Copy Number Variation Detection in Next Generation Sequencing Data

Por

Maria de Lurdes Gonçalves Caloba

Advisor: Viviana Vilar da Silva

Coordinator: Dario Joaquim Simões Loureiro dos Santos

Dissertação submetida à
UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO
para obtenção do grau de
MESTRE
em Biotecnologia para as Ciências da Saúde

Copy Number Variation Detection in Next Generation Sequencing Data

Por

Maria de Lurdes Gonçalves Caloba

Orientador: Viviana Vilar da Silva

Co-orientador: Dario Joaquim Simões Loureiro dos Santos

Dissertação submetida à
UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO
para obtenção do grau de
MESTRE
em Biotecnologia para as Ciências da Saúde, de acordo com o disposto no
DR – I série–A, Decreto-Lei n.º 74/2006 de 24 de Março e no
Regulamento de Estudos Pós-Graduados da UTAD
DR, 2.ª série – Deliberação n.º 2391/2007

Orientação Científica :

Viviana Vilar da Silva

do
Departamento Diagnóstico Molecular e Genómica Clínica
Centro de Genética Clínica
Porto

Dario Joaquim Simões Loureiro dos Santos

do
Departamento de Química
Escola de Ciências da Vida e do Ambiente
da Universidade de Trás-os-Montes e Alto Douro

”In the middle of difficulty lies opportunity” | ”No meio da dificuldade encontra-se a oportunidade.”

Einstein (1879 – 1955)

”Success is going from failure to failure without losing enthusiasm | Sucesso é ir de fracasso em fracasso sem perder o entusiasmo”

Winston Churchill(1874 – 1965)

Copy Number Variation Detection in Next Generation Sequencing Data

Maria de Lurdes Gonçalves Caloba

Submetido na Universidade de Trás-os-Montes e Alto Douro
para o preenchimento dos requisitos parciais para obtenção do grau de
Mestre em Biotecnologia para as Ciências da Saúde

Resumo — As variações genéticas no genoma humano podem ser desde grandes anomalias cromossómicas (aneuploidias segmentais), variações de um único nucleótido (SNVs) a pequenas inserções ou deleções (indels). Ganhos ou uma perdas do número de cópias vão corresponder, respectivamente, a duplicações ou deleções estruturais genómicas. Estes ganhos e perdas de cópias de genes são uma fonte comum de variação genética que têm sido implicados em muitas doenças genómicas. Os objetivos deste trabalho são a validação da deteção de CNVs com dados de NGS, o calculo da percentagem de diagnóstico com esta deteção tanto em amostras para paineis oncológicos como em WES, determinação de parâmetros de qualidade e respectivo limiar de deteção.

A validação da análise de CNVs com o programa VarSeq consistiu na junção de casos positivos (com deleções ou duplicações) e negativos confirmados por outro método, na sua análise para calcular a especificidade, sensibilidade, exatidão e valores preditivos positivo e negativo. Depois da validação do software obteve-se a percentagem de diagnóstico.

Foram incluidos neste estudo 902 pacientes para análise oncológica e o diagnóstico obtido foi de 2.54% para CNVs.

Para as amostras de exoma foram utilizadas um total de 540 amostras cuja análise resultou num diagnóstco para CNVs de 8.15%.

Foi possível detetar CNVs com uma sensibilidade de 99.15%, especificidade de 98.85% e exatidão de 98.90% para paineis oncológicos; para WES foi obtida uma sensibilidade de 87.50%, especificificidade de 99.30% e exatidão de 97.84%.

CNVs podem ser detetados com exatidão e com uma taxa de diagnóstico entre 3−8%

o que é relevante para a gestão clínica e para o aconselhamento genético tanto para os pacientes como para os seus familiares. Este estudo evidencia o potencial da sequenciação de nova geração como método para deteção de CNVs robusto e com uma boa relação custo-benefício, tanto para paineís oncológicos como para exomas.

Palavras Chave: Variação-Número-Cópias, Sequenciação-Nova-Geração, Variação-Genética, Genoma.

Copy Number Variation Detection in Next Generation Sequencing Data

Maria de Lurdes Gonçalves Caloba

Submetido na Universidade de Trás-os-Montes e Alto Douro
para o preenchimento dos requisitos parciais para obtenção do grau de
Mestre em Biotecnologia para as Ciências da Saúde

Abstract — Genetic variation in the human genome can range from large chromosomic anomalies to single nucleotide variations (SNVs), including structural variations, copy number variations (CNVs), small indels, and individual base alterations. Copy number gains or losses correspond to genomic structural duplications or deletions, respectively, and these alterations can directly influence genic dosage, which has direct implications in genomic diseases. This work focuses on patients that made oncologic genetic tests and in the detection of potential causal CNVs.

The main goals are the validation of CNVs detected with NGS data, calculation of the diagnostic yield with CNV detection in oncologic, and whole exome sequencing samples diagnostic. Determination of quality parameters and respective detection thresholds.

CNVs analysis validation involving CNV detection program VarSeq consisted of gathering positive (with deletions/duplications) and negative cases confirmed by another method, analyzing NGS data for the detection and calculating specificity, sensitivity, accuracy, and positive, and negative predictive values. After software validation, the diagnostic yield was calculated.

There were included in this study 902 patients from oncologic testing with a diagnostic yield for CNVs was 2.54%.

For WES samples, a total of 540 patients were analyzed with a diagnostic rate of 8.15% for CNVs.

CNV accurate detection is possible, with a sensitivity of 99.15%, specificity of 98.85% and accuracy of 98.90% for cancer panels. For exomes, accounting all the alterations, the sensitivity is 87.50%, the specificity achieved of 99.30% and the accuracy of

97.84%.

CNVs can accurately be detected and increase diagnostic yield by $3 - 8\%$, which is relevant for clinical management and genetic counseling to patients and their relatives. This study proves the potential of NGS as a reliable and affordable method to detect CNVs, both in target panel (as cancer panels) and WES.

Key Words: Copy-Number-Variation, Next-Generation-Sequencing, Genetic-Variation, Genome.

Acknowledgments

First of all, I need to recognize my gratitude to **Professor Doutor António Augusto Fontainhas Fernandes**, for the years of formation during my bachelor's and master's degree.

To **Professora Doutora Isabel Gaivão** for the patience and comprehension demonstrated with me along with this work, especially with all the complications she helped solve.

To **Viviana Vilar da Silva** for the guidance and advisory. Her positiveness, experience, and disponibility to help me in everything I needed through this path.

To **Professor Doutor Dario Joaquim Simões Loureiro dos Santos** for the acceptance to be my coordinator, the patience demonstrated, and guidance through this work.

To **CGC Genetics** (a Unilabs company) for offering me an internship in which I was able to develop my master's thesis.

To **Doutor Jorge Pinto Basto** and **Doutora Rita Cerqueira** for providing their experience and knowledge, develop strategies to improve my work, and being a support for my growth.

To **all my colleagues from CGC Genetics**, especially to the **molecular diagnostic team** for receiving me, for being there for me all the time, for their help in

everything I ever needed, for making my days happier, for the support and everything they taught me.

To all my **master's degree colleagues** for the fantastic moments shared, in particular to **Rodolfo**, he was my guardian angel during this journey.

To **all my friends** from these last couple of years, especially to **Rui, Marlene, Felipa, and my college family** for the friendship shared, the good moments, the comprehension and the help. They were my standing stones during the roughest times.

To **Eduardo Cristo**, he was a fundamental help, gave me motivation and support that were essential to accomplish my goals.

To my **closest family** for all the patience, guidance, comprehension, and effort. For trusting me, on my choices, for always wanting my continuous improvement and all the choices they had to make to proportionate the possibility of following my dreams. Without your help, I would never succeed.

General index

Table index

Figures index

Glossary, acronyms, and abbreviations

aCGH	*Array comparative genome hybridization*
BAM	*Binary alignment/map format*
BAM.bai	*Binary alignment/map index file*
BED	*Browser extensible data*
bp	*base pairs*
CGH	*Genomic comparative hybridization*
CNV	*Copy number variation*
CNP	*Copy number polymorphism*
DNA	*Desoxyribonucleic acid*
DOC	*Depth of coverage*
FISH	*Fluorescent in situ Hybridization*
FN	*False negatives*
FP	*False positives*
GC	*Guanine, cytosine*
HPO	*Human phenotype ontology*
IC	*Confidence interval*
IQR	*Interquartile range*
Kbp	*kilobase pairs*
Loci	*Specific position of a chromosome where a certain gene, or genetic marker, is positioned*

MLPA	*Multiplex ligation-dependent probe amplification*
NGS	*Next-generation sequencing*
NPV	*Negative predictive value*
PCR	*Polymerase chain reaction*
PEM	*Paired-end mapping*
PPV	*Positive predictive value*
qPCR	*Quantitative polymerase chain reaction*
R^2	*Coefficient of determination*
SNP	*Single nucleotide polymorphism*
SNV	*Single nucleotide variation*
TN	*True negatives*
TP	*True positives*
VCF	*Variant call format*
WES	*Whole-exome sequencing*

1 Introduction

The human genome has numerous forms of genetic variation, comprising single nucleotide variants (SNVs), small insertions or deletions (indels), copy number variations (CNVs), and large chromosomal-level changes (Tzeng et al. (2015), Zhao et al. (2013a), Redon et al. (2006), Pierce (2012)). These alterations may occur in a single chromosome (heterozygosity) or both homologous chromosomes (homozygosity) (Klug (2012), Weckselblatt and Rudd (2015), Conrad et al. (2009)).

Chromosome rearrangements include duplications, deletions, inversions, and translocations (Fig.1.1) of DNA structure, and can range from single exons of a gene to several genes (Fakhro et al. (2015), Legault et al. (2015), Pirooznia et al. (2015), HengWang and KaiYing (2014)).

1.1 Copy-Number Variations

Deletions and duplications with more than one kilobase (Kb) and less than five megabases (Mb) are known as copy number variations (CNVs). CNVs are an important and abundant source of genetic variation (Hehir-Kwa et al. (2015), Legault et al. (2015), Marcinkowska-Swojak et al. (2013), Pierce (2012), Redon et al. (2006),

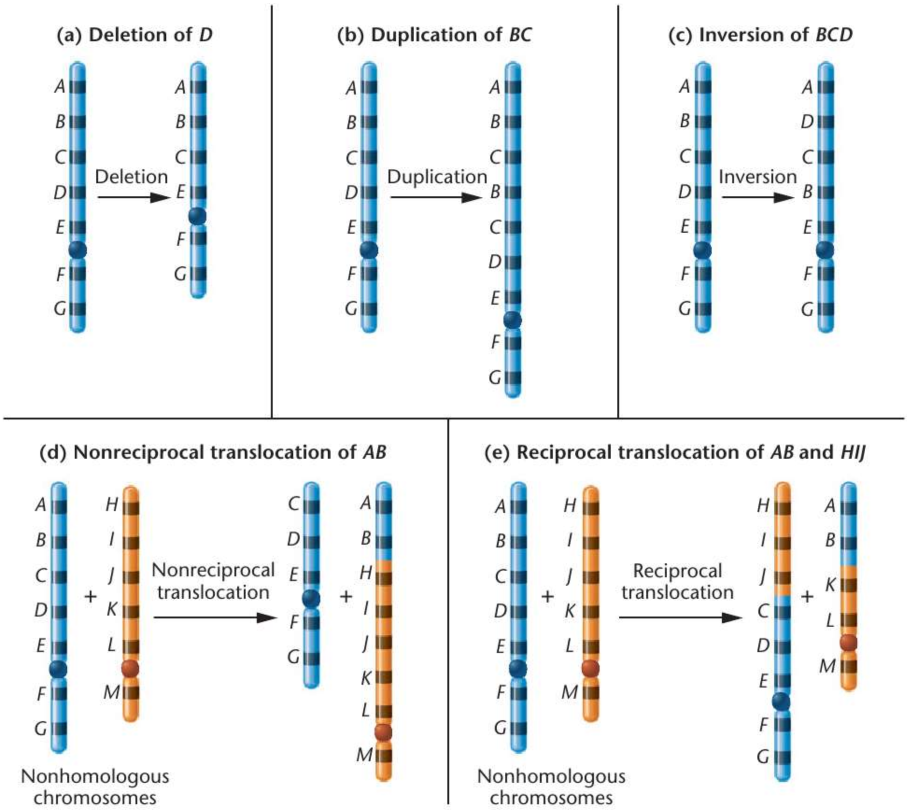


Figure 1.1 – The four main types of chromosome rearrangements: duplication, deletion, inversion, and translocation (Klug, 2012).

Valsesia et al. (2013)).

A significant portion of the genome of healthy individuals is susceptible to CNVs, and it is estimated that more than a thousand copy number variants are common in general population and with frequencies greater than 1% (Valsesia et al. (2013), Tzeng et al. (2015)).

It has been proved that some rare variants can be associated with mendelian disorder and cancer (Fakhro et al. (2015), Valsesia et al. (2013), Tzeng et al. (2015)) and such variants have already been described in disorders such as osteoporosis, congenital heart disease, autism, schizophrenia and hearing loss, (Fakhro et al. (2015), Hehir-Kwa et al. (2015), Kearney et al. (2011), Pierce (2012), Pirooznia et al. (2015), Zhao et al. (2013a)) breast, bladder, ovarian and colorectal cancer (Leary et al. (2008), Leary et al. (2008), Despierre et al. (2014), Xu et al. (2015), Silveira et al. (2014), Horpaopan et al. (2014), Bonberg et al. (2014), Foged et al. (2013)).

1.2 Copy number variations detection

CNV analysis is considered a standard approach for causal identification of developmental delay, autism spectrum disorders, or multiple congenital anomalies (Riggs et al., 2019).

Traditionally, CNVs detection were performed by chromosome microscopic observation. The first method used was the G-banded karyotyping. Fluorescent in situ hybridization (FISH) and fiber-FISH increased the resolution allowing both common and rare sub microscopic CNVs detection. However, these methodologies require intensive work from skilled professionals (Alkan et al. (2011), Miller et al. (2010), Valsesia et al. (2013), Foged et al. (2013), Zhao et al. (2013a), Weckselblatt and Rudd (2015)).

Later, the optimization of gene-specific customized assays, such as multiplex ligation-dependent probe amplification (MLPA) and qPCR, allowed CNVs detection at the molecular level (Roca et al., 2019). MLPA can evaluate a large number of loci based on PCR quantification fragments. With quantitative PCR, a large number of samples can be screened rapidly and accurately at a low cost per essay. They have the disadvantage of being limited to a small number of loci (Alkan et al., 2011).

In the context of whole-genome analysis, array comparative genome hybridization array (aCGH) is the gold standard for CNV detection. The ability to improve probe density increases the accuracy of the detection but it comes with the disadvantage of being more expensive (Alkan et al. (2011), Marcinkowska-Swojak et al. (2013), Valsesia et al. (2013)).

Similar to CGH technologies, SNP microarrays are also based on hybridization. SNP and CGH microarrays can detect from tenths to hundreds of events in a genome (Alkan et al. (2011), Valsesia et al. (2013)).

Hybridization methods are less effective in GC-rich regions and pseudogenes, reducing the accuracy of the detection (Alkan et al. (2011), Valsesia et al. (2013), Weckselblatt and Rudd (2015)).

The development of next-generation sequencing (NGS) and bioinformatic tools revolutionized genetic variation detection, mainly for high-throughput screening. In a unique assay, analysis of SNVs, indels, and CNVs is performed with reduced costs and lower turnaround time (Alkan et al. (2011), Legault et al. (2015), Sinha et al. (2015), Valsesia et al. (2013), Guo et al. (2014)).

1.2.1 NGS and CNVs

NGS platforms have probes that align randomly in the genome, in contrast with array-based approaches that are limited to targeted regions. This way, next-generation sequencing has progressed and is a popular strategy for characterizing CNVs, generating hundreds of millions of short reads in a single run (Metzker, 2009).

Among the advantages of NGS greater resolution and coverage, increased precision on the estimation of copy numbers, and breakpoint detections can be highlighted (Zhao et al., 2013b).

However, larger variants (like CNVs) are not detected as part of the data analysis routine. Even though several exome CNVs detection methods are available, they are difficult to use, and accuracy varies unpredictably between and within data sets (Sadedin et al. (2018), Zhao et al. (2013b)).

The use of NGS technologies conjugated with advanced bioinformatics processing has the potential to change the face of genetic diagnosis by offering faster, more affordable, and higher-resolution testing options (Hehir-Kwa et al., 2018).

Several algorithms have been developed to provide accurate detection of CNVs (Roca et al. (2019), HengWang and KaiYing (2014), Hehir-Kwa et al. (2018)). Some of them will be described in short.

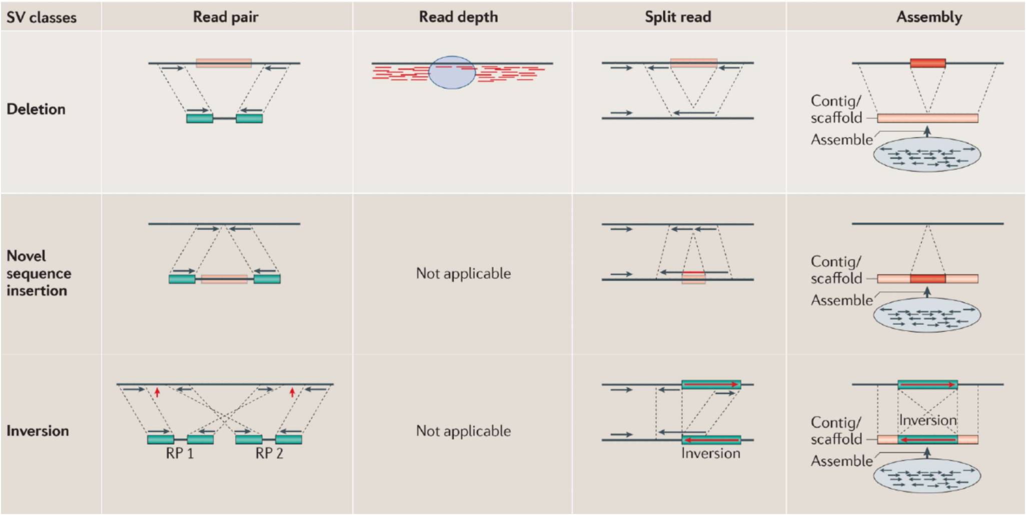


Figure 1.2 – Structural variation sequence signatures. Adapted from Alkan et al. (2011).

