CNVs covering large chromosomal regions while missing events that affect less than three exons. However, DECoN, which is based on ExomeDepth, provided similar results. We speculate that a higher number of TPs and thus higher sensitivity rather than precision would be preferable for clinical diagnosis at a cost of filtering few additional CNVs during downstream prioritization.

We then assessed the performance of VarGenius-HZD in a clinical context using targeted sequencing data from a cohort of unsolved IRD patients. Analysis of CNVs using

ExomeDepth and XHMM with such data turned out to be challenging. These tools detect hundreds of events, and filtering FPs was a tough task. We observed several false positives detected by ExomeDepth and XHMM, in agreement with current studies showing that state-of-the-art CNV-calling algorithms are influenced by different instrument outcomes and low-coverage samples, possibly due to the high number of off-target bases, duplicates, and low base quality. We speculated that CNV callers should deal with such issues, and, to reduce the false discovery rate, as a pre-processing step, it could be useful to remove outlier samples which have a high number of calls (e.g., >2 standard deviation).

After filtering, we could confirm, through experimental assays, five pathogenic HDs. Only VarGenius-HZD was able to detect all of them. In summary, XHMM lost all of them; ExomeDepth detected all except one but provided very low BF score, and hence they were initially excluded; HMZDelFinder detected all except one. One of the called HDs was instrumental in defining a new association of biallelic variants in the *RAX2* gene with autosomal recessive Retinitis pigmentosa [35].

5. Conclusions

In summary, the use of targeted sequencing data for CNV discovery, as well as the automation of this process (which currently requires programming skills) are of great importance. Here, we report an algorithm that could be useful to identify rare HDs, demonstrating that BoC is a valuable feature for their detection. Given the extensive use of targeted sequencing as a first-tier method for molecular genetic diagnosis, our work has a great importance for research and clinical practice. Our tool is available under GNU General Public License, version 3 at: https://github.com/frankMusacchia/VarGenius-HZD (accessed on 6 December 2021)

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12121979/s1, Table S1: HDs called in 1KGP data, Table S3: 1KGP HDs in VarGenius-HZD result, Table S2: 1KGP HDs in ExomeDepth result, Table S4: 1KGP HDs in HMZDelFinder result, Table S5: 1KGP HDs in DECoN result, Table S6: Single nucleotide homozygous variants found in 1KGP samples, Table S7: Summary statistics of HDs found and filtered in the VRCIRD cohort, Figure S1: IGV screenshot of a false positive detected in 1KGP data, Figure S2: Deletion of genes *CFHR1* and *CFHR3* in sample NA20798 of 1KGP data, Figure S3: HD of gene *UGT2B28* in NA18504 1KGP data.

Author Contributions: F.M. designed the algorithm and developed the tool; M.K. and A.T. coordinated the validation of CNVs and recruited the patients whose WES was used; V.P. developed the tool; M.P. performed PCR/RealTime validations for all detected events in patients; S.L. and S.B. (Sergi Beltran). substantially contributed with the knowledge of methods for CNV detection, analysis, and prioritization for rare diseases; G.C., V.N. and S.B. (Sandro Banfi) led the project and the patients' enrollment. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Fondazione Telethon (Grant No. GSP15001) and by the University of Campania 'Luigi Vanvitelli' under "VALERE: VAnviteLli pEr la RicErca" (project DisHetGeD).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical approval was given by the ethics committee of the Università degli Studi della Campania "Luigi Vanvitelli" (for adult protocol n. 8189/2015; for pediatric subjects protocol n. 500/2017).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study or their parents/legal guardians (for minors).

