

# CnvHunter: a new tool for highly accurate detection of single exon copy-number variants in next generation sequencing data, validated in 1554 samples from a targeted hereditary breast cancer panel

B. Auber<sup>1</sup>, G. Schmidt<sup>1</sup>, W. Hofmann<sup>1</sup>, M. Sturm<sup>2</sup>

<sup>1</sup>Department of Human Genetics, Hannover Medical School, Hannover, Germany

<sup>2</sup>Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

## Introduction

Copy number variants (CNVs) represent an important proportion of all detectable pathogenic variants in certain genes. Due to an ever increasing amount of genes in next generation sequencing (NGS) panels, the use of additional methods like multiplex ligation-dependent probe amplification (MLPA) for CNV detection becomes impractical. Using NGS data for the detection of CNVs would greatly reduce costs and yield important data of diagnostic value. However, high resolution single exon CNV detection is challenging, especially in data sets with variable sample sizes or different sequencing platforms used. CnvHunter is a CNV detection algorithm for targeted NGS data, i.e. panel or exome sequencing. CnvHunter was validated with MLPA data from 1554 hereditary breast cancer patients samples.

## Methods

DNA extraction was followed by library preparation (12 or 24 plexes, TruSightCancer®, Illumina, comprising 94 genes associated with hereditary tumor syndromes) and sequencing (MiSeq or NextSeq 500 Sequencer Illumina). Fastq files were generated utilizing bcl2fastq (Illumina). Mapping was performed using our in-house data analysis pipeline megSAP (<https://github.com/imgag/megSAP>). From the bam files, coverage profiles (average sequencing depth for each exon) were created. The input coverage profile of each sample was normalized by the mean sequencing depth. Afterwards, regions with too low average sequencing depth or too high variation in coverage were excluded from the analysis. For each sample, the most similar other 20 samples were detected using the correlation of normalized coverage profiles. From the most similar samples, a synthetic reference sample was constructed. It contained robust estimates of the average coverage and standard deviation for each exon. 55 samples were discarded due to low data quality. To detect CNVs for each sample, regions that are outliers in terms of coverage were detected based on the z-score, calculated using the synthetic reference sample. Around these outliers, adjacent regions were included if they exceed a second (lower) z-score cutoff. Adjacent regions with the same copy-number were merged into one copy-number event. CnvHunter results were compared to MLPA data available for five genes (*BRCA1/2*, *CHEK2*, *RAD51C/D*) from the 1554 samples.

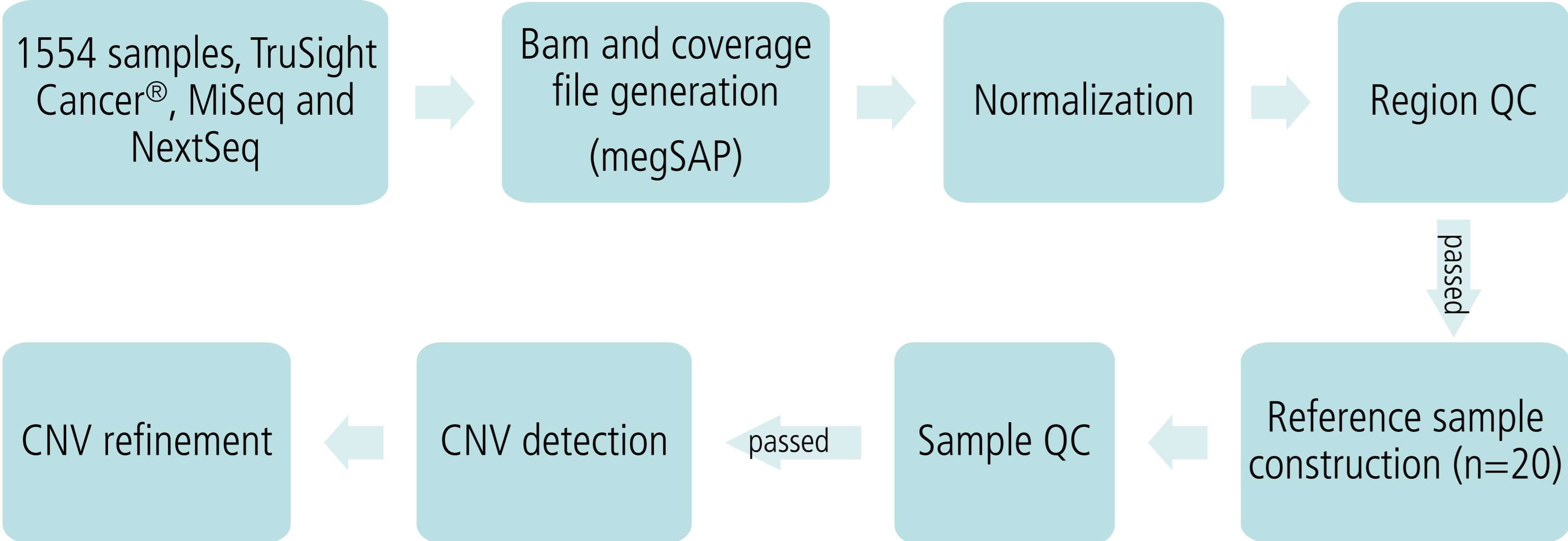


Figure 1. CnvHunter workflow.

Copy-number variants					
CNV calling information					
M43315_01 ref_correl: 0.982 (median: 0.986)					
M43315_01 cnvs: 1 (median: 0)					
M43315_01 qc_errors:					
CNVs					
Min. Size (KB):	0,00	Min regions:	1	Copy-number:	n/a
<input type="checkbox"/> target regions					
position	genes	size (KB)	region count	region_copy_numbers	region_zscores
chr17:56769974-56787382	RAD51C	17.409	5	1,1,1,1,1	-4.42,-4.3,-4.1,-4.42,-4.4

Figure 2. Overview of the copy-number variants in GSvar, the graphical user interface of our sequencing pipeline megSAP.

## Results

Samples	Samples failed QC	Regions True Pos	Regions False Pos	Regions True Neg	Regions False Neg	Sensitivity [%]	Specificity [%]	PPV [%]
1554	55	118	18	84581	1	99.2	99.9	86.9

Table 1. Performance metrics of CnvHunter compared to MLPA results

Gene	True Positives															False negative	
	BRCA1										CHEK2		RAD51C		RAD51D	BRCA1	
Deletion/ <i>duplication</i> exon	8-24	22	4-13	Dup 13	1-19	17	1-2	8	5-14	9-10	6-7	5-9	1-5	5	10	22	
Number of exons	17	1	10	1	20	1	2	1	10	2	2	5	5	1	1	1	
Number of samples	1	2	1	1	1	2	1	1	1	4	1	6	2	1	2	1	

Table 2. Overview of the true positive and true negative CNVs called with CnvHunter.

## Summary

In this highly heterogeneous validation data set, using MLPA, 28 CNVs were detected in 119 exons. CnvHunter detected CNVs in 118 exons (missing *BRCA1* Deletion Ex22) , including 9 single exon CNVs, thus achieving 99% sensitivity and 99% specificity and a positive predictive value of 87%. CnvHunter is a robust high resolution tool for CNV detection in heterogeneous NGS data with a very high sensitivity and specificity. CnvHunter and BedCoverage (auxiliary tool for coverage profile calculation) are available as part of the ngs-bits project (<https://github.com/imgag/ngs-bits>).