Read pair technologies (PEM)

Read pair approaches consist on sequence of interest fragmentation and its cloning into fosmids. Following, the alignment to the reference genome of the cloned fragments is performed using universal primers. This approach evaluates the length and orientation of paired-end reads and cluster pairs, which mapping is not compatible in span and/or orientation to the reference genome. Read pairs that map too distant define deletions, and those too close indicate duplications. Reads where there is only one end that clusters or others that have no match in the reference genome are variants flagged as novel insertions (they are not included in the reference genome), Fig. 1.2. This approach is the most widely used because it is a powerful tool. Ambiguous mapping assignments are challenging in repetitive regions because its precision relies on very tight distribution of the fragment sizes, which leads to a hard and expensive library construction (Alkan et al. (2011), Legault et al. (2015), Sinha et al. (2015), Valsesia et al. (2013), Zhao et al. (2013a)).

Read depth methods (DOC)

Read depth technologies have a random distribution, investigating the divergency in depth mapping and compares to its expected distribution. The base of this methodology relies on the fact that deletions will have reduced read depth, and on the contrary, duplications will show a higher read when compared to diploid regions (Fig. 1.2). This approach was first used to define cancer rearrangements with NGS and afterwards applied in segmental duplication and copy number absolute maps to the human genome. It can predict correctly absolute copy numbers but its resolution in breakpoints is usually poor. This method's main weakness is the influence that GC-content, library preparation variations, homologous regions, or low mappability can have in DOC differences between samples in a determined region. These factors will negatively influence the results causing an increase in false positives (Alkan et al. (2011), Legault et al. (2015), Roca et al. (2019), Sinha et al. (2015), Valsesia et al. (2013), de Ligt et al. (2013)).

Split read methods (SR)

Split read methods begin with a pair of reads in which the alignment occours. Part align to the reference genome exclusively and the other only maps partially or does not map at all. The focus of the last ones is breakpoints detection. These incompletely mapped reads are splited into multiple fragments. The first and last parts of each read are aligned to the reference genome independently. The remapping indicates the positions were deletions and insertions start and end. Furthermore, a line with continuous breaks indicates the deletion and if there are breaks in the reference genome, it corresponds to insertions. However, this aproach relies in the read length and can only be applied to unique regions in the reference genome and the alignment of small reads is difficult (Zhao et al. (2013b), Alkan et al. (2011), Valsesia et al. (2013)).

Sequence assembly

Theoretically, the structural variation could be analyzed in terms of copy, structure, and content if the reads could be long and sufficiently accurate to allow de novo assembly (Fig. 1.2). Sequence assembly methods have recently appeared and they usually use a combination of algorithms to local and de novo assembly, generating sequence contigs (DNA fragments reconstructed) that will be compared to a reference genome. A perfect approach would be capable of identifying thousands of variants if performed a de novo assembly and compared with a high-quality reference. Approaches that need this level of library construction, clone array, and end sequencing are too expensive and too laborious to be widely used. This type of assembly is promising and probably the most versatile by facilitating the genome comparison. However, the assembly would collapse in repeated regions and needs a minimum read coverage to detect overlapping fragments (Alkan et al. (2011), Valsesia et al. (2013), Zhao et al. (2013a)).

None of these approaches are comprehensive. When the same sample is tested with different methodologies, the results are inconsistent with many variants detected uniquely with one of the methods. Some softwares incorporate multiple methodologies improving sensitivity, specificity, and accuracy, combining read pair, read depth, and split read approaches. In this way, CNVs detection is more reliable, in part, because this junction puts the variants detected in the context of the population genetics (Alkan et al. (2011), Valsesia et al. (2013), Zhao et al. (2013a)).

CNV detection on NGS data can be performed from WES/WGS or NGS target panel. WES and WGS data offer a comprehensive study of the exome or genome required for several disorders. Usually, it is taken the approach to sequence samples at low coverage. It increases the cost efficiency but decreases the detection of structural variants. In order to improve mean coverage for clinically relevant genes, it is better to use target panels (Alkan et al. (2011), Guo et al. (2014), Zhao et al. (2013a)).

Another advantage of NGS technologies is the possibility of discovering several variant classes with one sequence experiment. In this way, it is possible to estimate the absolute copy number of duplicated regions in the human genome accurately. Characterizing and distinguishing them regions is essential to understand the effect of duplications in phenotypic differences (Alkan et al., 2011).

The most important NGS data drawback is its nature. Sequence reads are short and, due to the human genome complexity, lead to ambiguity. The solution would be to increase the specificity by enlarging inserts and reads. Additionally, NGS data requires investment in computational tools for storage and analysis and including more information requires more time spending on the analysis (Alkan et al. (2011), Guo et al. (2014), Zhao et al. (2013a)).

1.2.2 Software for CNVs detection

Bioinformatic tools to accurately detect CNVs from NGS data have been developed in the last years. Read depth approaches were successfully combined with whole-exome sequencing. Some examples are CONTRA, ExomeDepth, CoNIFER, cn.MOPS and XHMM (Zhao et al. (2013b), Tan et al. (2014), Zare et al. (2017)).

CNV Caller, from Golden Helix, is a recent software, released in 2017 that uses normalized DOC analysis. For coverage normalization, it uses a set of control samples. Matched reference controls are further used to overcome GC-content and mappability issues. The z-score is measured (number of standard-deviation in which a sample's coverage is from the mean reference sample coverage). The called CNVs were assigned accordingly to the probability of each targeted region exhibiting a diploid state or event: heterozygous deletion, homozygous deletion, or duplication. (Golden Helix, Inc., Bozeman, MT, goldenhelix.com).

1.3 Objectives

The main goal of this project was the implementation and consequent validation of CNV detection on NGS data and its application on clinical diagnosis. Therefore, the following were required:

- Detection and interpretation of CNVs using NGS data;
- Compare resultsfrom VarSeq with aCGH, MLPA, and qPCR data;
- Estimate the sensitivity, specificity, accuracy, negative and positive predicted values for cancer panels and WES;
- Determination of quality parameters and respective detection thresholds;
- Calculation of the diagnostic yield with CNV detection in oncologic, and whole exome sequencing samples diagnostic;
- Application of CNVs detection on diagnosis routines of cancer panels and WES.

2 Material and Methods

2.1 CNV software detection

In the present study VarSeq^{TM} v2.1.2 (Golden Helix, Inc., Bozeman, MT, goldenhelix.com) was used for detection of CNVs in a routine diagnostic using data from NGS.

VarSeq CNV Caller requests as input VCF, BAM, and BAM.bai files, generated with NGS from each sample under investigation and a BED file (to identify the target regions). This algorithm requires a set of, at least, ten control references for values normalization. The used controls were from probands of the corresponding performed test. Samples are flagged by the software if there is more than a 20% discrepancy from the test sample to the references. Samples with flags were not analyzed.

2.2 CNV validation set

For CNV validation samples that had been previously characterized, by other methods, as containing one clinically relevant associated CNV were selected. The validation process included twenty oncologic patients to evaluate the sensitivity, specificity, and accuracy, positive and negative predictive values. All CNV calls were divided by exon and compared with MLPA or qPCR results. For example, a deletion that was detected in exon 23 to 26 but the confirmation only detects the deletion from exon 23 to 25 would be considered a false positive. The exons detected by both methodologies would be classified true positives.

After the software validation for the cancer panel, positive samples were used to WES validation. This set includes six oncologic samples (previously tested for cancer panels and re-sequenced for WES) and additional positive samples previously performed by aCGH and qPCR (in total 30 samples).

2.3 Patient Samples

This study includes a worldwide population of patients referred to CGC Genetics between 2017 and 2019. This work focus on samples referred for NGS cancer panels and whole-exome sequencing. The NGS cancer panel includes 89 genes (summarized in Table 2.1).

A total of 902 patients were tested for NGS cancer panel and 540 for whole-exome sequencing to search disease-causing CNVs and define the adequate procedure to apply to the daily routine of CNVs detection. The NGS results were confirmed independently by MLPA, qPCR, or aCGH.

Table 2.1 – Genes present in cancer panels

Panel	Genes
Hereditary colorectal cancer	*APC, AXIN2, BMPR1A, CDH1, CHEK2, EPCAM, GALNT12, MLH1, MLH3, MSH2, MSH3, MSH6, MUTYH, PMS2, POLD1, POLE, PTEN, SMAD4, STK11, TGFBR2, TP53*
Ovarian and breast cancer	*ATM, BLM, BRCA1, BRCA2, BRIP1, CDH1, CDKN2A, CHEK2, EPCAM, FANCC, FANCM, MEN1, MLH1, MSH2, MSH6, MUTYH, NBN, PALB2, PMS2, PTEN, RAD51C, RAD51D, SLX4, STK11, TP53, NF1*
Gastric cancer	*CDH1, MLH1, MSH2, MSH6, PMS2, EPCAM*
Pancreatic cancer	*BRCA1,BRCA2,TP53,PALB2,STK11*
OncoRisk Expanded	*AIP, ALK, APC, ATM, BAP1, BLM, BMPR1A, BRCA1, BRCA2, BRIP1, BUB1B, CDC73, CDH1, CDK4, CDKN1C, CDKN2A, CEBPA, CEP57, CHEK2, CYLD, DDB2, DICER1, DIS3L2, EPCAM, ERCC2, ERCC3, ERCC4, ERCC5, EXT1, EXT2, EZH2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FH, FLCN, GATA2, GPC3, KIT, MAX, MEN1, MET, MLH1, MSH2, MSH6, MUTYH, NBN, NF1, NF2, NSD1, PALB2, PHOX2B, PMS2, PRF1, PRKAR1A, PTCH1, PTEN, RAD51C, RAD51D, RB1, RECQL4, RET, RUNX1, SBDS, SDHAF2, SDHB, SDHC, SDHD, SLX4, SMAD4, SMARCB1, STK11, SUFU, TMEM127, TP53, TSC1, TSC2, VHL, WRN, WT1, XPA, XPC*

2.4 Sequencing

Genomic DNA was extracted from whole blood samples or amniotic fluid. Next-generation sequencing (Illumina) of genomic DNA, upon the capture of target regions using oligonucleotide probes (Agilent Technologies). Oncologic samples were sequenced with the MiSeq system using QXT custom probes (NGS cancer panel), while WES samples were sequenced with NextSeq using Human All Exon V6 probes. Alignment and base calling were performed with the Burrows-Wheeler Aligner (BWA) and the Genome Analysis Toolkit (GATK), using as reference genome Homo sapiens (UCSC hg19). SNVs, indels, and CNVs were filtered and a structured analysis was performed to assess their pathogenicity and potential to explain the clinical phenotype. The variant classification was performed according to international recommendations (Richards et al. (2015), Rehm et al. (2013), Riggs et al. (2019)).

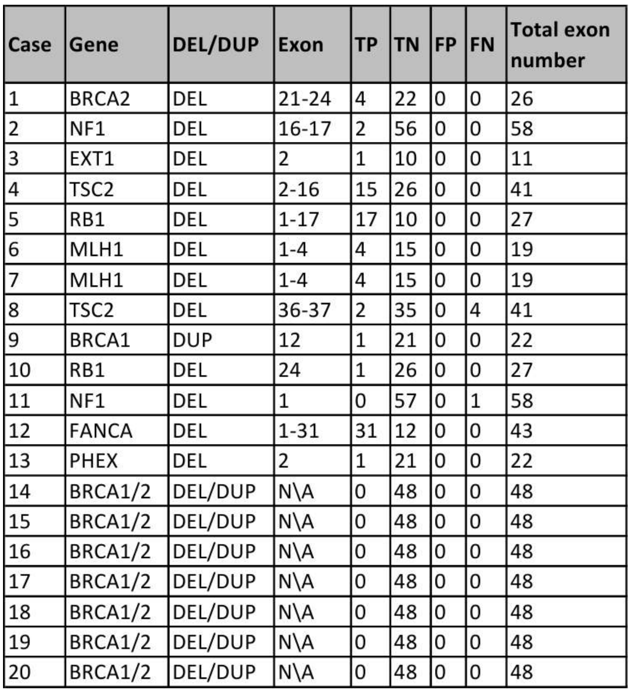
3 Results

3.1 Software Application for Oncology Data

The initial validation consisted of twenty cases, thirteen positives previously confirmed by MLPA and seven negatives for *BRCA1*/*BRCA2* genes (also confirmed by MLPA). Counting by exon number, these 20 cases correspond to 750 exons. The negative cases correspond each to 48 exons (total exons from *BRCA1* and *BRCA2*). For the true positives (TP) the total number of exons was 83, 662 for true negatives (TN), there were no false positives (FP) and 5 false negatives (FN), Table 3.1.

The obtained results were equal to the predicted ones except for two cases, the *NF1* deletion and the *TSC2* deletion (Table 3.1). The *NF1* alteration, detected by MLPA, comprises a 61bp deletion of exon 1. In NGS data, this region has low coverage due to the high GC content (more than 70%). The first exon of several genes is usually associated with a high CG-content that makes the amplification and sequencing more difficult. The *TSC2* alteration was detected larger than what was detected by MLPA. It was expected that this alteration would range from exon 36 to 37, but the software detected it from exon 36 to 42. These values can be explained by the low coverage of these exons and by their high GC content (around 60%).

Table 3.1 – Results confirmed by MLPA with and respective results from NGS data. del- deletion, dup- duplication, TP- true positives, TN- true negatives, FP- false positives, FN- false negatives



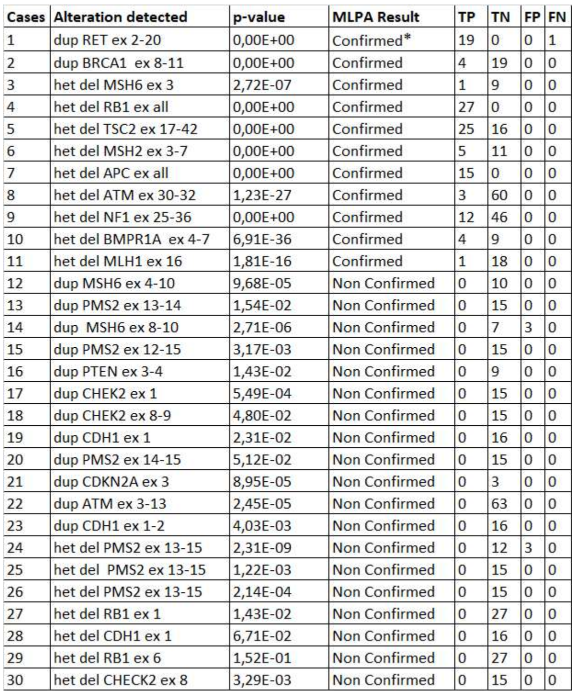
Case	Gene	DEL/DUP	Exon	TP	TN	FP	FN	Total exon number
1	BRCA2	DEL	21-24	4	22	0	0	26
2	NF1	DEL	16-17	2	56	0	0	58
3	EXT1	DEL	2	1	10	0	0	11
4	TSC2	DEL	2-16	15	26	0	0	41
5	RB1	DEL	1-17	17	10	0	0	27
6	MLH1	DEL	1-4	4	15	0	0	19
7	MLH1	DEL	1-4	4	15	0	0	19
8	TSC2	DEL	36-37	2	35	0	4	41
9	BRCA1	DUP	12	1	21	0	0	22
10	RB1	DEL	24	1	26	0	0	27
11	NF1	DEL	1	0	57	0	1	58
12	FANCA	DEL	1-31	31	12	0	0	43
13	PHEX	DEL	2	1	21	0	0	22
14	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
15	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
16	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
17	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
18	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
19	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48
20	BRCA1/2	DEL/DUP	N\A	0	48	0	0	48

For cancer panels, the values obtained for sensitivity were $98.81^{+1.16}_{-5.27}\%$, the specificity achieved was $99.34^{+0.48}_{-1.02}\%$. It was obtained a negative predictive value of $99.83^{+0.15}_{-0.98}\%$ and a positive predictive value of $95.40^{+2.82}_{-6.75}\%$ with an accuracy of $99.28^{+0.48}_{-0.96}\%$. These estimates are evaluated with a confidence interval of 95% and with the same interval henceforth (Table 3.2).

Table 3.2 – Exon number countably for quality detection evaluation

	Positive	Negative	Total
Positive Test	83	4	83
Negative Test	1	662	667
Total	88	662	750

Table 3.3 – Results gathered confirmed by MLPA and their respective p - value with the respective information. het del- heterozygotic deletion, dup- duplication, ex- exon, TP- true positives, TN- true negatives, FP- false positives, FN- false negatives. *the *RET* duplication detected by MLPA comprehends exons 1 to 20.



Cases	Alteration detected	p-value	MLPA Result	TP	TN	FP	FN
1	dup RET ex 2-20	0,00E+00	Confirmed*	19	0	0	1
2	dup BRCA1 ex 8-11	0,00E+00	Confirmed	4	19	0	0
3	het del MSH6 ex 3	2,72E-07	Confirmed	1	9	0	0
4	het del RB1 ex all	0,00E+00	Confirmed	27	0	0	0
5	het del TSC2 ex 17-42	0,00E+00	Confirmed	25	16	0	0
6	het del MSH2 ex 3-7	0,00E+00	Confirmed	5	11	0	0
7	het del APC ex all	0,00E+00	Confirmed	15	0	0	0
8	het del ATM ex 30-32	1,23E-27	Confirmed	3	60	0	0
9	het del NF1 ex 25-36	0,00E+00	Confirmed	12	46	0	0
10	het del BMPR1A ex 4-7	6,91E-36	Confirmed	4	9	0	0
11	het del MLH1 ex 16	1,81E-16	Confirmed	1	18	0	0
12	dup MSH6 ex 4-10	9,68E-05	Non Confirmed	0	10	0	0
13	dup PMS2 ex 13-14	1,54E-02	Non Confirmed	0	15	0	0
14	dup MSH6 ex 8-10	2,71E-06	Non Confirmed	0	7	3	0
15	dup PMS2 ex 12-15	3,17E-03	Non Confirmed	0	15	0	0
16	dup PTEN ex 3-4	1,43E-02	Non Confirmed	0	9	0	0
17	dup CHEK2 ex 1	5,49E-04	Non Confirmed	0	15	0	0
18	dup CHEK2 ex 8-9	4,80E-02	Non Confirmed	0	15	0	0
19	dup CDH1 ex 1	2,31E-02	Non Confirmed	0	16	0	0
20	dup PMS2 ex 14-15	5,12E-02	Non Confirmed	0	15	0	0
21	dup CDKN2A ex 3	8,95E-05	Non Confirmed	0	3	0	0
22	dup ATM ex 3-13	2,45E-05	Non Confirmed	0	63	0	0
23	dup CDH1 ex 1-2	4,03E-03	Non Confirmed	0	16	0	0
24	het del PMS2 ex 13-15	2,31E-09	Non Confirmed	0	12	3	0
25	het del PMS2 ex 13-15	1,22E-03	Non Confirmed	0	15	0	0
26	het del PMS2 ex 13-15	2,14E-04	Non Confirmed	0	15	0	0
27	het del RB1 ex 1	1,43E-02	Non Confirmed	0	27	0	0
28	het del CDH1 ex 1	6,71E-02	Non Confirmed	0	16	0	0
29	het del RB1 ex 6	1,52E-01	Non Confirmed	0	27	0	0
30	het del CHECK2 ex 8	3,29E-03	Non Confirmed	0	15	0	0

3.1.1 Quality parameters

The application of the software was applied to cases analyzed in CGC during my internship and all the calls tested by MLPA. A total of 30 samples, with CNVs detected by NGS and confirmed by MLPA, were used define quality thresholds (Table 3.3).