Data Availability Statement: 1000 Genomes Project data used for algorithms comparisons were downloaded from https://www.internationalgenome.org (accessed on 1 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boycott, K.M.; Vanstone, M.R.; Bulman, D.E.; MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* **2013**, *14*, 681–691. [CrossRef] [PubMed]
2. Wright, C.F.; FitzPatrick, D.R.; Firth, H.V. Paediatric genomics: Diagnosing rare disease in children. *Nat. Rev. Genet.* **2018**, *19*, 253–268. [CrossRef] [PubMed]
3. Demos, M.; Guella, I.; DeGuzman, C.; McKenzie, M.B.; Buerki, S.E.; Evans, D.M.; Toyota, E.B.; Boelman, C.; Huh, L.L.; Datta, A.; et al. Diagnostic Yield and Treatment Impact of Targeted Exome Sequencing in Early-Onset Epilepsy. *Front. Neurol.* **2019**, *10*, 434. [CrossRef]
4. Hartley, T.; Lemire, G.; Kernohan, K.D.; Howley, H.E.; Adams, D.R.; Boycott, K.M. New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases. *Annu. Rev. Genom. Hum. Genet.* **2020**, *21*, 351–372. [CrossRef] [PubMed]
5. Shashi, V.; McConkie-Rosell, A.; Rosell, B.; Schoch, K.; Vellore, K.; McDonald, M.; Jiang, Y.-H.; Xie, P.; Need, A.; Goldstein, D.B.; et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2014**, *16*, 176–182. [CrossRef]
6. Redon, R.; Ishikawa, S.; Fitch, K.R.; Feuk, L.; Perry, G.H.; Andrews, T.D.; Fiegler, H.; Shapero, M.H.; Carson, A.R.; Chen, W.; et al. Global variation in copy number in the human genome. *Nature* **2006**, *444*, 444–454. [CrossRef] [PubMed]
7. Yuan, B.; Wang, L.; Liu, P.; Shaw, C.; Dai, H.; Cooper, L.; Zhu, W.; Anderson, S.A.; Meng, L.; Wang, X.; et al. CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. *Genet. Med.* **2020**, *22*, 1633–1641. [CrossRef] [PubMed]
8. Vasson, A.; Leroux, C.; Orhant, L.; Boimard, M.; Toussaint, A.; Leroy, C.; Commere, V.; Ghiotti, T.; Deburgrave, N.; Saillour, Y.; et al. Custom oligonucleotide array-based CGH: A reliable diagnostic tool for detection of exonic copy-number changes in multiple targeted genes. *Eur. J. Hum. Genet.* **2013**, *21*, 977–987. [CrossRef] [PubMed]
9. Monroe, G.R.; Frederix, G.W.; Savelberg, S.M.; de Vries, T.I.; Duran, K.J.; van der Smagt, J.J.; Terhal, P.A.; van Hasselt, P.M.; Kroes, H.Y.; Verhoeven-Duif, N.M.; et al. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2016**, *18*, 949–956. [CrossRef] [PubMed]
10. Lelieveld, S.H.; Veltman, J.A.; Gilissen, C. Novel bioinformatic developments for exome sequencing. *Hum. Genet.* **2016**, *135*, 603. [CrossRef]
