HPST | Application Note

# Accurate detection of small copy-number variants (CNVs) using Agilent exon-proximal region designs and VarSome Clinical analysis tool

Tomas Hron, HPST s. r. o., Prague, The Czech Republic, tomas.hron@hpst.cz

## Abstract

Copy Number Variants (CNVs) are deletions or duplications in genomic DNA and represent a major source of variation in the human population. CNVs have been linked to numerous genetic disorders and assessing their importance is a standard part of modern clinical genetics. Although the current methods allow accurate detection of relatively long CNVs, small variants spanning only few exons are still very difficult to identify. Here, we evaluate a solution for the detection of exon-level CNVs by sequencing of Agilent target enrichment libraries followed by analysis in VarSome Clinical.

## Introduction

Next Generation Sequencing (NGS), particularly the targeted sequencing of exons, is becoming the predominant approach for the identification and characterization of genomic variants. In addition to short variants (SNPs and indels), targeted sequencing can also be used to detect CNVs. This is usually performed by read depth analysis assuming that the number of sequenced DNA fragments is proportional to the copy number of a particular genomic locus. This approach is, however, inaccurate in cases of small CNVs spanning only few or even a single exon. Target enrichment library designs developed by Agilent Technologies provide a solution to overcome this limitation by the inclusion of so-called “exon-proximal regions”. These regions provide additional information enabling higher resolution of CNV calling in and around the exons of interest.

Here, we evaluate the performance of this approach in combination with the VarSome Clinical platform, that enables simultaneous detection and classification of both CNVs and short variants.

## Materials and Methods

### Library design

NGS libraries were generated using Agilent SureSelect hybridisation-based target enrichment technology. To evaluate the benefit of designs with exon-proximal regions we compared i) a standard library targeting exon sequences only, and ii) a library targeting both exon and exon-proximal sequences (Figure 1). Both libraries cover 9 clinically relevant genes: DMD, PARK2, ATM, CDKL5, FBN1, HPRT1, PLP1, ERCC6 and MECP2.

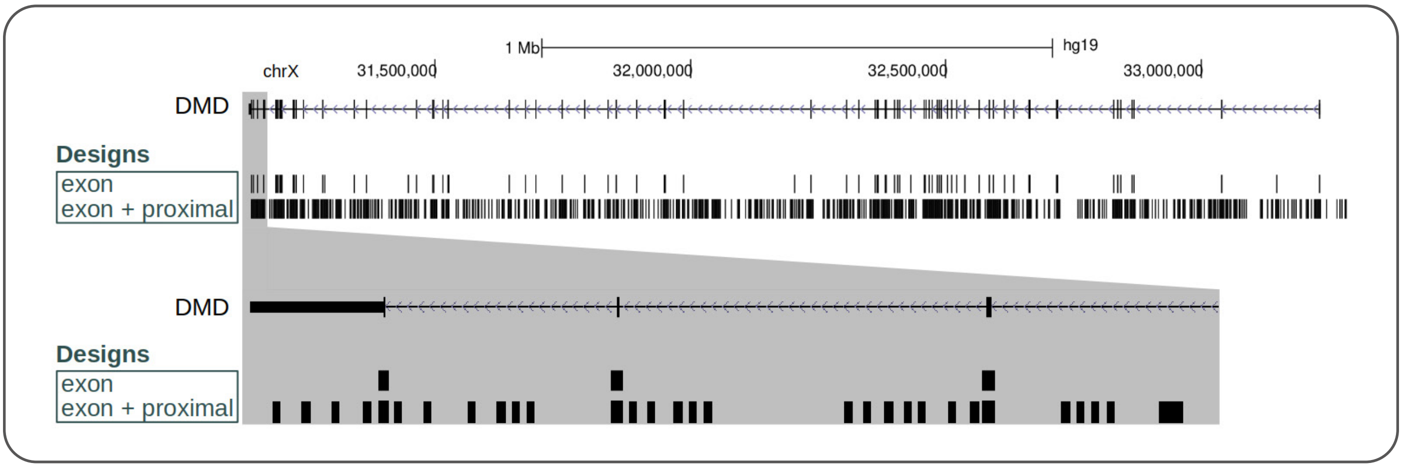


**Figure 1.** Illustration of standard exon library design and exon + exon proximal region design. This figure shows the regions covered by both library designs in the DMD gene.