To establish quality cut off values it was considered the use of the statistical variables z-score and p-value. Each one was independently evaluated in terms of deletions and duplications (Fig. 3.1, Fig. 3.2, Fig. 3.3, Fig. 3.4).

It was also evaluated the distribution the z-score space for deletions fixing the threshold close to higher value of the true positive, in this case -5. Using the same criteria for the p-value was defined 10^{-5}, with a good sensitivity but generating a false positive (Fig. 3.1 and Fig. 3.2).

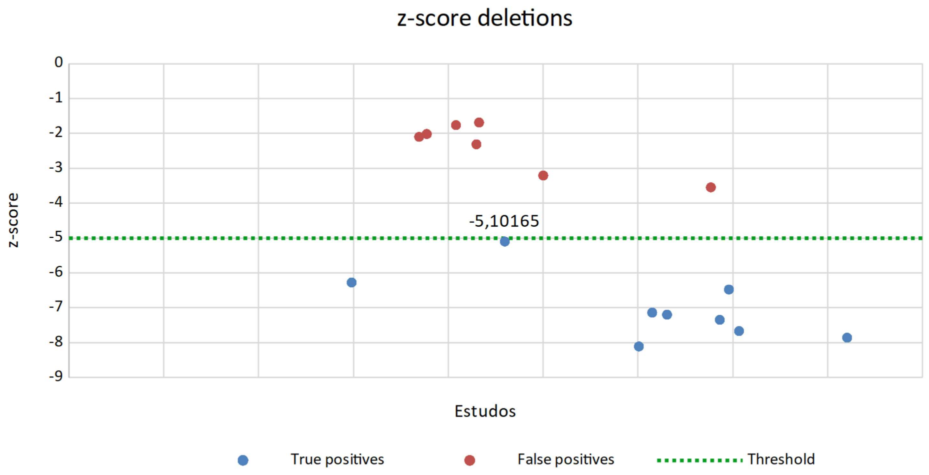


Figure 3.1 – $z - score$ distribution in deletions

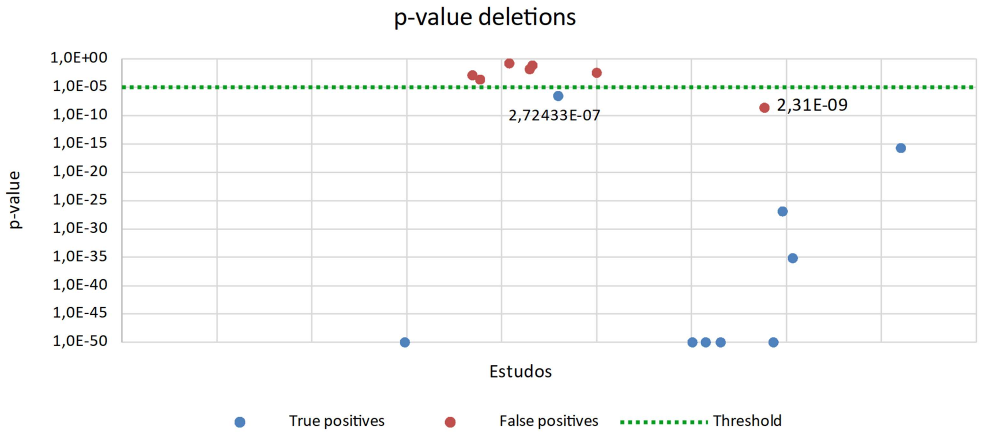


Figure 3.2 – p - value distribution in deletions

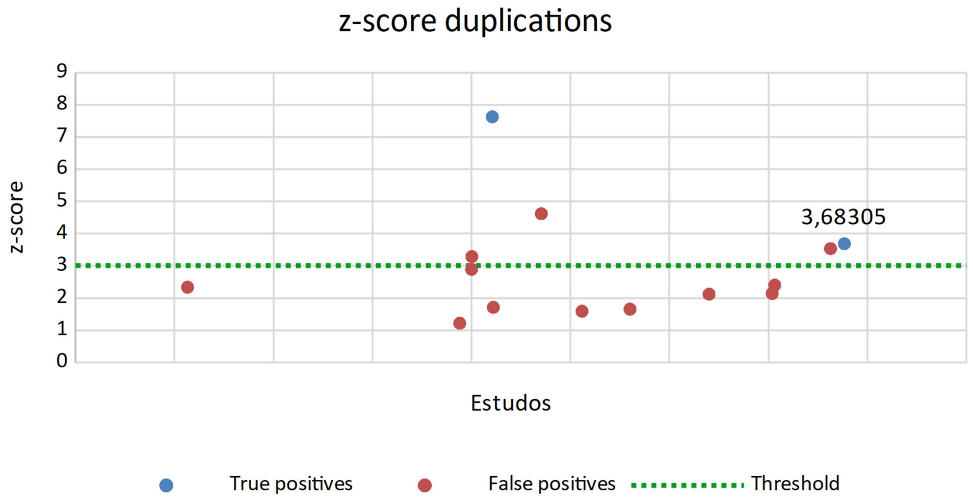


Figure 3.3 – $z - score$ distribution in duplications

Table 3.4 – Exon account with $|z - score| = 3$ for quality detection evaluation

z-score	Positives	Negatives	Total
Positive Test	116	11	127
Negative Test	1	508	509
Total	117	519	636

It was found an optimal value of three for the z-score only accounting duplications in order to achieve high sensitivity with loss specificity. For the p-value, only considering two duplications, it was also established a threshold of 10^{-5} adequate to determine the quality call of duplications (taking in account the threshold for deletions) (Fig. 3.3 and Fig. 3.4).

To clarify which of these would be the most practical cut-off value were calculated sensitivity and specificity for the total samples grouped by exons. The p-value was effectively used as a threshold. Despite sensitivity (99.15%) be the same for both z-score and p-value, the rate of false positives increases when z-score is used as the threshold (specificity decreases from 98.85% to 97.88% and positive predictive value decreases from 95.08% to 91.34%) (Table 3.5 and Table 3.4). In these circumstances, most of the cases would have zero CNV calls with some reaching one or two calls.

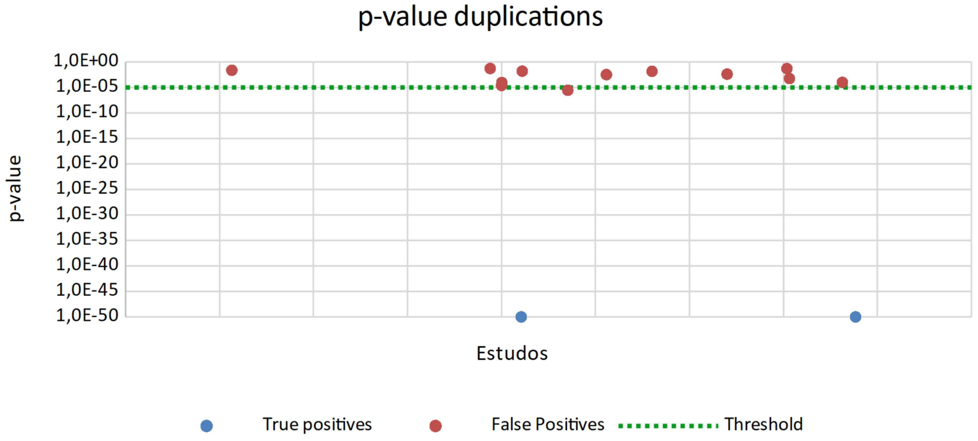


Figure 3.4 – p - value distribution in duplications

Table 3.5 – Exon account with p - value $= 10^{-5}$ for quality detection evaluation

p-value	Positives	Negatives	Total
Positive Test	116	6	122
Negative Test	1	514	515
Total	117	520	637

Additionally, with a p-value of 10^{-5}, for both duplications and deletions (this way sensitivity is ensured despite having few duplications), the only false negative was a duplication. NGS data detected it in exons 2 to 20 of *RET* gene, while MLPA confirmed a duplication on exons 1 to 20 of this gene. Once again, the NGS data does not detect a CNV in exon 1 due to the high GC content and consequent low coverage at this region.

Furthermore, there would be two false positives in genes *MSH6* and *PMS2*. The *PMS2* false positive from the case 24 probably is due to its pseudogene *PMS2CL* (with $>$ 99% similarity in exons 12–15), which interferes with NGS capture and sequencing reads alignment. The *MSH6* duplication presented a borderline p-value (2.71×10^{-6}), probably due to the low coverage associated to a high CG content (almost 50%). Additionally, the detection of duplications is more challenging than deletions, mainly in regions with low coverage. The rest of the calls were equal to the MLPA result (Table 3.3).

The sensitivity of $99.15^{+0.83}_{-3.82}\%$, specificity of $98.85^{+0.73}_{-1.34}\%$, accuracy of $98.90^{+0.66}_{-1.15}\%$, PPV of $95.08\_4.97^{+2.54}\%$, and NPV of $99.81^{+0.16}_{1.16-}\%$ were achieved for the p-value of 10^{-5} (Table 3.5). With these metrics, some interesting discoveries were made regarding the quality of the calls. Surprisingly, the software allows the detection of two alterations in mosaicism in *ATM* and *RET* genes. These mosaics envolved many exons which probably made its detection possible. Probably smaller alterations in mosaicism would not be detected. However, it is not possible to predict the frequency of each CNV detection, the MLPA confirmed that both alterations were mosaics of 25%.

Using the validated threshold, 23 patients out of 902 tested positive for this screening, giving a diagnostic yield of 2.54%.

3.2 Software Application for WES Data

In contrast to oncologic cases in exomes there is the problem of the total number of calls that make impossible the analysis and confirmation of all the detected variants. Through the process of WES validation, it was established a p-value threshold equal to 10^{-4}. This was possible by testing different samples and trying to find a balance between the number of analyzed variants (Tables from A to A.6, Figures from A.1 to A.6) and the sensitivity (Table 3.6) of the detection. The calculations were not made by exons count, like in oncologic samples, because some CNVs detected comprised several genes. Comparing the p-values achieved with the cancer kit with the ones from v6 with a threshold of 10^{-4}, the only call that would not pass the filter would be the *BRCA1* duplication. To lower the threshold to values in which this alteration could be detected was not feasible.

The WES statistical data was calculated with five CNVs detected in other NGS kits and now sequenced for WES (v6 kit). Additionally, a duplication detected only by MLPA was also included to increase the number of duplications tested since these are the more challenging ones (Table 3.6). Other CNVs were added to the validation, mainly CNVs confirmed by CGH, in the attempt to have a representative number

Table 3.6 – p-value comparison between the initial kit used and exome from the CNV software detection

Kit	Alteration	Gene	p-Value
Cancer.1	Het. Deletion	$NF1$	6.11×10^{-3}
v6	Het. Deletion	$NF1$	1.36×10^{-5}
Cancer.1	Duplication	$BRCA1$	2.69×10^{-11}
v6	Duplication	$BRCA1$	1.00×10^{-1}
6.1	Het. Deletion	$ABCA3$	1.72×10^{-13}
v6	Het. Deletion	$ABCA3$	8.43×10^{-34}
6.1	Het. Deletion	$GLI3$	1.71×10^{-40}
v6	Het. Deletion	$GLI3$, $INHBA$	0
9.1	Duplication	$PMP22$	1.42×10^{-3}
v6	Duplication	$CDRT4$, $CDRT15$, $COX10$, $HS3ST3B1$ $HS3ST3B1$, $PMP22$, $TEKT3$, $TVP23C$, $TVP23C - CDRT4$	0
MLPA	Duplication	DMD	$---$
v6	Duplication	$DMD, FTHL17$	0

of CNVs to validate the VarSeq software in WES data.

With the p-value threshold established to 10^{-4}, the average value of CNVs to analyze would be 100 variants at the top but with the phenotype filter would be reduced massively, even though the number of CNVs obtained remains challenging in some samples. During the validation process, it was found that CNVs screening is extremely sensitive to variations in the lab procedure and sample quality. Therefore, it was essential to establish the metrics that define the reliable samples and allows tthe exclusion of low-quality data that compromises the analysis. For example, choosing three different random samples, their mean in CNVs totals detection would be around 2000, when there is some error during the lab procedure, the total of CNVs would duplicate. When the error occurred during the assay preparation, the total number would almost triple, and a sample from a degraded sample would have 1000 more in average.

Table 3.7 – Quality detection in WES with a p-value threshold of 10^{-4}.

	Positive	Negative	Total
Positive Test	18	1	19
Negative Test	4	9	13
Total	22	10	32

One way to have an additional threshold regarding the total number of detected CNVs is by computing the percentage of low-quality calls. When this percentage of flagged calls, compared to the total calls, is low, between 50% and 60%, the CNV number detection is higher. Percentages around $80 - 90\%$ represent a lower number of CNVs detected and are associated with higher quality data. There was a direct correlation between the total number of copies detected and the percentage of flags retrieved by the software. Other tests were made in the attempt to find a correlation between coverage at 10×, with coverage uniformity or coverage mean. However, it was not found a linear correlation between the total number of detected CNVs and them (Figures 3.5, 3.6, and 3.7).

The value achieved for sensitivity was $81.82^{+12.99}_{-21.60}\%$, for specificity was $90.00^{+9.75}_{-34.5}\%$, negative predictive value was $69.23^{+15.6}_{-21.71}\%$ and the positive was $94.74^{+9.75}_{-34.5}\%$, the determined accuracy was $84.38^{+10.34}_{-17.17}\%$ (Table 3.7).

After some experimentation, sensitivity, specificity, accuracy, PPV, and NPV were calculated and an increase of the initial threshold to 1×10^{-7} was accessed and proved to be good without messing with sensitivity and specificity (Table 3.8, Table 3.9). The main difference between a p-value of 1×10^{-4} and 1×10^{-7} relies on the CNVs that had only one exon detected. In calls bigger than one exon, the specificity and sensitivity would remain the same. As well as CGH data, the main goal of WES is not to detect single exon CNVs, therefore several diagnostic labs does not reported them. Furthermore, the total number of CNV calls with that p-value would be significantly reduced.

With these values, if looking for the total number of alterations, the sensitivity

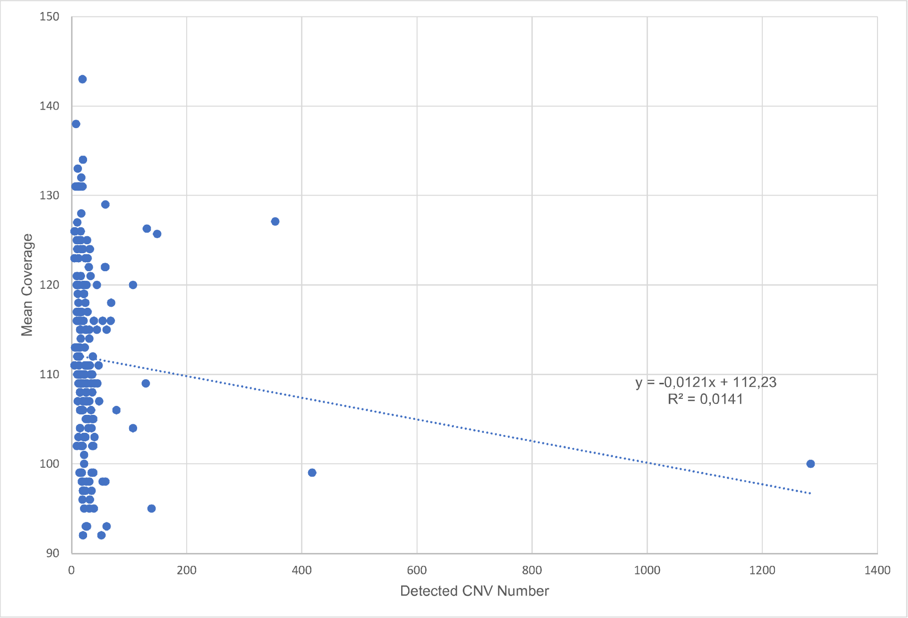


Figure 3.5 – Correlation between mean coverage and the detected number of CNVs. The value of R^2 is low, which indicates no significant relation between both quantities. This is confirmed by the existence of values with the same coverage rate and the number of detected CNVs that differ by two orders of magnitude.

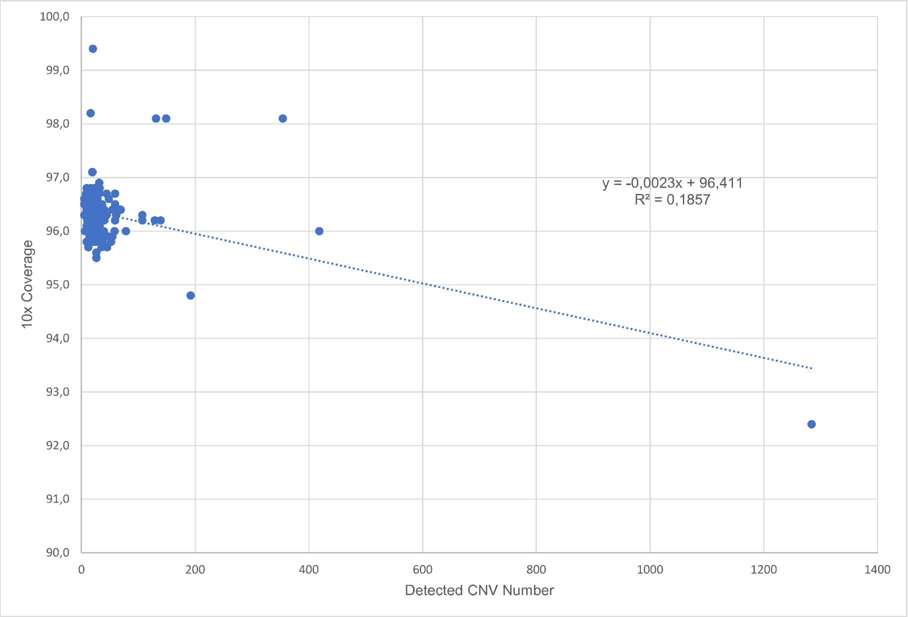


Figure 3.6 – The relation between $10\times$ coverage depth and the detected number of CNVs. It was not found a significant correlation.