11. Stark, Z.; Tan, T.Y.; Chong, B.; Brett, G.R.; Yap, P.; Walsh, M.; Yeung, A.; Peters, H.; Mordaunt, D.; Cowie, S.; et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet. Med.* **2016**, *18*, 1090–1096. [CrossRef] [PubMed]
12. Pfundt, R.; del Rosario, M.; Vissers, L.E.; Kwint, M.P.; Janssen, I.M.; de Leeuw, N.; Yntema, H.G.; Nelen, M.R.; Lugtenberg, D.; Kamsteeg, E.-J.; et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.* **2017**, *19*, 667–675. [CrossRef] [PubMed]
13. Fromer, M.; Purcell, S.M. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr. Protoc. Hum. Genet.* **2014**, *81*, 7.23.1–7.23.21. [CrossRef]
14. Krumm, N.; Sudmant, P.H.; Ko, A.; O'Roak, B.J.; Malig, M.; Coe, B.P.; NHLBI Exome Sequencing Project; Quinlan, A.R.; Nickerson, D.A.; Eichler, E.E. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **2012**, *22*, 1525–1532. [CrossRef]
15. Li, J.; Lupat, R.; Amarasinghe, K.C.; Thompson, E.R.; Doyle, M.A.; Ryland, G.L.; Tothill, R.W.; Halgamuge, S.K.; Campbell, I.G.; Gorringe, K.L. CONTRA: Copy number analysis for targeted resequencing. *Bioinformatics* **2012**, *28*, 1307–1313. [CrossRef]
16. Plagnol, V.; Curtis, J.; Epstein, M.; Mok, K.Y.; Stebbings, E.; Grigoriadou, S.; Wood, N.W.; Hambleton, S.; Burns, S.O.; Thrasher, A.J.; et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **2012**, *28*, 2747–2754. [CrossRef]
17. Auer, P.L.; Reiner, A.P.; Wang, G.; Kang, H.M.; Abecasis, G.R.; Altshuler, D.; Bamshad, M.J.; Nickerson, D.A.; Tracy, R.P.; Rich, S.S.; et al. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *Am. J. Hum. Genet.* **2016**, *99*, 791. [CrossRef]
18. Guo, Y.; Sheng, Q.; Samuels, D.C.; Lehmann, B.; Bauer, J.A.; Pietenpol, J.; Shyr, Y. Comparative Study of Exome Copy Number Variation Estimation Tools Using Array Comparative Genomic Hybridization as Control. *BioMed Res. Int.* **2013**, *2013*, 1–7. [CrossRef]
19. Hong, C.S.; Singh, L.N.; Mullikin, J.C.; Biesecker, L.G. Assessing the reproducibility of exome copy number variations predictions. *Genome Med.* **2016**, *8*, 82. [CrossRef] [PubMed]
20. Feng, Y.; Chen, D.; Wang, G.-L.; Zhang, V.W.; Wong, L.-J.C. Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing. *Genet. Med.* **2015**, *17*, 99. [CrossRef]
21. Samarakoon, P.S.; Sorte, H.S.; Stray-Pedersen, A.; Rødningen, O.K.; Rognes, T.; Lyle, R. cnvScan: A CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data. *BMC Genom.* **2016**, *17*, 51. [CrossRef]
22. Tan, R.; Wang, Y.; Kleinstein, S.E.; Liu, Y.; Zhu, X.; Guo, H.; Jiang, Q.; Allen, A.S.; Zhu, M. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum. Mutat.* **2014**, *35*, 899–907. [CrossRef]
23. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef]