## Samples

Twenty-four Coriell Institute reference samples were sequenced on Illumina HiSeq2000 (2x100 cycles). In each sample, there was a known CNV mutation spanning 1 to 4 exons (Table 1). All these CNVs are located in the genes covered by the library design used in this study, and thus can be theoretically identified in the data.

# Results

### Analysis of exon-proximal region libraries in VarSome Clinical enables robust exon-level CNV detection.

The CNV detection module of VarSome Clinical uses the ExomeDepth software. This tool enables sensitive CNV calling from targeted sequencing data. Along with each CNV call, it provides a log likelihood score which can be used for efficient filtering of false positive (FP) calls. To determine the appropriate threshold for our data we measured the number of FPs for various log likelihood score values. Any CNV calls that spanned coding exons but were not included in the Coriell Institute's annotations were considered FPs, under the assumption that CNVs in coding regions are unlikely. Based on our results, a score >50 is sufficient to filter almost all false positive calls (Table 1, #FP). Importantly, although the majority of annotated CNVs can be identified by both exon and exon-proximal libraries, only exon-proximal library provides true positive calls with scores > 50 (Table 1, Detected with Score > 50). This enables CNV detection with analytical sensitivity >85% without compromising precision (Figure 2).

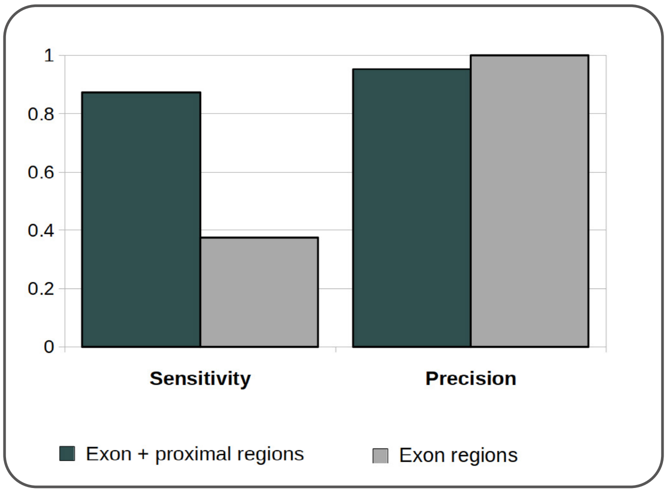


**Figure 2.** Analytical sensitivity and precision estimated for library designs containing either only exon regions or exon regions + proximal regions. Data were analyzed on VarSome Clinical. The threshold score for positive CNV calls is 50.