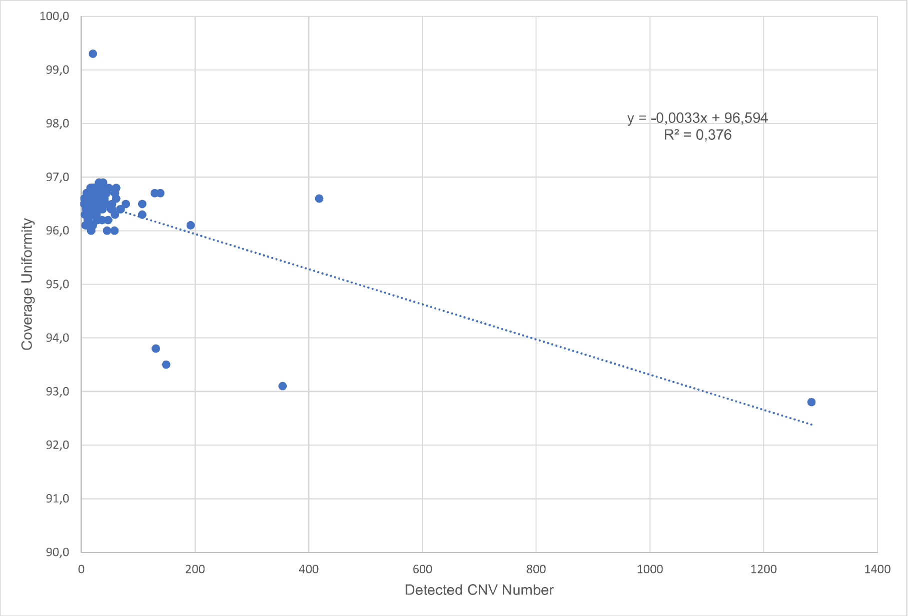


Figure 3.7 – Coverage uniformity vs CNV number detection. Despite the higher value of the coefficient of determination when compared with Fig. 3.5 and Fig. 3.6, it is not significant as a reliable indicator between both quantities.

Table 3.8 – Evaluation of quality detection in WES with a p-value threshold of 10^{-7} without alterations smaller than one target.

	Positive	Negative	Total
Positive Test	34	2	36
Negative Test	2	281	283
Total	36	283	319

Table 3.9 – Evaluation of quality detection in WES with a p-value threshold of 10^{-7} with all the targets.

	Positive	Negative	Total
Positive Test	35	2	37
Negative Test	5	282	287
Total	40	284	324

is $87.50\%^{+8.31}_{-14.30}$, the specificity achieved was $99.30^{+0.61}_{-1.82}$, a negative predictive value of 98.26% and the positive 94.59%, and accuracy of $97.84^{+1.29}_{-2.24}$. Furthermore, if we look only to the detected CNVs with more than one target (the majority), we can see an increase in these parameters. Sensitivity would be of $94.44^{+4.88}_{-13.10}\%$, the specificity achieved $99.29^{+1.82}_{-0.62}\%$, the negative predictive value computed 99.29% and the positive was 94.44%, the determined accuracy $98.75^{+0.91}_{-2.00}\%$ all with an interval of confidence of 95%.

These computed values were possible by aggregating every positive and negative confirmed calls and comparing their p-values (Table A.7). The application of CNVs detection in whole-exome sequencing was confirmed by CGH, qPCR, or MLPA.

Using the validated threshold, 44 patients out of 540 tested positive for this screening, which gives a diagnostic yield of 8.15%.

3.3 CNV detection limitations

After some testing, it was found some issues that constitutes limitations and compromise CNV detection:

- When only part of the exon is involved in an alteration;
- The first exon of each gene and GC rich regions are associated with low coverage;
- Sequences with pseudogenes or regions with high homology;
- The smaller deletion detected had $61bp$,so there was no way of excluding that deletions smaller than $60bp$ would not be detected;
- The smaller duplication detected had $35bp$, so there was no way of excluding that duplications smaller than $34bp$, would not be detected;
- The duplications fase is not possible to distinguish (heterozygosity and homozygosity);

- It is expected that lower quality samples (prenatal, blood from hematologic patients, and highly degraded DNA) generate lower quality NGS data. In these cases, CNV analysis may not be possible to perform;

3.4 Clinical Cases

This section will present some specific cases where CNVs detection had an impact on the diagnosis.

3.4.1 Case A - Dominant gene: *CRX*

Sample from an 18 years old female for a whole-exome analysis, including parents samples (WES Trio). The patient presented retinal dystrophy; her father and sister were also affected. The analysis integrated the parents and sister.

The heterozygous deletion NM_000554.4: c.(100+1_101-1)_(c.900+1_?)del, that comprises at least exons 3 and 4 of the *CRX* gene (chr19), was detected by CNVs analysis (Fig. 3.8) and confirmed by MLPA. This deletion was reported in the literature in families with retinitis pigmentosa Bravo-Gil et al. (2016), Martin-Merida et al. (2018). The parental and sister samples study indicated that the variant was inherited from the father, and it is also present in the sister, confirming the disease in the family. Therefore, with the available information, this should be classified as a pathogenic variant.

Pathogenic variants in the *CRX* gene cause autosomal dominant cone-rod retinal dystrophy-2 (MIM 120970), as well as Leber congenital amaurosis 7 (MIM 613829), with an autosomal dominant pattern of inheritance (Adam et al., 1993).

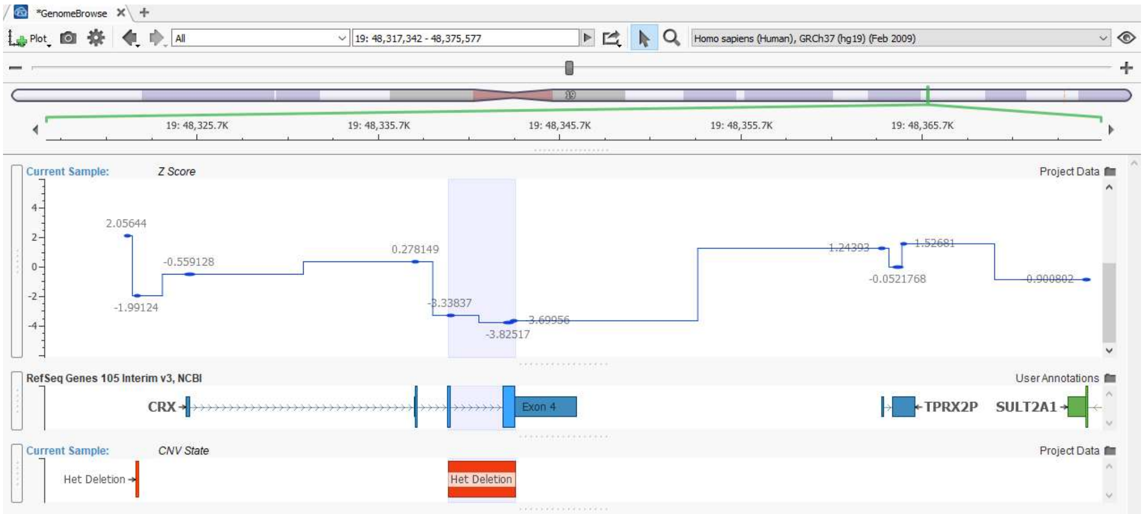


Figure 3.8 – CRX deletion depiction from VarSeq

3.4.2 Case B - Recessive gene: *CLN3*

Sample from a 5 years old male for whole-exome analysis, including parents samples (WES trio). The patient presented rapidly progressive retinal dystrophy and his physician suspected neuronal ceroid lipofuscinosis, specifically related to the *CLN3* gene. He had no family members affected and no parental consanguinity.

The NM_001042432.1: c.988G>T p.(Val330Phe) variant, detected in heterozygosity in the *CLN3* gene (chr.16), was described in the literature in a patient with neuronal ceroid lipofuscinosis (Munroe et al., 1997). It was also reported in dbSNP (rs386833744) and ClinVar databases as a likely pathogenic variant (ID:56296). It is located in a highly conserved residue and the bioinformatic analysis suggested it may be deleterious. Additionally, functional studies support its pathogenicity Gachet et al. (2005), Haines et al. (2009). The study of the parents indicated that the variant was inherited from the mother. With the available information, the variant should be classified as pathogenic.

The heterozygous NM_001042432.1: c.(790+1_791-1)_(1056+1_1057-1)del, that comprises at least exons 11 to 14 of *CLN3* gene, was detected by CNVs analysis (Fig. 3.9) and confirmed by MLPA. This deletion was reported in the literature in a patient with neuronal ceroid lipofuscinosis (Kousi et al., 2012) and functional studies

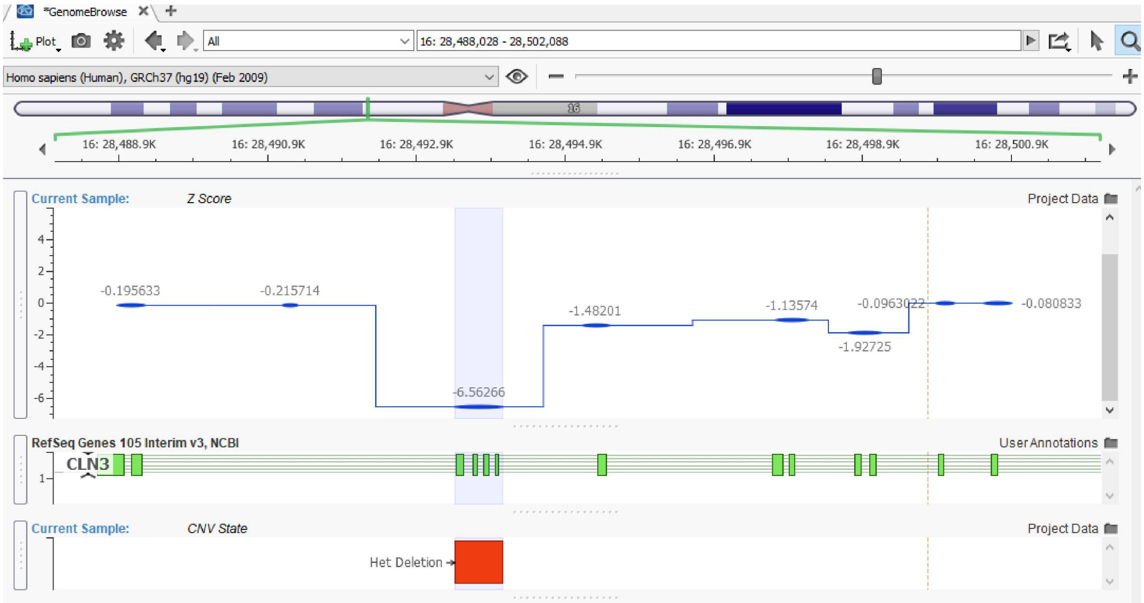


Figure 3.9 – CLN3 deletion depiction from VarSeq

support its pathogenicity (Haines et al., 2009). The study of the parents indicated that the variant was inherited from the father. Therefore, with the available information, the variant should be classified as pathogenic.

Pathogenic variants in the *CLN3* gene cause neuronal ceroid lipofuscinosis (MIM 204200) with an autosomal recessive pattern of inheritance. The study of the parents indicated that the variants are in different alleles (*trans*).

3.4.3 Case C - X-linked dominant gene: *MECP2*

Sample from a 3 years old female for whole-exome analysis. The patient presented epileptic encephalopathy, psychomotor developmental delay, normal somatometric parameters, stereotypy, tapering fingers, umbilicated nipples, fifth foot finger short and levels of FV, and FVII in the lower limit of normality and her physician suspected of a Rett-like disorder. She had no family history and no parental consanguinity.

The heterozygous NM_0004992.3: c.(26+1_27-1)_(1051_1214)del, that comprises at least exon 3 and part of the 4 of *MECP2* gene (chr.X), was detected by CNVs

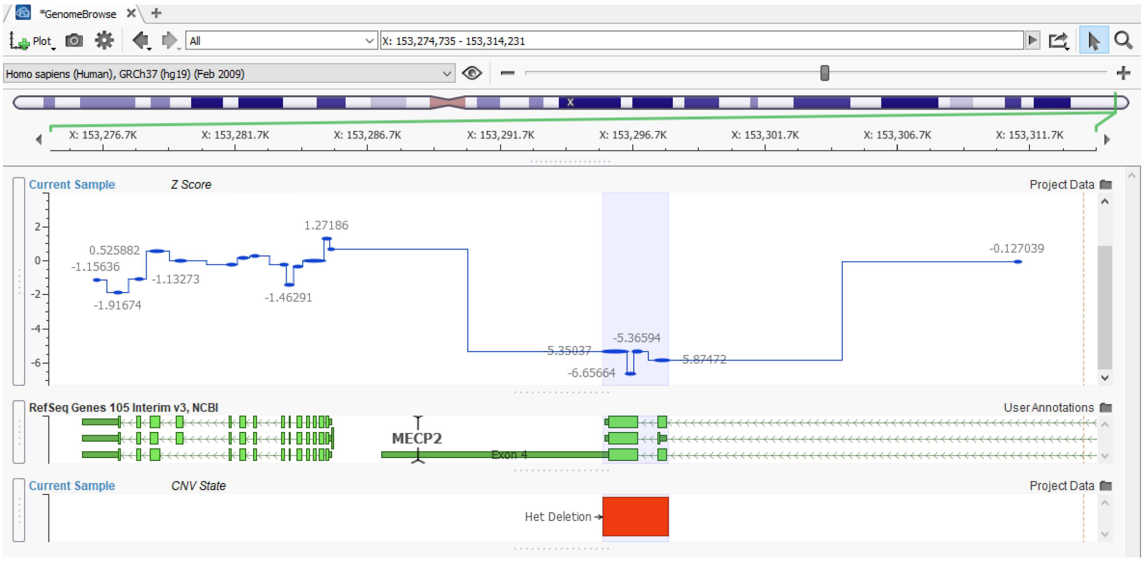


Figure 3.10 – MECP2 deletion depiction from VarSeq

analysis (Fig. 3.10) and confirmed by MLPA. Similar deletions were described in the literature on patients with Rett Syndrome Schollen et al. (2003), Zahorakova et al. (2007), Kobayashi et al. (2012) accordingly this deletion should be classified as pathogenic.

Pathogenic variants in the *MECP2* gene cause Rett Syndrome (MIM 312750) with an X-linked dominant pattern of inheritance and apparently compatible with this patient phenotype.

3.4.4 Case D - X-linked recessive gene: *STS*

Sample from a 41 years old male for whole exome analysis. The patient presented a cognitive deficit, facial dysmorphism, ichthyosis, cataracts, and epilepsy.

The hemizygous deletion NM_015506.2: c.(?_-1)_(*1_?)del, that comprises at least *STS* gene was detected by CNV analysis (Fig. 3.11) and confirmed by MLPA. This deletion was reported in the literature in a patient with ichthyosis Bonifas et al. (1987), Diociaiuti et al. (2016), most of the cases being reported with X-linked

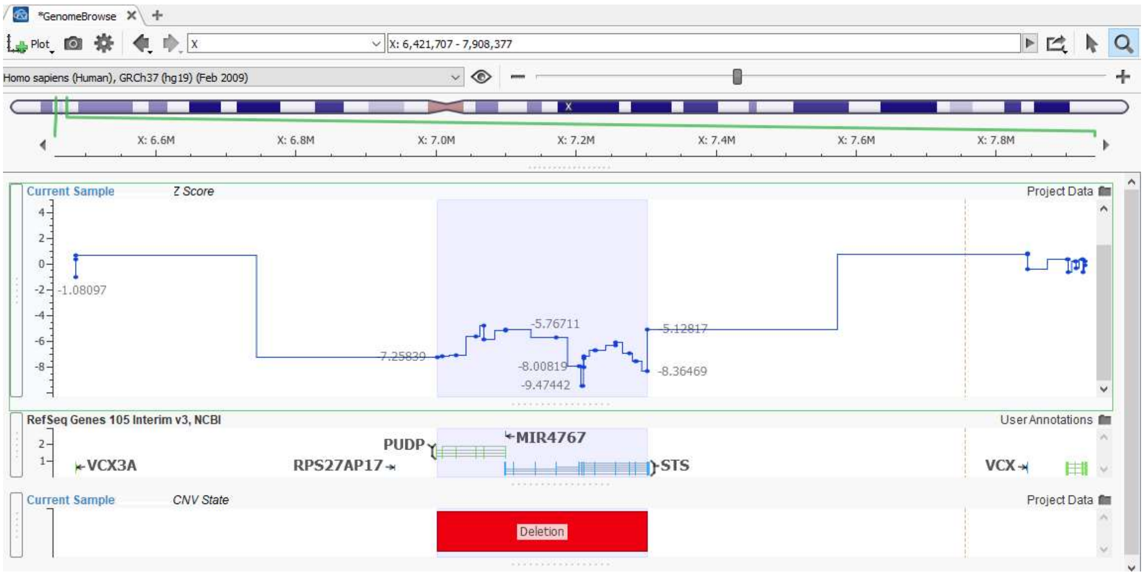


Figure 3.11 – STS deletion depiction from VarSeq

icthyosis (around 90%) presented the complete deletion of *STS* gene (Hernández-Martín et al., 1999). With the available information, this should be classified as a pathogenic variant.

Pathogenic variants in the *STS* gene cause ichthyosis with an X-linked recessive pattern of inheritance (MIM 308100) characterized by progressive cutaneous manifestations since childhood. Additionally, it was described, cases with intellectual deficit and corneal opacity.