24. Moreno-Cabrera, J.M.; del Valle, J.; Castellanos, E.; Feliubadaló, L.; Pineda, M.; Brunet, J.; Serra, E.; Capellà, G.; Lázaro, C.; Gel, B. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur. J. Hum. Genet.* **2020**, *28*, 1645–1655. [CrossRef] [PubMed]
25. Yao, R.; Zhang, C.; Yu, T.; Li, N.; Hu, X.; Wang, X.; Wang, J.; Shen, Y. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol. Cytogenet.* **2017**, *10*, 30. [CrossRef]
26. Sadedin, S.P.; Ellis, J.A.; Masters, S.L.; Oshlack, A. Ximmer: A system for improving accuracy and consistency of CNV calling from exome data. *GigaScience* **2018**, *7*, giy112. [CrossRef] [PubMed]
27. Chiang, T.; Liu, X.; Wu, T.-J.; Hu, J.; Sedlazeck, F.J.; White, S.; Schaid, D.; de Andrade, M.; Jarvik, G.P.; Crosslin, D.; et al. Atlas-CNV: A validated approach to call single-exon CNVs in the eMERGESeq gene panel. *Genet. Med.* **2019**, *21*, 2135–2144. [CrossRef] [PubMed]
28. 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]
29. Gambin, T.; Akdemir, Z.C.; Yuan, B.; Gu, S.; Chiang, T.; Carvalho, M.B.; Shaw, C.; Jhangiani, S.; Boone, P.M.; Eldomery, M.K.; et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.* **2017**, *45*, 1633–1648. [CrossRef]
30. 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.; Sijmons, R.H.; et al. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum. Mutat.* **2016**, *37*, 457–464. [CrossRef]
31. Musacchia, F.; Ciolfi, A.; Mutarelli, M.; Bruselles, A.; Castello, R.; Pinelli, M.; Basu, S.; Banfi, S.; Casari, G.; Tartaglia, M.; et al. VarGenius executes cohort-level DNA-seq variant calling and annotation and allows to manage the resulting data through a PostgreSQL database. *BMC Bioinform.* **2018**, *19*, 477. [CrossRef]
32. Ellingford, J.M.; Campbell, C.; Barton, S.; Bhaskar, S.; Gupta, S.; Taylor, R.L.; Sergouniotis, P.I.; Horn, B.; Lamb, J.A.; Michaelides, M.; et al. Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur. J. Hum. Genet.* **2017**, *25*, 719–724. [CrossRef]
33. 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*, S1. [CrossRef] [PubMed]
34. Garcia, M.; Juhos, S.; Larsson, M.; Olason, P.I.; Martin, M.; Eisfeldt, J.; DiLorenzo, S.; Sandgren, J.; Díaz De Ståhl, T.; Ewels, P.; et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **2020**, *9*, 63. [CrossRef]
35. Van de Sompele, S.; Smith, C.; Karali, M.; Corton, M.; Van Schil, K.; Peelman, F.; Cherry, T.; Rosseel, T.; Verdin, H.; Derolez, J.; et al. Biallelic sequence and structural variants in RAX2 are a novel cause for autosomal recessive inherited retinal disease. *Genet. Med. Off. J. Am. Coll. Med Genet.* **2019**, *21*, 1319–1329. [CrossRef]
36. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: New York, NY, USA, 2013; Volume 43, pp. 11.10.1–11.10.33. [CrossRef]
37. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595. [CrossRef]
38. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
39. Geoffroy, V.; Herenger, Y.; Kress, A.; Stoetzel, C.; Piton, A.; Dollfus, H.; Muller, J. AnnotSV: An integrated tool for structural variations annotation. *Bioinformatics* **2018**, *34*, 3572–3574. [CrossRef] [PubMed]
40. Cheng, J.; Demeulemeester, J.; Wedge, D.C.; Vollan, H.K.M.; Pitt, J.J.; Russnes, H.G.; Pandey, B.P.; Nilsen, G.; Nord, S.; Bignell, G.R.; et al. Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat. Commun.* **2017**, *8*, 1–14. [CrossRef]
41. Cox, C.; Bignell, G.; Greenman, C.; Stabenau, A.; Warren, W.; Stephens, P.; Davies, H.; Watt, S.; Teague, J.; Edkins, S.; et al. A survey of homozygous deletions in human cancer genomes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4542–4547. [CrossRef]
42. Lupski, J.R.; Belmont, J.W.; Boerwinkle, E.; Gibbs, R.A. Clan genomics and the complex architecture of human disease. *Cell* **2011**, *147*, 32–43. [CrossRef]
43. Valduga, M.; Philippe, C.; Lambert, L.; Bach-Segura, P.; Schmitt, E.; Masutti, J.P.; François, B.; Pinaud, P.; Vibert, M.; Jonveaux, P. WWOX and severe autosomal recessive epileptic encephalopathy: First case in the prenatal period. *J. Hum. Genet.* **2015**, *60*, 267–271. [CrossRef] [PubMed]
44. Levy, S.E.; Myers, R.M. GG17CH05-Levy Advancements in Next-Generation Sequencing. *Annu. Rev. Genom. Hum. Genet.* **2016**, *17*, 95–115. [CrossRef] [PubMed]
45. Samarakoon, P.; Sorte, H.; Kristiansen, B.; Skodje, T.; Sheng, Y.; Tjønnfjord, G.E.; Stadheim, B.; Stray-Pedersen, A.; Rødningen, O.; Lyle, R. Identification of copy number variants from exome sequence data. *BMC Genom.* **2014**, *15*, 661. [CrossRef] [PubMed]
46. Zare, F.; Dow, M.; Monteleone, N.; Hosny, A.; Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinform.* **2017**, *18*, 286. [CrossRef] [PubMed]