**Table 1.**

| | | | Exon + proximal regions | | | | Exon regions | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample (Coriell ID) | Affected Gene | Reported CNV | Detected with Score > 50 | #FP * (Score > 50) | Detected with Score > 10 | #FP * (Score > 10) | Detected with Score > 50 | #FP * (Score > 50) | Detected with Score > 10 | #FP * (Score > 10) |
| NA11254 | ATM | deletion (exons 18-20) | ☑ | 0 | ☑ | 1 | ☒ | 0 | ☑ | 1 |
| NA23710 | CDKL5 | deletion (exons 17, 18) | ☑ | 1 | ☑ | 2 | ☑ | 0 | ☑ | 1 |
| NA04315 | DMD | deletion (exon 44) | ☑ | 0 | ☑ | 1 | ☒ | 0 | ☑ | 1 |
| NA05115 | DMD | deletion (exon 45) | ☑ | 0 | ☑ | 0 | ☑ | 0 | ☑ | 0 |
| NA23127 | DMD | duplication (exons 27, 28) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☒ | 0 |
| NA23159 | DMD | duplication (exon 17) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☑ | 1 |
| NA21939 | FBN1 | deletion (exons 42, 43) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☑ | 0 |
| NA06804 | HPRT1 | duplication (exons 2, 3) | ☑ | 0 | ☑ | 1 | ☒ | 0 | ☑ | 1 |
| NA21920 | MECP2 | deletion (exons 1-4) | ☑ | 0 | ☑ | 0 | ☑ | 0 | ☑ | 0 |
| NA23459 | MECP2 | deletion (exon 3, 4) | ☒ | 0 | ☒ | 1 | ☒ | 0 | ☒ | 1 |
| NA23599 | MECP2 | deletion (exons 3, 4) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☑ | 0 |
| NA23635 | MECP2 | deletion (exons 3, 4) | ☑ | 0 | ☑ | 1 | ☒ | 0 | ☑ | 1 |
| NA23648 | MECP2 | deletion (exons 1-4) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☒ | 0 |
| NA23654 | MECP2 | deletion (exons 3, 4) | ☑ | 0 | ☑ | 0 | ☑ | 0 | ☑ | 0 |
| NA23675 | MECP2 | duplication (exons 1-4) | ☑ | 0 | ☑ | 1 | ☑ | 0 | ☑ | 0 |
| NA23676 | MECP2 | duplication (exons 1-4) | ☑ | 0 | ☑ | 1 | ☑ | 0 | ☑ | 1 |
| ND01037 | PARK2 | deletion (exon 4) | ☑ | 0 | ☑ | 1 | ☑ | 0 | ☑ | 0 |
| ND01038 | PARK2 | deletion (exon 4) | ☒ | 0 | ☒ | 2 | ☒ | 0 | ☒ | 0 |
| ND01040 | PARK2 | deletion (exon 4) | ☑ | 0 | ☑ | 1 | ☒ | 0 | ☑ | 0 |
| ND06284 | PARK2 | deletion (exon 3) | ☑ | 0 | ☑ | 0 | ☒ | 0 | ☑ | 0 |
| ND07278 | PARK2 | deletion (exons 3, 4) | ☑ | 0 | ☑ | 2 | ☒ | 0 | ☑ | 1 |
| ND08917 | PARK2 | deletion (exons 5, 6) | ☒ | 0 | ☒ | 0 | ☒ | 0 | ☒ | 2 |
| ND35201 | PARK2 | deletion (exon 3) | ☑ | 0 | ☑ | 0 | ☑ | 0 | ☑ | 0 |
| NA13434 | PLP1 | deletion (exons 3, 4) | ☑ | 0 | ☑ | 0 | ☑ | 0 | ☑ | 0 |

* FP – False positives. CNV calls outside coding regions were not considered

## Exon-proximal regions increase the accuracy of CNV breakpoint assignment.

Read depth analysis of targeted sequencing data does not provide the exact position of CNV breakpoints. Instead, it predicts the minimal CNV length where the CNV borders are the first and the last bin in a genomic segment with aberrant read depth. The additional information provided by the exon-proximal design can theoretically give a more accurate picture of the real CNV, which is usually much longer than the estimate. This is supported by the fact that the length of all true positive CNV calls is higher in the exon-proximal design compared to the exon-only design (Figure 3). Besides increasing sensitivity, exon-proximal regions can therefore also improve the accuracy of of CNV breakpoint detection.

## The intuitive interface of VarSome Clinical enables easy inspection of individual CNV calls.

Each CNV call can be visualized in the context of genes and other variants present in public databases. In addition to this, the user can also inspect a plot of observed/expected sequencing coverage to evaluate the reliability of each call. The example in Figure 4 shows a duplication of a single exon in the DMD gene. It nicely illustrates that inclusion of exon-proximal regions provides much more information for the calling algorithm and thus enhances CNV detection.