Additionally, were detected two variants in *MMACHC* gene, with an autosomal recessive pattern of inheritance and characterized by variable phenotype and age of onset, cause methylmalonic aciduria and homocystinuria type cblC (MIM 277400), apparently compatible with this patient phenotype. Usually, this patient genotype is associated with a late onset (Morel et al., 2006), usually with neuropsychiatric alterations, progressive cognitive delay, and/or thromboembolisms, possibly justifying the facial dysmorphism and epilepsy.

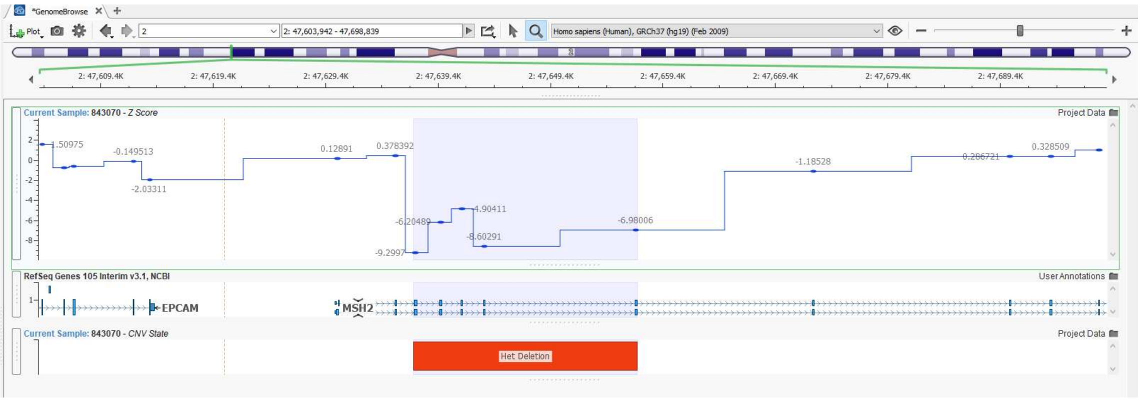


Figure 3.12 – MSH2 deletion depiction from VarSeq

3.4.5 Case E - Dominant gene: *MSH2*

Sample from a 34 years old female for oncologic analysis. The patient was diagnosed with hereditary nonpolyposis colorectal cancer and had a positive family history.

The NM_0000251.2: *c.*(366+1_367−1)_(*c.*1276+1_1277−1)del, detected in heterozygosity in the *MSH2* gene that comprises exons 3 to 7 was detected by CNV analysis (Fig. 3.12) and confirmed by MLPA.This deletion is described in the literature in patients with nonpolyposis colorectal cancer (Papp et al. (2007), Alqahtani et al. (2018)). With the available information, this should be classified as a pathogenic variant. This result was confirmed by MLPA.

Pathogenic variants in the *MSH2* gene cause hereditary nonpolyposis colorectal cancer type 1 (MIM 120435), with an autosomal dominant pattern of inheritance.

3.4.6 Case F - Dominant gene: *BRCA1*

Sample from a 64 years old female for oncologic analysis. The patient was diagnosed at 62 years old with serous high-grade bilateral ovarian cancer and had a positive family history for breast and pancreatic cancer.

The NM_0007294.3: $c.(547 + 1_548 - 1)_(4185 + 1_4186 - 1)$dup, detected in heterozygosity in the *BRCA1* gene that comprises exons 8 to 11 was detected by CNVs

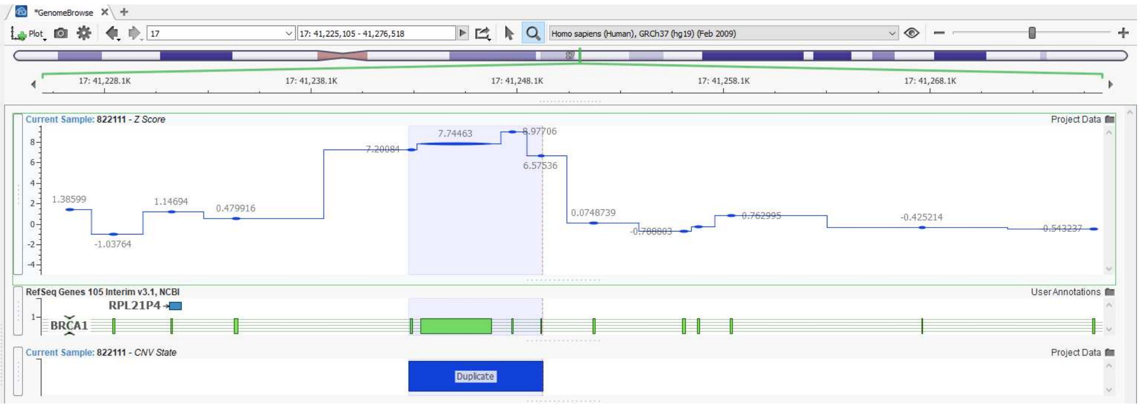


Figure 3.13 – BRCA1 duplication depiction from VarSeq

analysis (Fig. 3.13) and confirmed by MLPA. This duplication is described in the literature in patients with oviduct cancer (Arnold et al., 2014). With the available information, this should be classied as a likely pathogenic variant.

4 Discussion

With this study it was possible to define applicable thresholds that can lead to high sensitivity and specificity: 99.15% and 98.85% for oncology panels and 87.50% and 90.30% for WES when calculating with the total values, 94.44% and 99.29% when the CNVs with only one target are removed from the calculation. These metrics are similar to those obtained by Fortier et al. (2018), where they achieve a sensitivity of 97.6% for cancer panel with the same bioinformatic tool. Furthermore, Iacocca et al. (2017b) had successfully used this tool obtaining 100% sensitivity when compared to the golden standard (MLPA). This was due to the fact that their minimum threshold was defined as 300bp. If we apply the same principle to our exome and oncologic samples, we would obtain the same result. The sensitivity obtained during this work was considerably higher than the ones obtained with other softwares. For example, CoNIFER has a value around 84% for cancer panels with some studies reporting 14.6% for exomes. XHMM achieves 22.2% with some studies with 92.6% for cancer panels. Also CNVnator ranges from 87.7% to 96% (Fortier et al. (2018), Yao et al. (2017), Abyzov et al. (2011)).

Despite the results, there are some limitations to this method. When the alteration involves only part of the exon, or is small, or involves the first exon, or has a high GC content, the software has some problems identifying the alteration correctly. In fact, the first exon of several genes is usually associated with a high CG-content that makes the amplification and sequencing more difficult. Some studies refer the ideal GC content is 30% (Shen et al. (2019), Roca et al. (2019)). Another disadvantage is the sequences with pseudogenes such as the *PMS2* gene. Some studies exclude all the genes with pseudogenes of the analysis since they consider that NGS is not a reliable method to detect CNVs in these genes (Mu et al., 2018). The last limitation would be the results expected with lower quality samples (prenatal, blood from hematologic patients, and highly degraded DNA) that generate lower quality NGS data. In these cases, CNV analysis may not be possible to perform due to the total number of calls being too elevated and having too many false positives.

It was detected two large mosaicism events with VarSeq in oncologic samples. It is not known the result from smaller mosaicism alterations. To detect this type of mosaicism it would be probably needed greater coverage (Baert-Desurmont et al., 2018). The duplicated alterations are the most challenging to detect according to Newman et al. (2015) and to the results obtained once that were detected few duplications and small duplications like *BRCA1* were not detected. Furthermore, false positives like *MSH6* had borderline values.

In spite of the known limitations, including CNVs in the method allows the increase of the diagnostic yield. It was achieved 2.54% for cancer panels and 8.15% for WES in our sample. These values are higher than what has been previously obtained by others. For example, Pfundt et al. (2016) achieved a mean yield of 2% for exome analysis but lower compared to Iacocca et al. (2017a) achieving almost 10% for familial hypercholesterolemia or Conrad et al. (2009) that achieves 5%. These differences might be explained by the fact that different diseases have different mutational mechanisms besides this work has a small cohort that can bias the results.

5 Conclusions

There have been several studies trying to evaluate the behavior of CNV detection algorithms. It has been proved that CNV analysis is a good approach to the identification of genomic alterations. However, the specificity and sensitivity are highly dependent on the algorithm used to do the prevision affecting the false positives and false negatives rate. CNV accurate detection is possible, with a sensitivity of 99.15% and specificity of 98.85%, for cancer and sensitivity of 87.50% and specificity of 90.30% for WES considering the total number of samples. The sensitivity is heightened when restricting to alterations bigger than 300pb. For this threshold, the sensitivity increases for up to 7%.

This detection has its limitations like when it is involved only part of the exon, or this is small, or involves the first exon, or has a high GC content, sequences with pseudogenes, and the results expected from lower quality samples (prenatal, blood from hematologic patients, and highly degraded DNA) are problematic too.

Copy number variants can accurately be detected and increase diagnostic yield by $3-8\%$, which is relevant for clinical management and genetic counseling to patients and their relatives. This study proves the potential of NGS as a reliable and affordable method to detect CNVs, both in target panel (as cancer panels) and WES.

Next-generation DNA sequencing use is primarily SNVs and small indels. However, it can also be used to detect CNVs and provide a valuable addition to the information retrieved from the data. This methodology allowed to increase the number of patients with a positive result and without this technic it could take years to find out the causative alteration unless the physician had a very strong suspicion of a gene-related disorder. The results from this study are promising but the usage of other algorithms simultaneously, as well as long read sequencing, might be used to improve the CNVs detection performance from NGS data.

Bibliography

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984. 33

Adam, M. P., Ardinger, H. H., Pagon, R. A., Wallace, S. E., Bean, L. J. H., Stephens, K., and Amemiya, A. (1993). Genereviews. 26

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–76. xvi, 3, 4, 5, 6, 7, 8

Alqahtani, M., Edwards, C., Buzzacott, N., Carpenter, K., Alsaleh, K., Alsheikh, A., Abozeed, W., Mashhour, M., Almousa, A., Housawi, Y., Al Hawwaj, S., and Iacopetta, B. (2018). Screening for lynch syndrome in young saudi colorectal cancer patients using microsatellite instability testing and next generation sequencing. *Familial cancer*, 17:197–203. 31

Arnold, A. G., Otegbeye, E., Fleischut, M. H., Glogowski, E. A., Siegel, B., Boyar, S. R., Salo-Mullen, E., Amoroso, K., Sheehan, M., Berliner, J. L., Stadler, Z. K., Kauff, N. D., Offit, K., Robson, M. E., and Zhang, L. (2014). Assessment of individuals with brca1 and brca2 large rearrangements in high-risk breast and ovarian cancer families. *Breast cancer research and treatment*, 145:625–634. 32

Baert-Desurmont, S., Coutant, S., Charbonnier, F., Macquere, P., Lecoquierre, F., Schwartz, M., Blanluet, M., Vezain, M., Lanos, R., Quenez, O., Bou, J., Bouvignies, E., Fourneaux, S., Manase, S., Vasseur, S., Mauillon, J., Gerard, M., Marlin, R., Bougeard, G., Tinat, J., Frebourg, T., and Tournier, I. (2018). Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. *European Journal of Human Genetics*, 26(11):1597–1602. 34

Bonberg, N., Pesch, B., Behrens, T., Johnen, G., Taeger, D., Gawrych, K., Schwentner, C., Wellhäußer, H., Kluckert, M., Leng, G., Nasterlack, M., Oberlinner, C., Stenzl, A., and Brüning, T. (2014). Chromosomal alterations in exfoliated urothelial cells from bladder cancer cases and healthy men: a prospective screening study. *BMC Cancer*, 14(1). 3

Bonifas, J. M., Morley, B. J., Oakey, R. E., Kan, Y. W., and Epstein, E. H. (1987). Cloning of a cdna for steroid sulfatase: frequent occurrence of gene deletions in patients with recessive x chromosome-linked ichthyosis. *Proceedings of the National Academy of Sciences of the United States of America*, 84:9248–9251. 29

Bravo-Gil, N., Méndez-Vidal, C., Romero-Pérez, L., González-del Pozo, M., Rodríguez-de la Rúa, E., Dopazo, J., Borrego, S., and Antiñolo, G. (2016). Improving the management of inherited retinal dystrophies by targeted sequencing of a population-specific gene panel. *Scientific reports*, 6:23910. 26

Caloba, M., Silva, V., Cerqueira, R., and Pinto Basto, J. (2020). P12-copy number variations analysis of ngs data in germline oncology testing, proceedings of the 23rd annual meeting of the portuguese society of human genetics. *Medicine*, 99(9):e19291. xvii, 60

CGC (2019). Retrieved on 02/02/2019 from www.cgcgenetics.com/pt/sobre-o-cgc.

Conrad, D. F., , Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2009). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712. 1, 34

de Ligt, J., Boone, P. M., Pfundt, R., Vissers, L. E., Richmond, T., Geoghegan, J., O'Moore, K., de Leeuw, N., Shaw, C., Brunner, H. G., Lupski, J. R., Veltman, J. A., and Hehir-Kwa, J. Y. (2013). Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat*, 34(10):1439–48. 6

Despierre, E., Moisse, M., Yesilyurt, B., Sehouli, J., Braicu, I., Mahner, S., Castillo-Tong, D. C., Zeillinger, R., Lambrechts, S., Leunen, K., Amant, F., Moerman, P., Lambrechts, D., and Vergote, I. (2014). Somatic copy number alterations predict response to platinum therapy in epithelial ovarian cancer. *Gynecologic Oncology*, 135(3):415–422. 2

Diociaiuti, A., El Hachem, M., Pisaneschi, E., Giancristoforo, S., Genovese, S., Sirleto, P., Boldrini, R., and Angioni, A. (2016). Role of molecular testing in the multidisciplinary diagnostic approach of ichthyosis. *Orphanet journal of rare diseases*, 11:4. 29

Fakhro, K. A., Yousri, N. A., Rodriguez-Flores, J. L., Robay, A., Staudt, M. R., Agosto-Perez, F., Salit, J., Malek, J. A., Suhre, K., Jayyousi, A., Zirie, M., Stadler, D., Mezey, J. G., and Crystal, R. G. (2015). Copy number variations in the genome of the qatari population. *BMC Genomics*, 16:834. 1, 2

Foged, N. T., Brügmann, A., and Jørgensen, J. T. (2013). The her2 cish pharmdx™ kit in the assessment of breast cancer patients for anti-her2 treatment. *Expert Review of Molecular Diagnostics*, 13(3):233–242. 3

Fortier, N., Rudy, G., and Scherer, A. (2018). Detection of CNVs in NGS data using VS-CNV. In *Methods in Molecular Biology*, pages 115–127. Springer New York. 33

Gachet, Y., Codlin, S., Hyams, J. S., and Mole, S. E. (2005). btn1, the schizosaccharomyces pombe homologue of the human batten disease gene cln3, regulates vacuole homeostasis. *Journal of cell science*, 118:5525–5536. 27

Guo, Y., Zhao, S., Lehmann, B. D., Sheng, Q., Shaver, T. M., Stricker, T. P., Pietenpol, J. A., and Shyr, Y. (2014). Detection of internal exon deletion with exon del. *BMC Bioinformatics*, 15(1). 4, 7, 8

Haines, R. L., Codlin, S., and Mole, S. E. (2009). The fission yeast model for the lysosomal storage disorder batten disease predicts disease severity caused by mutations in cln3. *Disease models and mechanisms*, 2:84–92. 27, 28

Hehir-Kwa, J. Y., Pfundt, R., and Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert Rev Mol Diagn*, 15(8):1023–32. 1, 2

Hehir-Kwa, J. Y., Tops, B. B. J., and Kemmeren, P. (2018). The clinical implementation of copy number detection in the age of next-generation sequencing. *Expert Review of Molecular Diagnostics*, 18(10):907–915. 4

HengWang, D. and KaiYing (2014). Copy number variation detection using next generation sequencing read counts. *BMCBioinformatics*, 15(109). 1, 4

Hernández-Martín, A., González-Sarmiento, R., and De Unamuno, P. (1999). X-linked ichthyosis: an update. *The British journal of dermatology*, 141:617–627. 30

Horpaopan, S., Spier, I., Zink, A. M., Altmüller, J., Holzapfel, S., Laner, A., Vogt, S., Uhlhaas, S., Heilmann, S., Stienen, D., Pasternack, S. M., Keppler, K., Adam, R., Kayser, K., Moebus, S., Draaken, M., Degenhardt, F., Engels, H., Hofmann, A., Nöthen, M. M., Steinke, V., Perez-Bouza, A., Herms, S., Holinski-Feder, E., Fröhlich, H., Thiele, H., Hoffmann, P., and Aretz, S. (2014). Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *International Journal of Cancer*, 136(6):E578–E589. 3

Iacocca, M., Wang, J., Dron, J., Robinson, J., Mcintyre, A., Cao, H., and Hegele, R. (2017a). Use of next-generation sequencing to detect ldlr gene copy number variation in familial hypercholesterolemia. *Journal of lipid research*, 58. 34

Iacocca, M. A., Wang, J., Dron, J. S., Robinson, J. F., McIntyre, A. D., Cao, H., and Hegele, R. A. (2017b). Use of next-generation sequencing to detectLDLRgene copy number variation in familial hypercholesterolemia. *Journal of Lipid Research*, 58(11):2202–2209. 33

Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F., South, S. T., and Working Group of the American College of Medical Genetics Laboratory Quality Assurance, C. (2011). American college of medical genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med*, 13(7):680–5. 2

Klug, W. (2012). *Concepts of genetics*. Pearson Education, San Francisco. xvi, 1, 2

Kobayashi, Y., Ohashi, T., Akasaka, N., and Tohyama, J. (2012). Congenital variant of rett syndrome due to an intragenic large deletion in mecp2. *Brain and development*, 34:601–604. 29

Kousi, M., Lehesjoki, A.-E., and Mole, S. E. (2012). Update of the mutation spectrum and clinical correlations of over 360 mutations in eight genes that underlie the neuronal ceroid lipofuscinoses. *Human mutation*, 33:42–63. 27

Leary, R. J., Lin, J. C., Cummins, J., Boca, S., Wood, L. D., Parsons, D. W., Jones, S., Sjoblom, T., Park, B.-H., Parsons, R., Willis, J., Dawson, D., Willson, J. K. V., Nikolskaya, T., Nikolsky, Y., Kopelovich, L., Papadopoulos, N., Pennacchio, L. A., Wang, T.-L., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proceedings of the National Academy of Sciences*, 105(42):16224–16229. 2

Legault, M. A., Girard, S., Lemieux Perreault, L. P., Rouleau, G. A., and Dube, M. P. (2015). Comparison of sequencing based cnv discovery methods using monozygotic twin quartets. *PLoS One*, 10(3):e0122287. 1, 4, 5, 6

Marcinkowska-Swojak, M., Uszczynska, B., Figlerowicz, M., and Kozlowski, P.

(2013). An mlpa-based strategy for discrete cnv genotyping: Cnv-mirnas as an example. *Hum Mutat*, 34(5):763–73. 1, 3

Martin-Merida, I., Aguilera-Garcia, D., Fernandez-San, J. P., Blanco-Kelly, F., Zurita, O., Almoguera, B., Garcia-Sandoval, B., Avila-Fernandez, A., Arteche, A., Minguez, P., Carballo, M., Corton, M., and Ayuso, C. (2018). Toward the mutational landscape of autosomal dominant retinitis pigmentosa: A comprehensive analysis of 258 spanish families. *Investigative ophthalmology and visual science*, 59:2345–2354. 26

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46. 4

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., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., Watson, M. S., Martin, C. L., and Ledbetter, D. H. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet*, 86(5):749–64. 3

Morel, C. F., Lerner-Ellis, J. P., and Rosenblatt, D. S. (2006). Combined methylmalonic aciduria and homocystinuria (cblc): phenotype-genotype correlations and ethnic-specific observations. *Molecular genetics and metabolism*, 88:315–321. 30

MT: Golden Helix, I. Bozeman, varseq^{TM} (version 2.1.2) [software]. *http://www.goldenhelix.com.*

Mu, W., Li, B., Wu, S., Chen, J., Sain, D., Xu, D., Black, M. H., Karam, R., Gillespie, K., Hagman, K. D. F., Guidugli, L., Pronold, M., Elliott, A., and Lu, H.-M. (2018). Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. *Genetics in Medicine*, 21(7):1603–1610. 34

Munroe, P. B., Mitchison, H. M., O'Rawe, A. M., Anderson, J. W., Boustany, R. M., Lerner, T. J., Taschner, P. E., de Vos, N., Breuning, M. H., Gardiner, R. M., and Mole, S. E. (1997). Spectrum of mutations in the batten disease gene, cln3. *American journal of human genetics*, 61:310–316. 27

Newman, S., Hermetz, K. E., Weckselblatt, B., and Rudd, M. K. (2015). Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *The American Journal of Human Genetics*, 96(2):208–220. 34

Papp, J., Kovacs, M. E., and Olah, E. (2007). Germline mlh1 and msh2 mutational spectrum including frequent large genomic aberrations in hungarian hereditary non-polyposis colorectal cancer families: implications for genetic testing. *World journal of gastroenterology*, 13:2727–2732. 31

Pfundt, R., Rosario, M., Vissers, L., Kwint, M., Janssen, I., Leeuw, N., Yntema, H., Nelen, M., Lugtenberg, D., Kamsteeg, E.-J., Wieskamp, N., Stegmann, A., Stevens, S., Rodenburg, R., Simons, A., Mensenkamp, A., Rinne, T., Gilissen, C., Scheffer, H., and Hehir-Kwa, J. (2016). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in medicine : official journal of the American College of Medical Genetics*, 19. 34

Pierce, B. A. (2012). *Genetics Conceptual Approach*. W. H. Freeman and Company, 4th edition. 1, 2

Pirooznia, M., Goes, F. S., and Zandi, P. P. (2015). Whole-genome cnv analysis: advances in computational approaches. *Front Genet*, 6:138. 1, 2

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., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–54. 1

Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., Friez, M. J., Funke, B. H., Hegde, M. R., Lyon, E., of the American College of Medical Genetics, W. G., and Commitee, G. L. Q. A. (2013). Acmg clinical laboratory standards for next-generation sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15:733–747. 13

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., and Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17:405–424. 13

Riggs, E. R., , Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S., and Martin, C. L. (2019). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genetics in Medicine*, 22(2):245–257. 3, 13

Roca, I., Gonzalez-Castro, L., Maynou, J., Palacios, L., Fernandez, H., Couce, M. L., and Fernandez-Marmiesse, A. (2019). Pattrec: An easy-to-use cnv detection tool optimized for targeted ngs assays with diagnostic purposes. *Genomics*. 3, 4, 6, 34

Sadedin, S. P., Ellis, J. A., Masters, S. L., and Oshlack, A. (2018). Ximmer: a system for improving accuracy and consistency of cnv calling from exome data. *GigaScience*, 7(10). 4

Schollen, E., Smeets, E., Deflem, E., Fryns, J. P., and Matthijs, G. (2003). Gross rearrangements in the mecp2 gene in three patients with rett syndrome: implications for routine diagnosis of rett syndrome. *Human mutation*, 22:116–120. 29

Shen, W., Szankasi, P., Durtschi, J., Kelley, T. W., and Xu, X. (2019). Genome-wide copy number variation detection using NGS: Data analysis and interpretation. In *Methods in Molecular Biology*, pages 113–124. Springer New York. 34

Silveira, S. M., da Cunha, I. W., Marchi, F. A., Busso, A. F., Lopes, A., and Rogatto, S. R. (2014). Genomic screening of testicular germ cell tumors from monozygotic twins. *Orphanet Journal of Rare Diseases*, 9(1). 2

Sinha, R., Samaddar, S., and De, R. K. (2015). Cnv-ch: A convex hull based segmentation approach to detect copy number variations (cnv) using next-generation sequencing data. *PLoS One*, 10(8):e0135895. 4, 5, 6

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A. S., and Zhu, M. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, 35(7):899–907. 8

Tzeng, J. Y., Magnusson, P. K., Sullivan, P. F., Swedish Schizophrenia, C., and Szatkiewicz, J. P. (2015). A new method for detecting associations with rare copy-number variants. *PLoS Genet*, 11(10):e1005403. 1, 2

Valsesia, A., Mace, A., Jacquemont, S., Beckmann, J. S., and Kutalik, Z. (2013). The growing importance of cnvs: New insights for detection and clinical interpretation. *Front Genet*, 4:92. 2, 3, 4, 5, 6, 7

Weckselblatt, B. and Rudd, M. K. (2015). Human structural variation: Mechanisms of chromosome rearrangements. *Trends in Genetics*, 31(10):587–599. 1, 3, 4

Xu, H., Zhu, X., Xu, Z., Hu, Y., Bo, S., Xing, T., and Zhu, K. (2015). Non-invasive analysis of genomic copy number variation in patients with hepatocellular carcinoma by next generation DNA sequencing. *Journal of Cancer*, 6(3):247–253. 2

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., Wang, J., and Shen, Y. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*, 10(1). 33

Zahorakova, D., Rosipal, R., Hadac, J., Zumrova, A., Bzduch, V., Misovicova, N., Baxova, A., Zeman, J., and Martasek, P. (2007). Mutation analysis of the mecp2 gene in patients of slavic origin with rett syndrome: novel mutations and polymorphisms. *Journal of human genetics*, 52:342–348. 29

Zare, F., Dow, M., Monteleone, N., Hosny, A., and Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, 18(1). 8

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013a). Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14 Suppl 11:S1. 1, 2, 3, 5, 7, 8

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013b). Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(S11). 4, 6, 8

A Appendix

Table A.1 – CNV total number detected with flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	26	5	7	9	22	39	100	228	551	1249	2236
B	8	3	5	11	31	80	199	422	787	1558	3104
C	33	11	14	21	42	125	226	369	602	1311	2754
D	22	7	7	13	26	42	94	187	450	1315	2163
E	12	5	5	13	15	44	73	200	482	1322	2171
Mean	20	6	8	13	27	66	138	281	574	1351	2486
Totals	20	26	34	47	75	141	279	560	1135	2486	

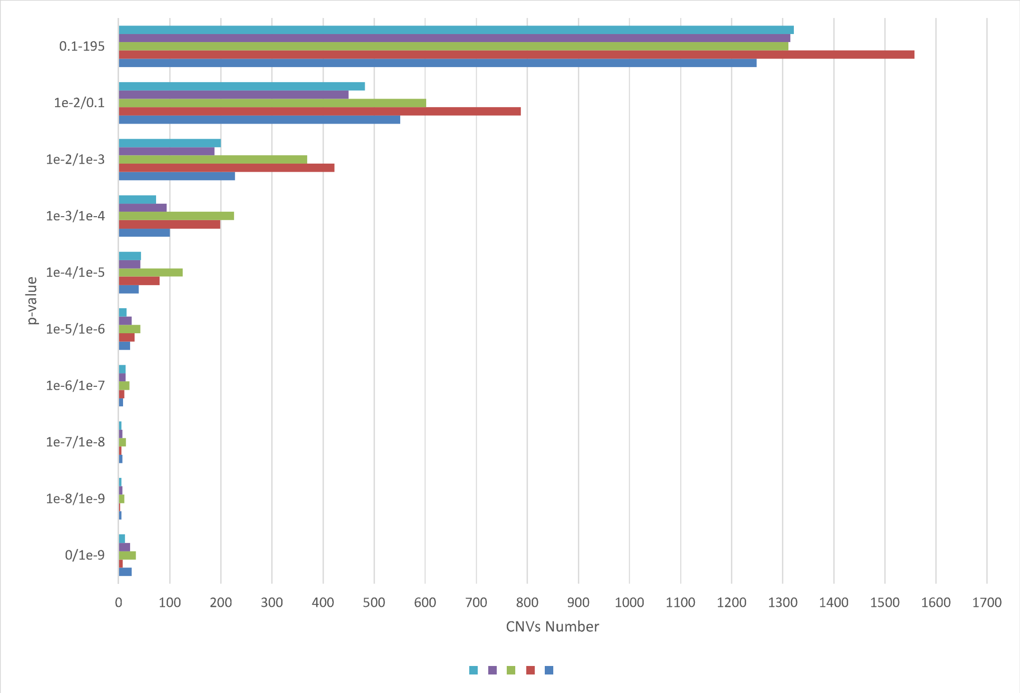


Figure A.1 – CNV total distribution by p-value with flags

Table A.2 – CNV number with HPO detected with flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	2	0	0	0	0	0	2	5	8	13	30
B	3	1	1	0	2	4	14	25	42	20	112
C	1	0	2	0	3	12	8	22	21	18	87
D	0	1	0	0	0	2	9	16	14	33	75
E	0	1	0	0	0	4	5	8	31	34	83
Mean	1	1	1	0	1	4	8	15	23	24	77
Totals	1	2	2	2	3	8	15	31	54	77	

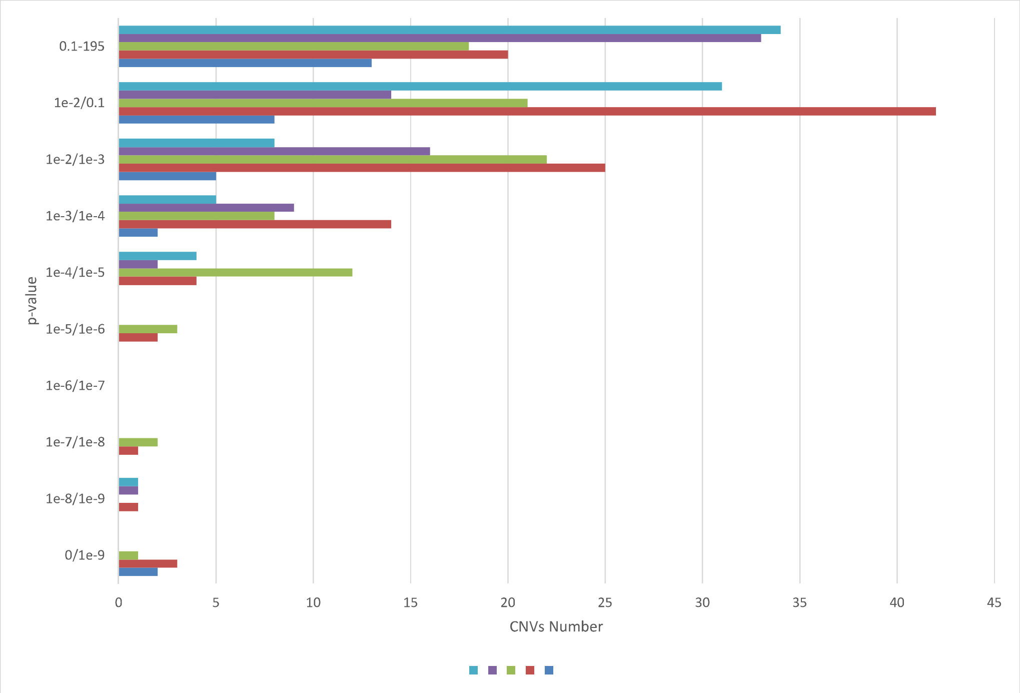


Figure A.2 – CNV distribution with HPO by p-value with flags

Table A.3 – CNV number without HPO detected with flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	24	5	7	9	22	39	98	223	543	1236	2206
B	5	2	4	11	29	76	185	397	745	1538	2992
C	32	11	12	21	39	113	218	347	581	1293	2667
D	22	6	7	13	26	40	85	171	436	1282	2088
E	12	4	5	13	15	40	68	192	451	1288	2088
Mean	19	6	7	13,4	26	62	131	266	551	1327	2408
Totals	19	25	32	45	71	133	264	530	1081	2408	

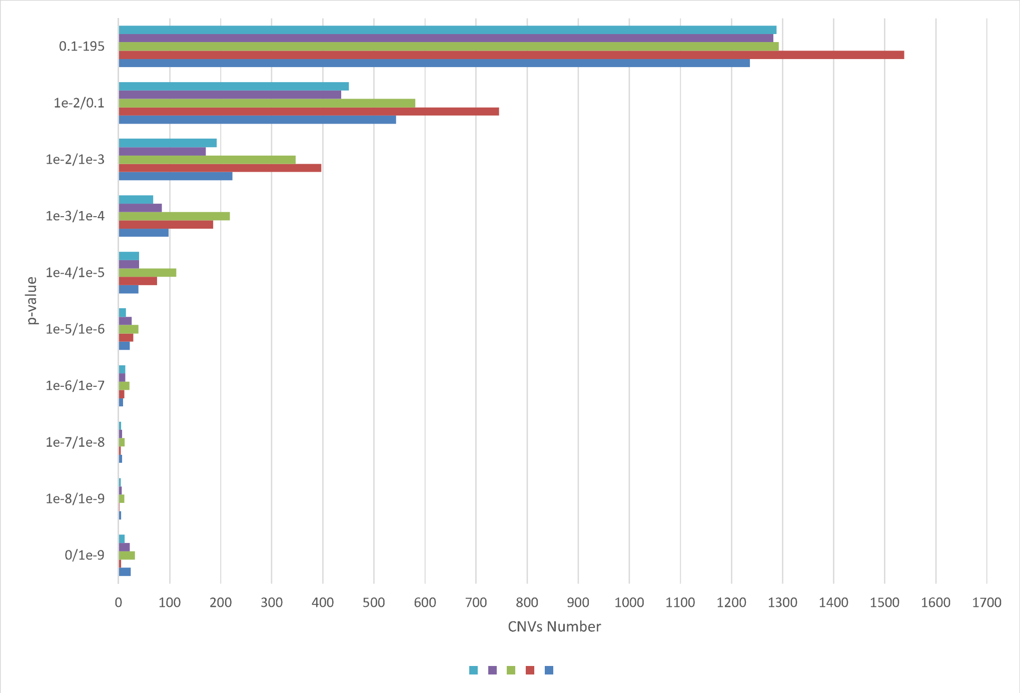


Figure A.3 – CNV distribution without HPO by p-value with flags

Table A.4 – CNV total number detected without flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	8	3	2	5	7	11	19	30	3	16	104
B	6	1	2	5	18	48	114	170	57	12	433
C	15	6	7	10	15	44	99	93	28	10	327
D	17	1	4	2	8	16	33	62	10	18	171
E	5	1	1	6	6	20	23	56	19	16	153
Mean	10,2	2,4	3,2	5,6	10,8	27,8	57,6	82,2	23,4	14,4	237,6
Totals	10,2	12,6	15,8	21,4	32,2	60	117,6	199,8	223,2	237,6	

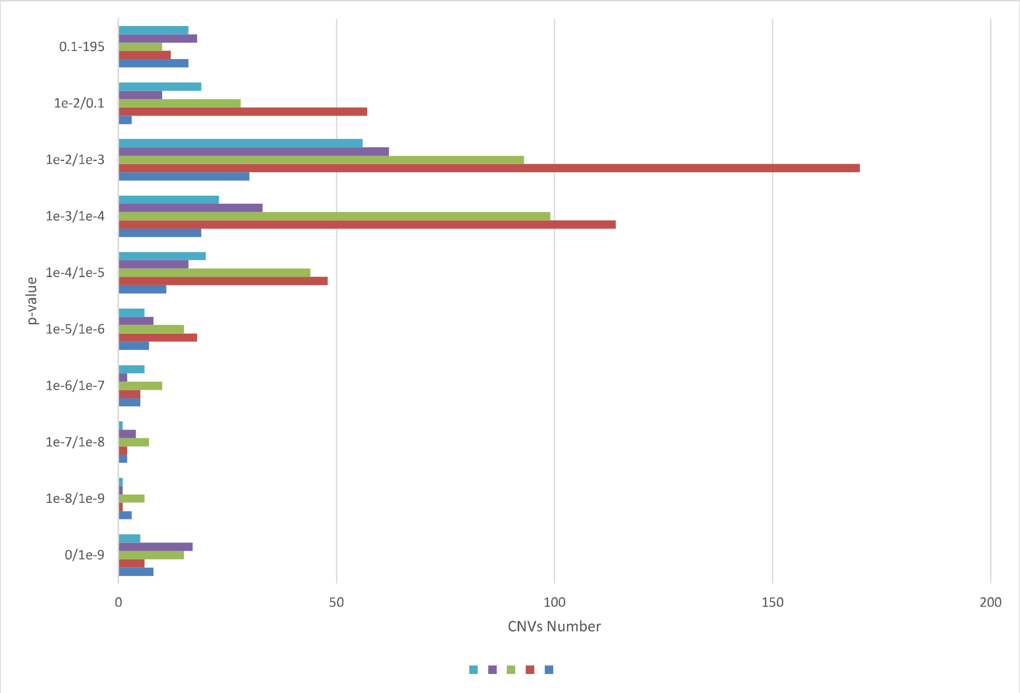


Figure A.4 – CNV total distribution by p-value with flags

Table A.5 – CNV number with HPO detected without flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	0	1	0	0	0	0	2	5	0	0	8
B	2	1	0	0	2	3	9	10	6	1	34
C	1	0	1	0	1	5	3	5	0	0	16
D	1	0	0	0	0	0	1	3	0	0	5
E	0	1	0	0	0	2	4	2	2	0	11
Mean	0,8	0,6	0,2	0	0,6	2	3,8	5	1,6	0,2	14,8
Totals	0,8	1,4	1,6	1,6	2,2	4,2	8	13	14,6	14,8	