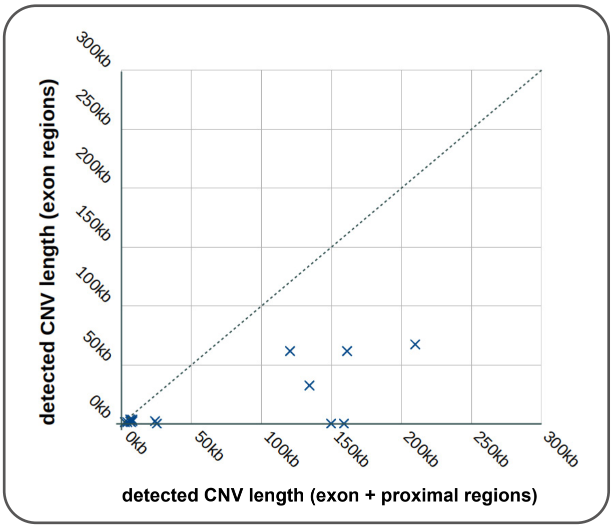
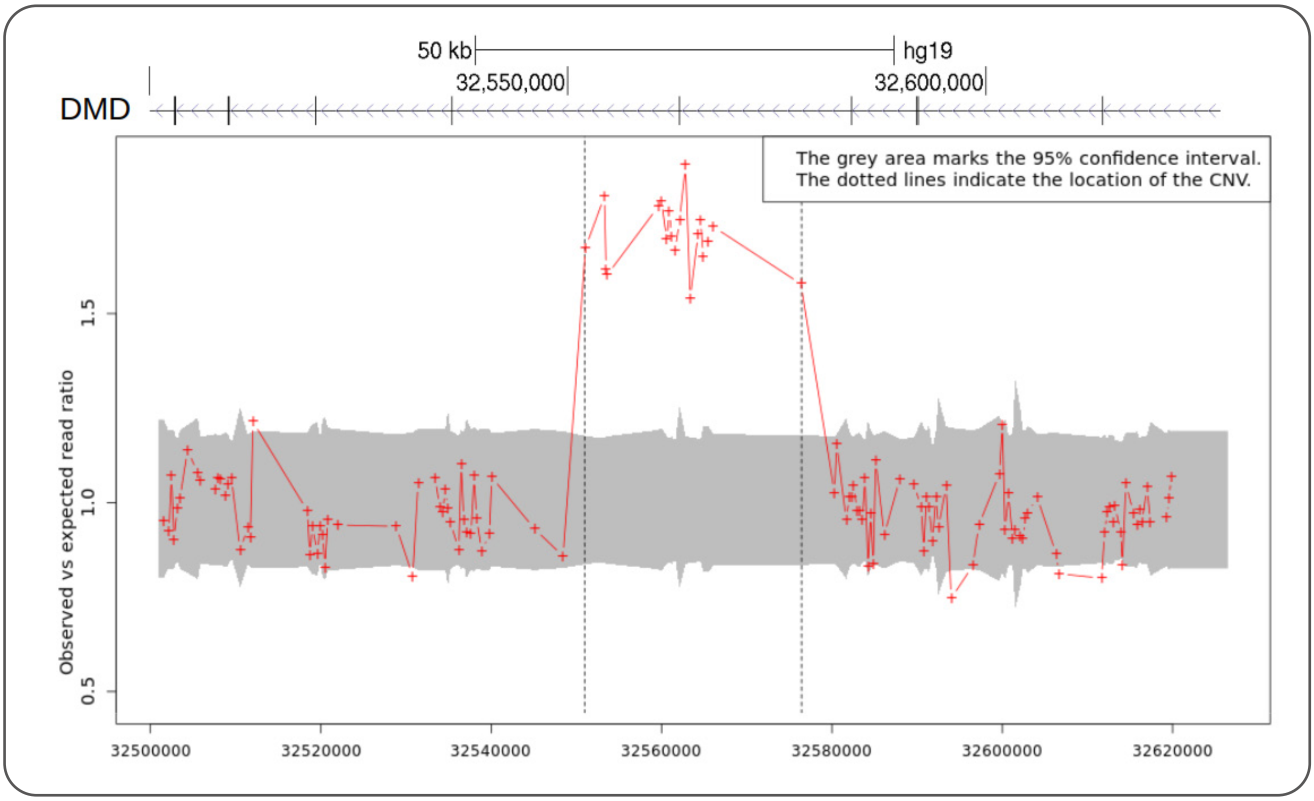


**Figure 3.** Estimates of minimal CNV length called in exon and exon + proximal region libraries, respectively.

(A)



(B)