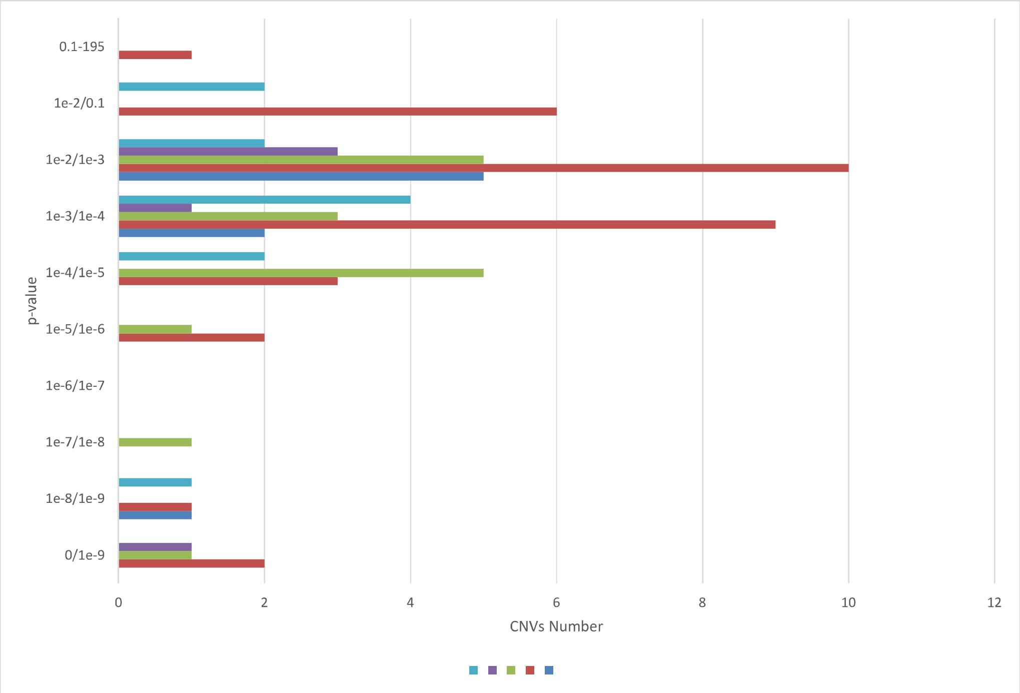


Figure A.5 – CNV distribution with HPO by p-value without flags

Table A.6 – CNV number without HPO detected without flags among p-value distribution

Class	1	2	3	4	5	6	7	8	9	10	
Study	0 10^{-9}	10^{-9} 10^{-8}	10^{-8} 10^{-7}	10^{-7} 10^{-6}	10^{-6} 10^{-5}	10^{-5} 10^{-4}	10^{-4} 10^{-3}	10^{-3} 10^{-2}	10^{-2} 10^{-1}	10^{-1} 195	Total
A	8	2	2	5	7	11	17	25	3	16	96
B	4	0	2	5	16	45	105	160	51	11	399
C	14	6	6	10	14	39	96	88	28	10	311
D	16	1	4	2	8	16	32	59	10	18	166
E	5	0	1	6	6	18	19	54	17	16	142
Mean	9,4	1,8	3	5,6	10,2	25,8	53,8	77,2	21,8	14,2	222,8
Totals	9,4	11,2	14,2	19,8	30	55,8	109,6	186,8	208,6	222,8	

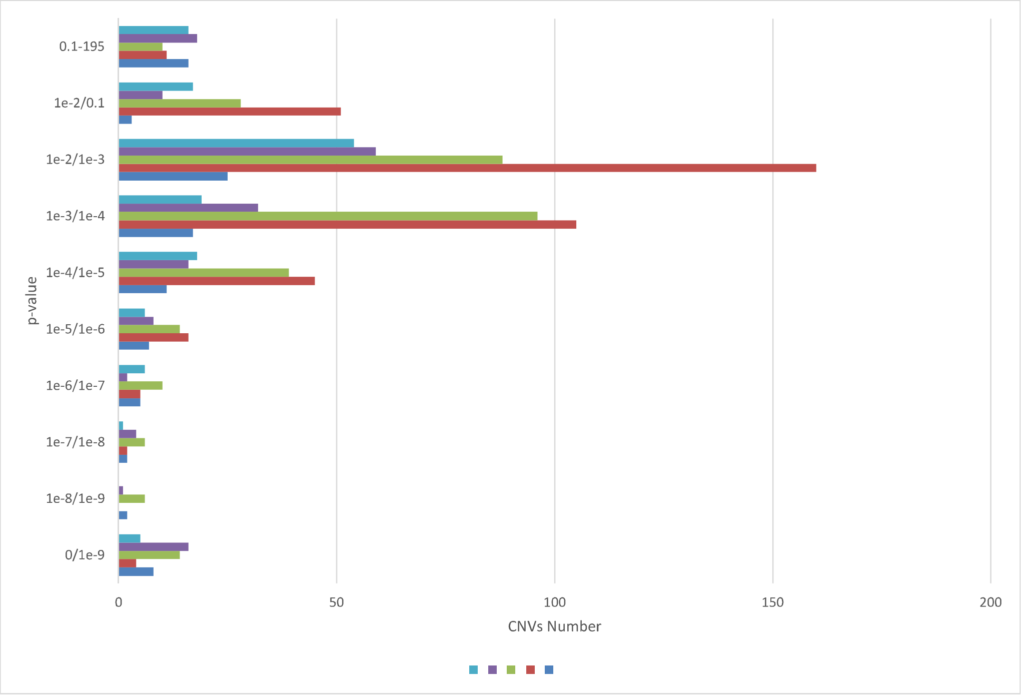
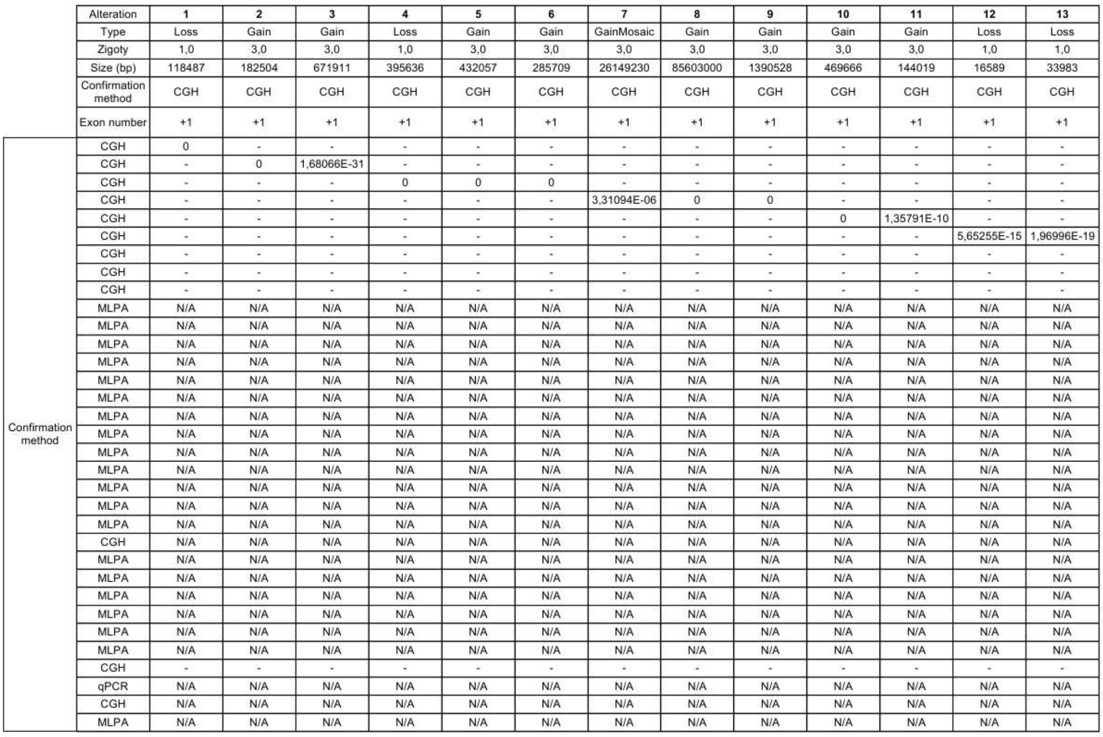


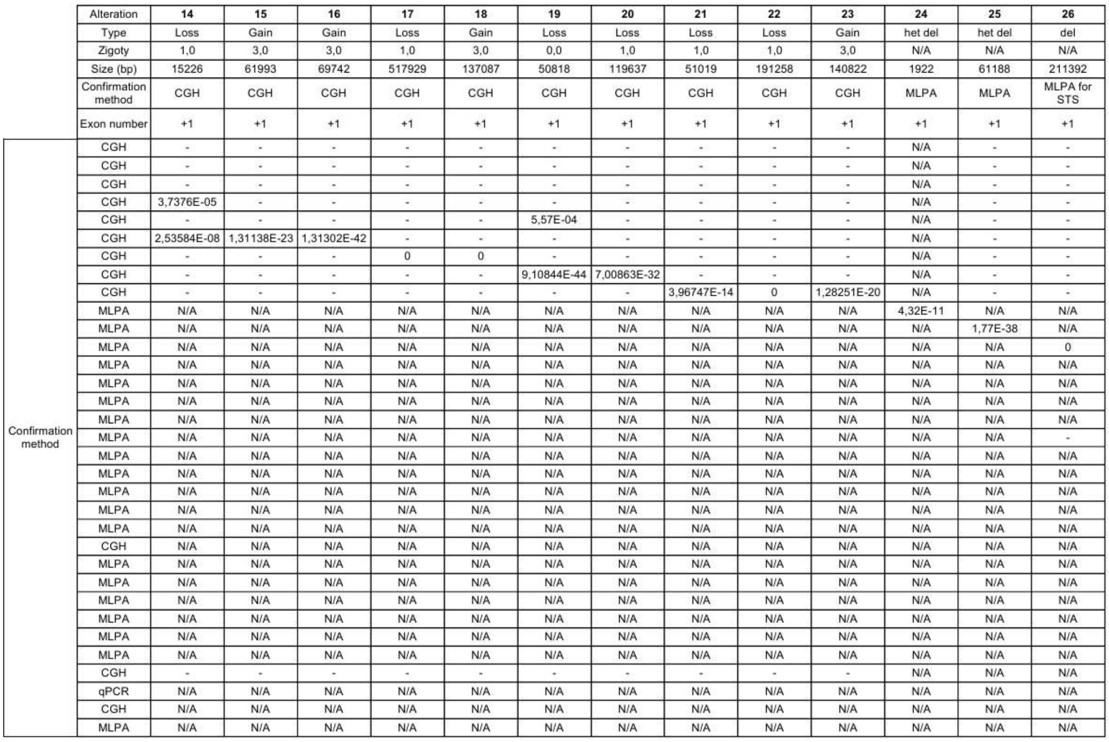
Figure A.6 – CNV without HPO distribution by p-value without flags

Table A.7 – Results gathered from WES confirmed CNVs and their respective p - value with the respective information



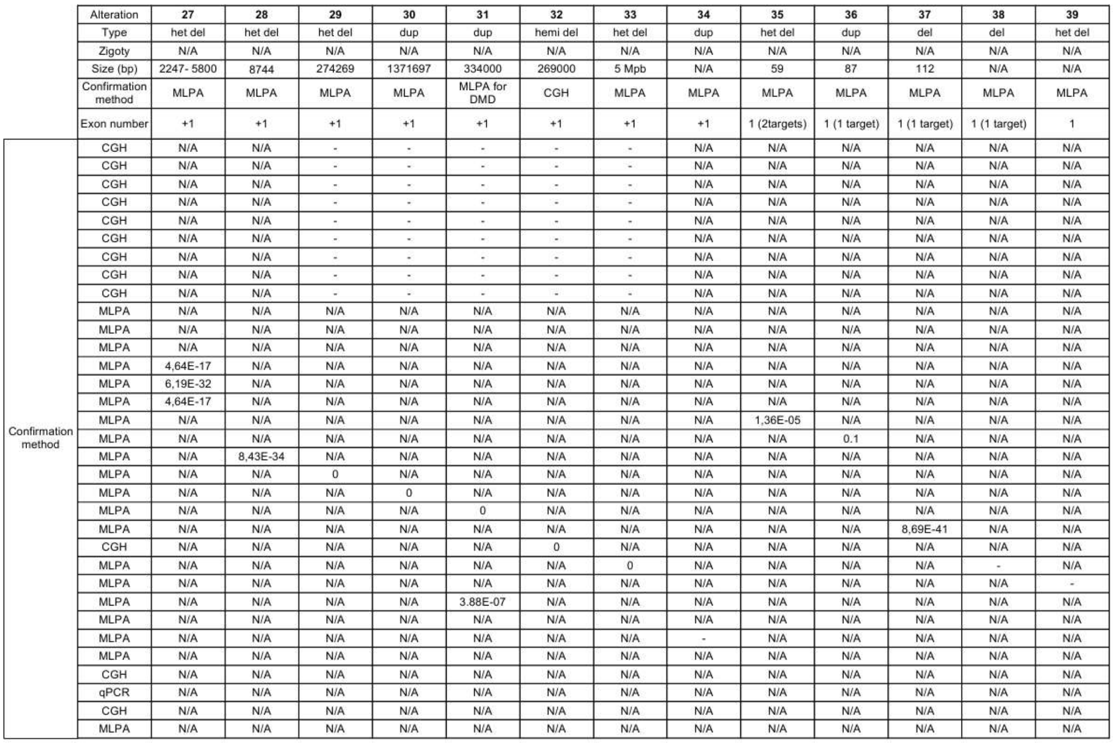
	Alteration	1	2	3	4	5	6	7	8	9	10	11	12	13
	Type	Loss	Gain	Gain	Loss	Gain	Gain	GainMosaic	Gain	Gain	Gain	Gain	Loss	Loss
	Zigoty	1,0	3,0	3,0	1,0	3,0	3,0	3,0	3,0	3,0	3,0	3,0	1,0	1,0
	Size (bp)	118487	182504	671911	395636	432057	285709	26149230	85603000	1390528	469666	144019	16589	33983
	Confirmation method	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH
	Exon number	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
Confirmation method	CGH	0	-	-	-	-	-	-	-	-	-	-	-	-
	CGH	-	0	1,68066E-31	-	-	-	-	-	-	-	-	-	-
	CGH	-	-	-	0	0	0	-	-	-	-	-	-	-
	CGH	-	-	-	-	-	-	3,31094E-06	0	0	-	-	-	-
	CGH	-	-	-	-	-	-	-	-	-	0	1,35791E-10	-	-
	CGH	-	-	-	-	-	-	-	-	-	-	-	5,65255E-15	1,96996E-19
	CGH	-	-	-	-	-	-	-	-	-	-	-	-	-
	CGH	-	-	-	-	-	-	-	-	-	-	-	-	-
	CGH	-	-	-	-	-	-	-	-	-	-	-	-	-
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	-	-	-	-	-	-	-	-	-	-	-	-	-
	qPCR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table A.8 – Continuation of table A.7



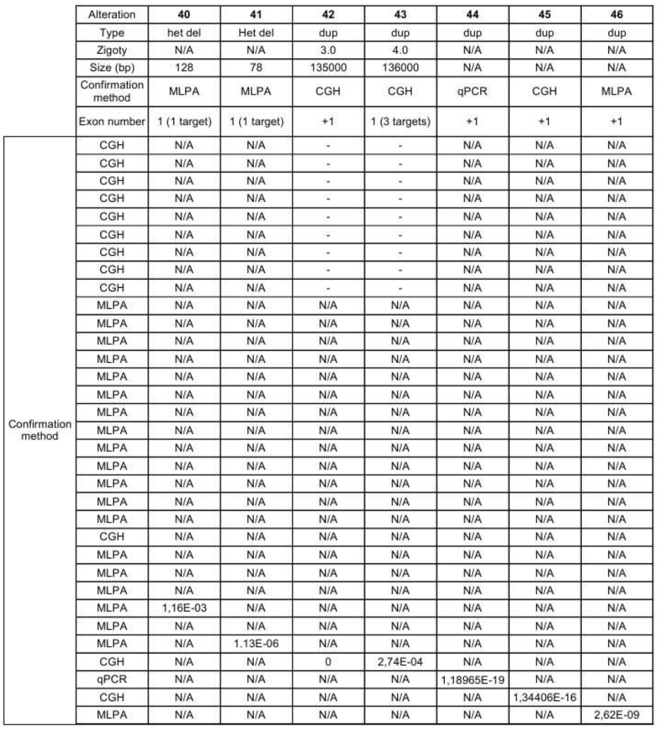
	Alteration	14	15	16	17	18	19	20	21	22	23	24	25	26
	Type	Loss	Gain	Gain	Loss	Gain	Loss	Loss	Loss	Loss	Gain	het del	het del	del
	Zigoty	1,0	3,0	3,0	1,0	3,0	0,0	1,0	1,0	1,0	3,0	N/A	N/A	N/A
	Size (bp)	15226	61993	69742	517929	137087	50818	119637	51019	191258	140822	1922	61188	211392
	Confirmation method	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	CGH	MLPA	MLPA	MLPA for STS
	Exon number	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
Confirmation method	CGH	-	-	-	-	-	-	-	-	-	-	N/A	-	-
	CGH	-	-	-	-	-	-	-	-	-	-	N/A	-	-
	CGH	-	-	-	-	-	-	-	-	-	-	N/A	-	-
	CGH	3,7376E-05	-	-	-	-	-	-	-	-	-	N/A	-	-
	CGH	-	-	-	-	-	5,57E-04	-	-	-	-	N/A	-	-
	CGH	2,53584E-08	1,31138E-23	1,31302E-42	-	-	-	-	-	-	-	N/A	-	-
	CGH	-	-	-	0	0	-	-	-	-	-	N/A	-	-
	CGH	-	-	-	-	-	9,10844E-44	7,00863E-32	-	-	-	N/A	-	-
	CGH	-	-	-	-	-	-	-	3,96747E-14	0	1,28251E-20	N/A	-	-
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4,32E-11	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1,77E-38	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	-	-	-	-	-	-	-	-	-	-	N/A	N/A	N/A
	qPCR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table A.9 – Continuation of table A.7



	Alteration	27	28	29	30	31	32	33	34	35	36	37	38	39
	Type	het del	het del	het del	dup	dup	hemi del	het del	dup	het del	dup	del	del	het del
	Zigoty	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Size (bp)	2247- 5800	8744	274269	1371697	334000	269000	5 Mpb	N/A	59	87	112	N/A	N/A
	Confirmation method	MLPA	MLPA	MLPA	MLPA	MLPA for DMD	CGH	MLPA	MLPA	MLPA	MLPA	MLPA	MLPA	MLPA
	Exon number	+1	+1	+1	+1	+1	+1	+1	+1	1 (2targets)	1 (1 target)	1 (1 target)	1 (1 target)	1
Confirmation method	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	4,64E-17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	6,19E-32	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	4,64E-17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1,36E-05	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.1	N/A	N/A	N/A
	MLPA	N/A	8,43E-34	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8,69E-41	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	-	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-
	MLPA	N/A	N/A	N/A	N/A	3.88E-07	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	qPCR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table A.10 – Continuation of table A.7



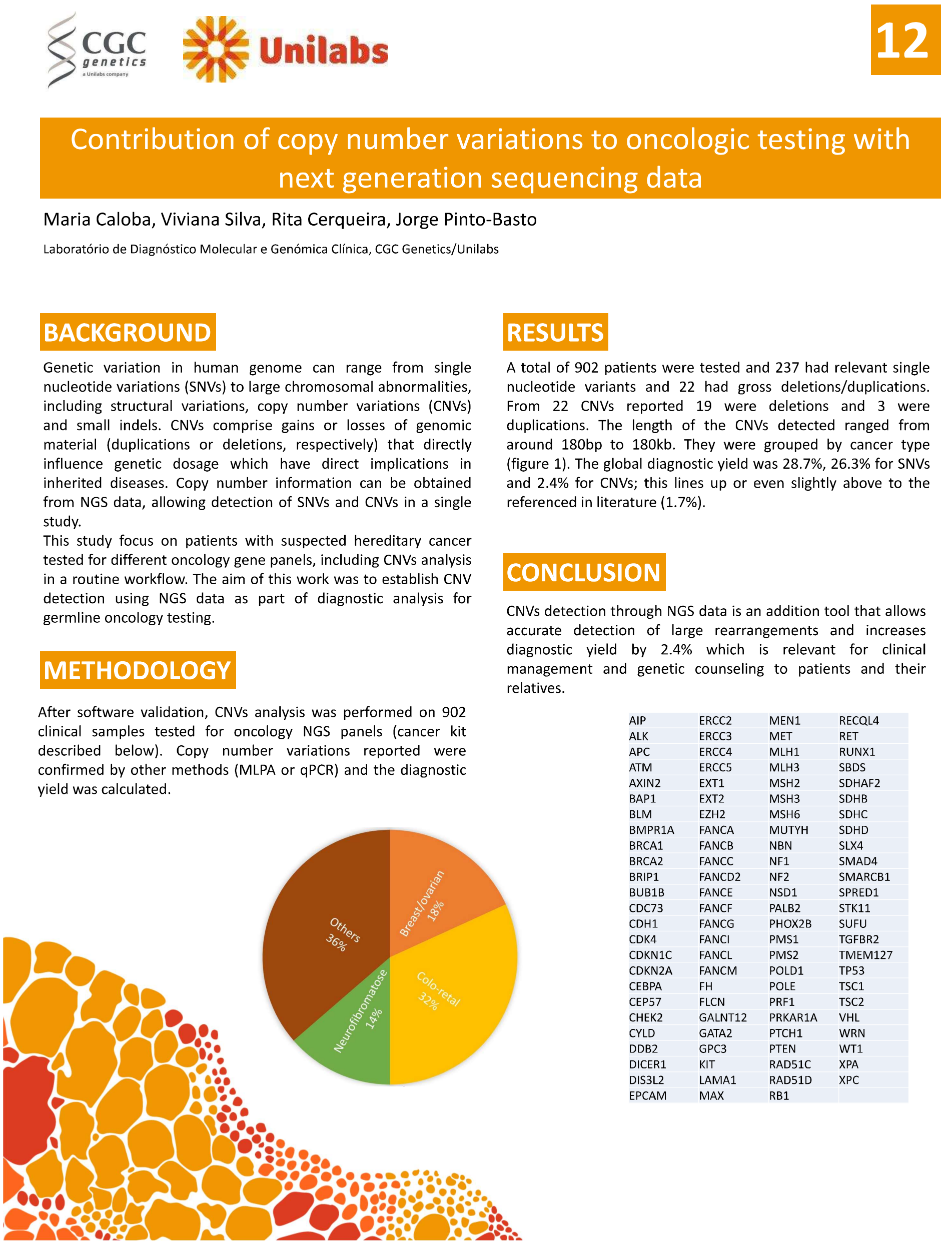
	Alteration	**40**	**41**	**42**	**43**	**44**	**45**	**46**
	Type	het del	Het del	dup	dup	dup	dup	dup
	Zigoty	N/A	N/A	3.0	4.0	N/A	N/A	N/A
	Size (bp)	128	78	135000	136000	N/A	N/A	N/A
	Confirmation method	MLPA	MLPA	CGH	CGH	qPCR	CGH	MLPA
	Exon number	1 (1 target)	1 (1 target)	+1	1 (3 targets)	+1	+1	+1
Confirmation method	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	CGH	N/A	N/A	-	-	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	1,16E-03	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MLPA	N/A	1.13E-06	N/A	N/A	N/A	N/A	N/A
	CGH	N/A	N/A	0	2,74E-04	N/A	N/A	N/A
	qPCR	N/A	N/A	N/A	N/A	1,18965E-19	N/A	N/A
	CGH	N/A	N/A	N/A	N/A	N/A	1,34406E-16	N/A
	MLPA	N/A	N/A	N/A	N/A	N/A	N/A	2,62E-09

Table A.11 – Identification of the genes present in each alteration from table A.7

Alteration	Genes
1	*XG, GYG2*
2	*KCNT2, MIR4735, CFH*
3	*C5orf64, LOC101928651, LOC100506526, KIF2A*
4	*LYRM4-AS1, LYRM4, FARS2, LOC101927972*
5	*SHANK2, FLJ42102, DHCR7, NADSYN1, MIR6754, KRTAP5-7, KRTAP5-8, KRTAP5-9, KRTAP5-10, KRTAP5-11*
6	*EML5, TTC8*
7	*SRY, RPS4Y1, ZFY, LINC00278, TGIF2LY, PCDH11Y, TTTY23, TTTY23B, TSPY2, LINC00280, TTTY1, TTTY1B, TTTY2, TTTY2B, TTTY21B, TTTY21, TTTY7B, TTTY7, TTTY8, TTTY8B, AMELY, TBL1Y, PRKY, TTTY16, TTTY12, TTTY18, TTTY19, TTTY11, RBMY1A3P, TTTY20, TSPY1, TSPY10, FAM197Y5P, FAM197Y2P, TSPY8, TSPY4, TSPY3, RBMY3AP, TTTY22, GYG2P1, TTTY15, USP9Y, DDX3Y, UTY, TMSB4Y, VCY1B, VCY, NLGN4Y, NLGN4Y-AS1, FAM41AY2, FAM41AY1, FAM224B, FAM224A, XKRY2, XKRY, CDY2B, CDY2A, HSFY2, HSFY1, TTTY9A, TTTY9B, TTTY14, CD24, BCORP1, TXLNGY, KDM5D, TTTY10, EIF1AY, RPS4Y2, PRORY, RBMY2EP, RBMY1B, RBMY1A1, RBMY1D, RBMY1E, TTTY13, PRY2, PRY, LOC101929148, TTTY6, TTTY6B, RBMY1F, RBMY1J, TTTY5, RBMY2FP, LOC100652931, TTTY17B, TTTY17C, TTTY17A, TTTY4C, TTTY4, TTTY4B, BPY2C, BPY2B, BPY2, DAZ1, DAZ4, DAZ3, DAZ2, TTTY3B, TTTY3, CDY1B, CDY1, CSPG4P1Y, GOLGA2P2Y, GOLGA2P3Y*
8	*PPP1R3E, BCL2L2, BCL2L2-PABPN1, PABPN1, SLC22A17, EFS, IL25, CMTM5, MYH6, MIR208A, MYH7, MHRT, MIR208B, NGDN, THTPA, ZFHX2, AP1G2, LOC102724814, JPH4, DHRS2, DHRS4-AS1, DHRS4, DHRS4L2, DHRS4L1, CARMIL3, CPNE6, NRL, PCK2, DCAF11, FITM1, PSME1, EMC9, PSME2, MIR7703, RNF31*
9	*BCR, CES5AP1, ZDHHC8P1, LOC101929374, LOC388882, IGLL1, DRICH1, GUSBP11, RGL4, ZNF70, VPREB3, C22orf15, CHCHD10, MMP11, SMARCB1, DERL3, SLC2A11, MIF-AS1, MIF, GSTT2B, GSTT2, DDTL, DDT, GSTTP1, LOC391322, GSTT1-AS1, GSTT1, GSTTP2, CABIN1, SUSD2, GGT5, POM121L9P, SPECC1L, SPECC1L-ADORA2A, ADORA2A, ADORA2A-AS1, UPB1, GUCD1, SNRPD3, GGT1, LRRC75B, BCRP3, POM121L10P*

Table A.12 – Continuation of table A.11

Alteration	Genes
10	*BCL2L14, MIR1244-3, MIR1244-4, MIR1244-2, MIR1244-1, LRP6, MANSC1, LOH12CR2, BORCS5, DUSP16*
11	*LINC00850, MAGEA8-AS1, MAGEA8*
12	*FAM47E, FAM47E-STBD1*
13	*GSTA1*
14	*DMBT1*
15	*FKBP5*
16	*SYCE1, SPRNP1*
17	*TUBGCP5, CYFIP1, NIPA2, NIPA1, LOC283683, WHAMMP3, GOLGA8IP, HERC2P2*
18	*USP50, TRPM7*
19	*CDK11B, SLC35E2B, MMP23A, CDK11A, SLC35E2*
20	*SLC2A14, SLC2A3*
21	*HMGCLL1*
22	*ASCC3*
23	*LOC101927746, IFT22, COL26A1*
24	*CLN3*
25	*MECP2*
26	*PUDP,STS*
27	*CRX*
28	*ABCA3*
29	*GLI3,INHBA*
30	*CDRT4, CDRT15, COX10, HS3ST3B1, PMP22, TEKT3, TVP23C, TVP23C-CDRT4*
31	*DMD,FTHL17*
32	*GRIA3*
33	*Includes UBE3A*
34	*Negative for COL2A1*
35	*NF1*
36	*BRCA1*
37	*LAMA2*
38	*Negative for SPG11*
39	*Negative for USP9X*
40	*SALL4*
41	*CSMD1*
42	*UMOD, POILT, ACSM5, ACSM2A*
43	*LINC00850,MAGEA8-AS1, MAGEA8*
44	*IL1RAPL1*
45	*DHH,LMBR1L,RHEBL1,TUBA1A,TUBA1B*
46	*TSC2*



12

Contribution of copy number variations to oncologic testing with next generation sequencing data

Maria Caloba, Viviana Silva, Rita Cerqueira, Jorge Pinto-Basto

Laboratório de Diagnóstico Molecular e Genómica Clínica, CGC Genetics/Unilabs

BACKGROUND

Genetic variation in human genome can range from single nucleotide variations (SNVs) to large chromosomal abnormalities, including structural variations, copy number variations (CNVs) and small indels. CNVs comprise gains or losses of genomic material (duplications or deletions, respectively) that directly influence genetic dosage which have direct implications in inherited diseases. Copy number information can be obtained from NGS data, allowing detection of SNVs and CNVs in a single study.

This study focus on patients with suspected hereditary cancer tested for different oncology gene panels, including CNVs analysis in a routine workflow. The aim of this work was to establish CNV detection using NGS data as part of diagnostic analysis for germline oncology testing.

METHODOLOGY

After software validation, CNVs analysis was performed on 902 clinical samples tested for oncology NGS panels (cancer kit described below). Copy number variations reported were confirmed by other methods (MLPA or qPCR) and the diagnostic yield was calculated.

RESULTS

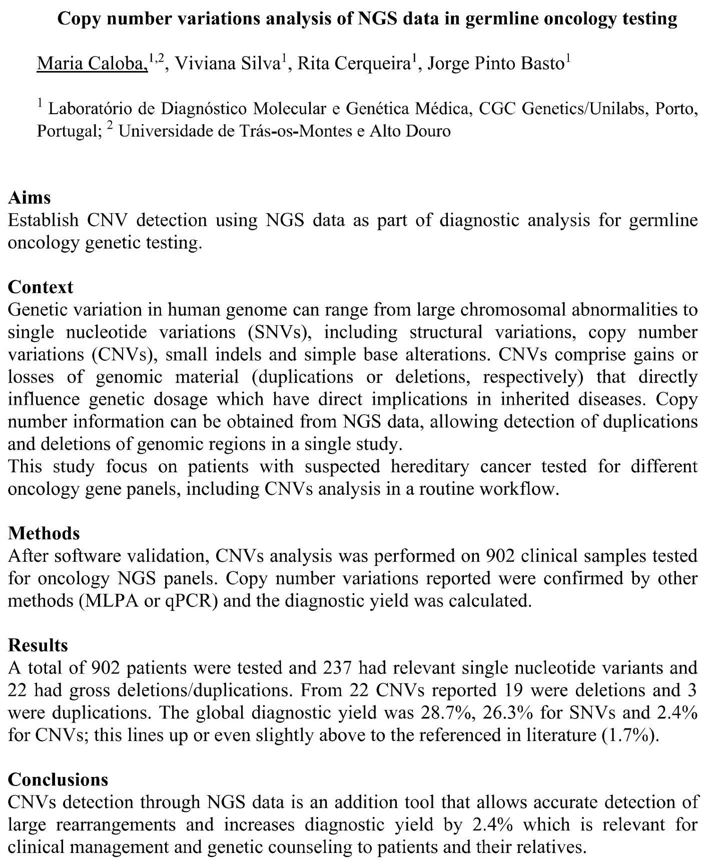
A total of 902 patients were tested and 237 had relevant single nucleotide variants and 22 had gross deletions/duplications. From 22 CNVs reported 19 were deletions and 3 were duplications. The length of the CNVs detected ranged from around 180bp to 180kb. They were grouped by cancer type (figure 1). The global diagnostic yield was 28.7%, 26.3% for SNVs and 2.4% for CNVs; this lines up or even slightly above to the referenced in literature (1.7%).

CONCLUSION

CNVs detection through NGS data is an addition tool that allows accurate detection of large rearrangements and increases diagnostic yield by 2.4% which is relevant for clinical management and genetic counseling to patients and their relatives.

AIP	ERCC2	MEN1	RECQL4
ALK	ERCC3	MET	RET
APC	ERCC4	MLH1	RUNX1
ATM	ERCC5	MLH3	SBDS
AXIN2	EXT1	MSH2	SDHAF2
BAP1	EXT2	MSH3	SDHB
BLM	EZH2	MSH6	SDHC
BMPR1A	FANCA	MUTYH	SDHD
BRCA1	FANCB	NBN	SLX4
BRCA2	FANCC	NF1	SMAD4
BRIP1	FANCD2	NF2	SMARCB1
BUB1B	FANCE	NSD1	SPRED1
CDC73	FANCF	PALB2	STK11
CDH1	FANCG	PHOX2B	SUFU
CDK4	FANCI	PMS1	TGFBR2
CDKN1C	FANCL	PMS2	TMEM127
CDKN2A	FANCM	POLD1	TP53
CEBPA	FH	POLE	TSC1
CEP57	FLCN	PRF1	TSC2
CHEK2	GALNT12	PRKAR1A	VHL
CYLD	GATA2	PTCH1	WRN
DDB2	GPC3	PTEN	WT1
DICER1	KIT	RAD51C	XPA
DIS3L2	LAMA1	RAD51D	XPC
EPCAM	MAX	RB1	

Figure A.7 – Poster presented at 23rd annual meeting of the portuguese society of human genetics



Copy number variations analysis of NGS data in germline oncology testing

Maria Caloba,[1,2], Viviana Silva[1], Rita Cerqueira[1], Jorge Pinto Basto[1]

[1] Laboratório de Diagnóstico Molecular e Genética Médica, CGC Genetics/Unilabs, Porto, Portugal; [2] Universidade de Trás-os-Montes e Alto Douro

Aims

Establish CNV detection using NGS data as part of diagnostic analysis for germline oncology genetic testing.

Context

Genetic variation in human genome can range from large chromosomal abnormalities to single nucleotide variations (SNVs), including structural variations, copy number variations (CNVs), small indels and simple base alterations. CNVs comprise gains or losses of genomic material (duplications or deletions, respectively) that directly influence genetic dosage which have direct implications in inherited diseases. Copy number information can be obtained from NGS data, allowing detection of duplications and deletions of genomic regions in a single study.

This study focus on patients with suspected hereditary cancer tested for different oncology gene panels, including CNVs analysis in a routine workflow.

Methods

After software validation, CNVs analysis was performed on 902 clinical samples tested for oncology NGS panels. Copy number variations reported were confirmed by other methods (MLPA or qPCR) and the diagnostic yield was calculated.

Results

A total of 902 patients were tested and 237 had relevant single nucleotide variants and 22 had gross deletions/duplications. From 22 CNVs reported 19 were deletions and 3 were duplications. The global diagnostic yield was 28.7%, 26.3% for SNVs and 2.4% for CNVs; this lines up or even slightly above to the referenced in literature (1.7%).

Conclusions

CNVs detection through NGS data is an addition tool that allows accurate detection of large rearrangements and increases diagnostic yield by 2.4% which is relevant for clinical management and genetic counseling to patients and their relatives.

Figure A.8 – Abstract (Caloba et al., 2020) published in the journal Medicine (Wolters Kluwer; IF 2017: 2.028)