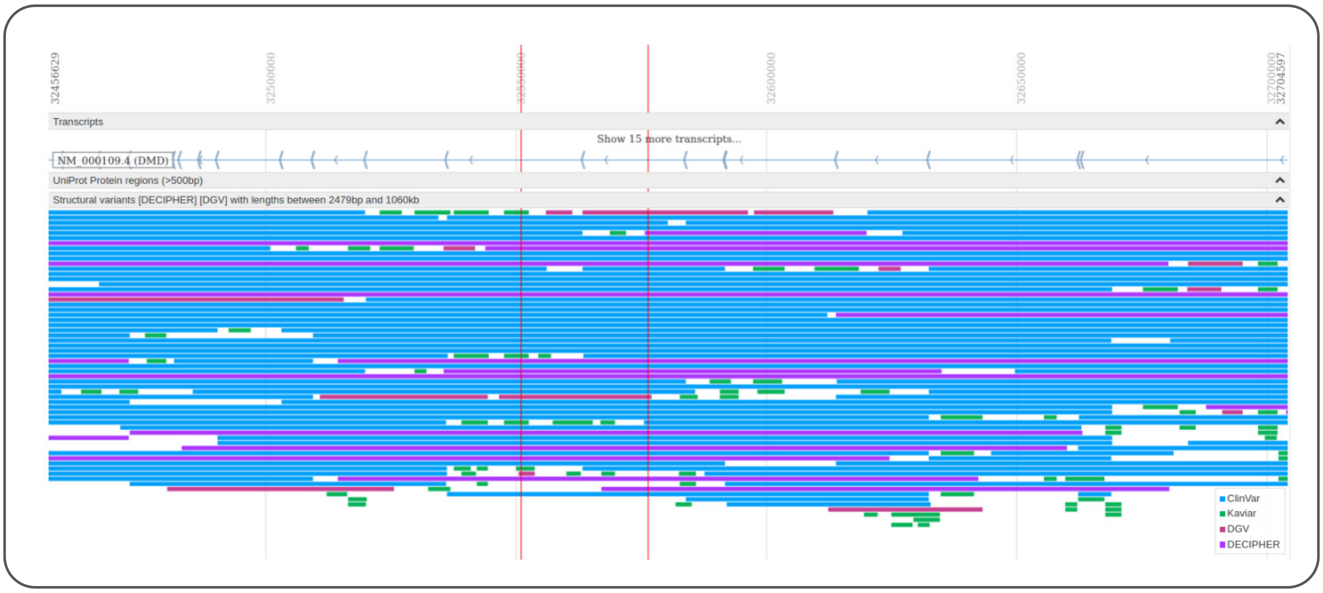

32456629
32500000
32550000
32600000
32650000
32700000
32704597
Transcripts
Show 15 more transcripts...
NM_000109.4 (DMD)
UniProt Protein regions (>500bp)
Structural variants [DECIPHER] [DGV] with lengths between 2479bp and 1060kb
ClinVar
Kaviar
DGV
DECIPHER

**Figure 4.** Visualization of CNV calling results in VarSome Clinical. (A) Observed/expected sequencing coverage ratio plot. The plot shows a single-exon duplication in the DMD gene. Each dot in the chart represents one genomic region covered in a design. Dashed lines indicate the duplicated region and the diagram above the chart shows the DMD gene structure.

**Acknowledgments**

Here we would like to acknowledge Sandra Melo and whole Diagnostics and Genomics Group of Agilent Technologies (Santa Clara, CA 95051) who developed the concept of exon-proximal region designs and provide us the data.

**References**

**HPST, s. r. o.,** is a distributor of VarSome Clinical platform and an authorized distributor of Agilent Technologies in the Czech Republic in the fields of chromatography and mass spectrometry, dissolution, molecular and atomic spectroscopy and not least in the field of molecular biology and genomics. The company also provides authorized service for Agilent Technologies instruments.

www.hpst.cz

**Agilent Technologies, Inc.,** is a leader in life sciences, diagnostics and applied chemical markets. The company provides laboratories worldwide with instruments, services, consumables, applications and expertise, enabling customers to gain the insights they seek. Agilent's expertise and trusted collaboration give them the highest confidence in our solutions.

www.agilent.cz

**VarSome.com** is a community-driven project featuring an aggregated knowledge base consisting of 50+ cross-referenced public data resources and contributions from its community of more than 200‘000 users worldwide.

Christos Kopanos, Vasilis Tsiolkas, Alexandros Kouris, Charles E Chapple, Monica Albarca Aguilera, Richard Meyer, Andreas Massouras, VarSome: the human genomic variant search engine, *Bioinformatics*, Volume 35, Issue 11, 1 June 2019, Pages 1978–1980

https://doi.org/10.1093/bioinformatics/bty897

**VarSome Clinical** is a CE-IVD certified and HIPAA-compliant platform allowing fast and accurate variant discovery, annotation, interpretation, and reporting for NGS data for whole genomes, exomes, and gene panels.

https://saphetor.com/varsome-editions/varsome-clinical/