CytoSure

An approach for determination of copy number variation using short-read next-generation sequencing

Kaajal Reeves[a], Mafalda Bourbon[b,c], Sandra Kachhia[a], Douglas Hurd[a], James Reid[a], Duarte Molha[a], Darren Houniet[a] and John Cousin[a]

[a]Oxford Gene Technology (OGT), [b]Unidade de I&D, [c]BioISI-Biosystems & Integrative Sciences Institute.

A Sysmex Group Company

What binds us, makes us.

Introduction

The ability to determine Copy Number Variation (CNV) from short-read Next Generation Sequencing (NGS) data would enable laboratories to determine both CNV and Single Nucleotide Variation (SNV) simultaneously in one assay. To date, most NGS CNV analysis approaches have been designed for whole genome/exome sequencing and besides being less robust than standard array comparative genomic hybridisation (aCGH), they are not suitable for small targeted NGS panels. In this study, we describe a method for detecting exonic CNVs using targeted NGS panels and a bioinformatics approach. This is demonstrated with the *LDLR* gene, involved in Familial Hypercholesterolaemia (FH) which is thought to have a prevalence between 1/500 and 1/200[1]; and the *DMD* gene on the X chromosome, involved in Duchenne muscular dystrophy (DMD) which has an estimated prevalence of 1/3500 of male births[2].

Methods

In this study, samples underwent library preparation, short-read sequencing and bioinformatics analysis to determine intragenic CNV status. CNVs in samples were confirmed by aCGH.

Samples

Two sample cohorts were analysed in this project:

1. For *LDLR*, 48 samples (5 with confirmed CNVs) were supplied by the Instituto Nacional de Saúde Doutor Ricardo Jorge. CNV status was previously determined by the institute using Multiplex Ligation-dependent Probe Amplification (MLPA) analysis for the *LDLR* gene. All samples were processed blind at Oxford Gene Technology (OGT).

2. For *DMD*, 50 samples were purchased from the Coriell Institute for Medical Research with confirmed copy number status, 29 of which had known CNVs on the *DMD* gene, and 21 which were copy neutral.

Library preparation and sequencing

An overview of the library preparation and sequencing workflow is shown in Figure. 1.

In brief, sheared DNA was amplified using the SureSeq™ NGS Library Preparation kit (OGT) and 500ng of each library was hybridised overnight to an OGT-developed panel of biotinylated oligonucleotides using the SeqCap EZ Hybridization and wash kit (Roche Nimblegen). The hybridised DNA fragments were captured to streptavidin beads, washed and subsequently amplified with the SureSeq NGS Library Preparation kit using indexing primers. A 4nM pool consisting of 24 individual DNA libraries was loaded into a v2 300 cycle Miseq cartridge (Illumina) and run on an Illumina Miseq.

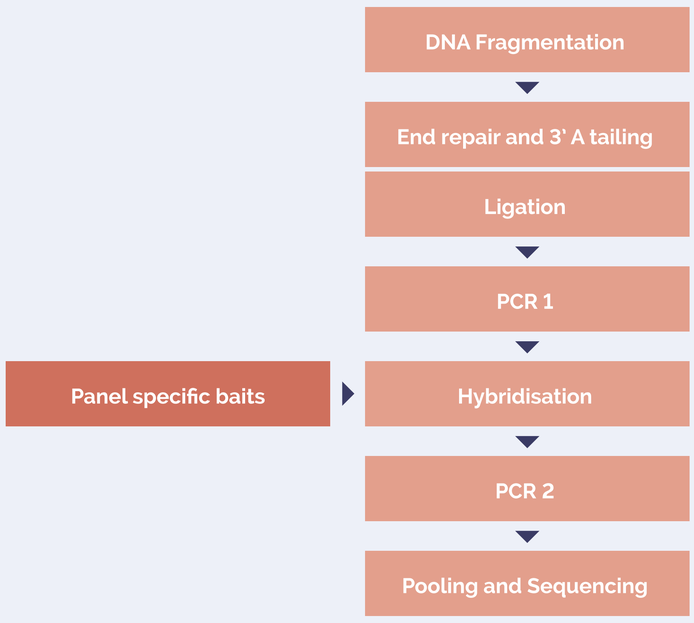

Figure 1: Workflow of SureSeq NGS library preparation, from DNA to sequencer.

A proprietary bait design approach is used in all OGT-developed panels to improve uniformity and depth of coverage across all target regions, important in accurately calling regions with CNVs. The two panel designs used in this experiment as are follows:

A SureSeq FH panel targeting 4 genes and 2 SNPs implicated in the condition (*LDLR*, *PCSK9*, *LDLRAP1*, *APOB*, rs2306283 and rs4149056)

A SureSeq DMD panel targeting all coding regions of the *DMD* gene

Bioinformatics analysis

Sequence analysis including germline CNV detection was performed using OGT's NGS analysis software, Interpret, (Hg 19 was used as the reference genome on IGV[3]) from FASTQ files obtained from the sequencer.

CNV detection algorithms used by Interpret were developed in-house by OGT, and use read-depth analysis with pre-determined parameters to determine copy-number status.

aCGH

All samples determined by NGS analysis as containing CNVs were confirmed using aCGH.

High-resolution microarrays (OGT) were designed with high density of probes within the genes of interest. Samples were processed and labelled using the CytoSure® labelling kit (OGT), and analysis was performed using CytoSure Interpret software (OGT).

Results

Uniformity of coverage is important in CNV calling from NGS and was high across all targets – examples shown in Figure 2.

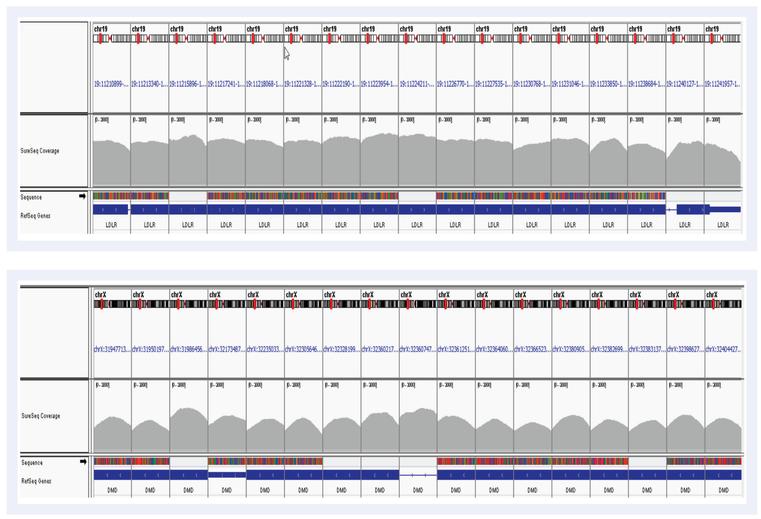

Figure 2: Uniformity of coverage in the *LDLR* gene (top) and part of the *DMD* gene (bottom) shown on IGV. Read depth across targets is indicated in grey.

LDLR

Using the FH panel, 48 samples were processed in total, and the OGT algorithm correctly called CNVs in 5 samples. All of these CNVs were confirmed by the aCGH method. The remaining 43 samples were correctly called negative, concordant with previous MLPA assays.

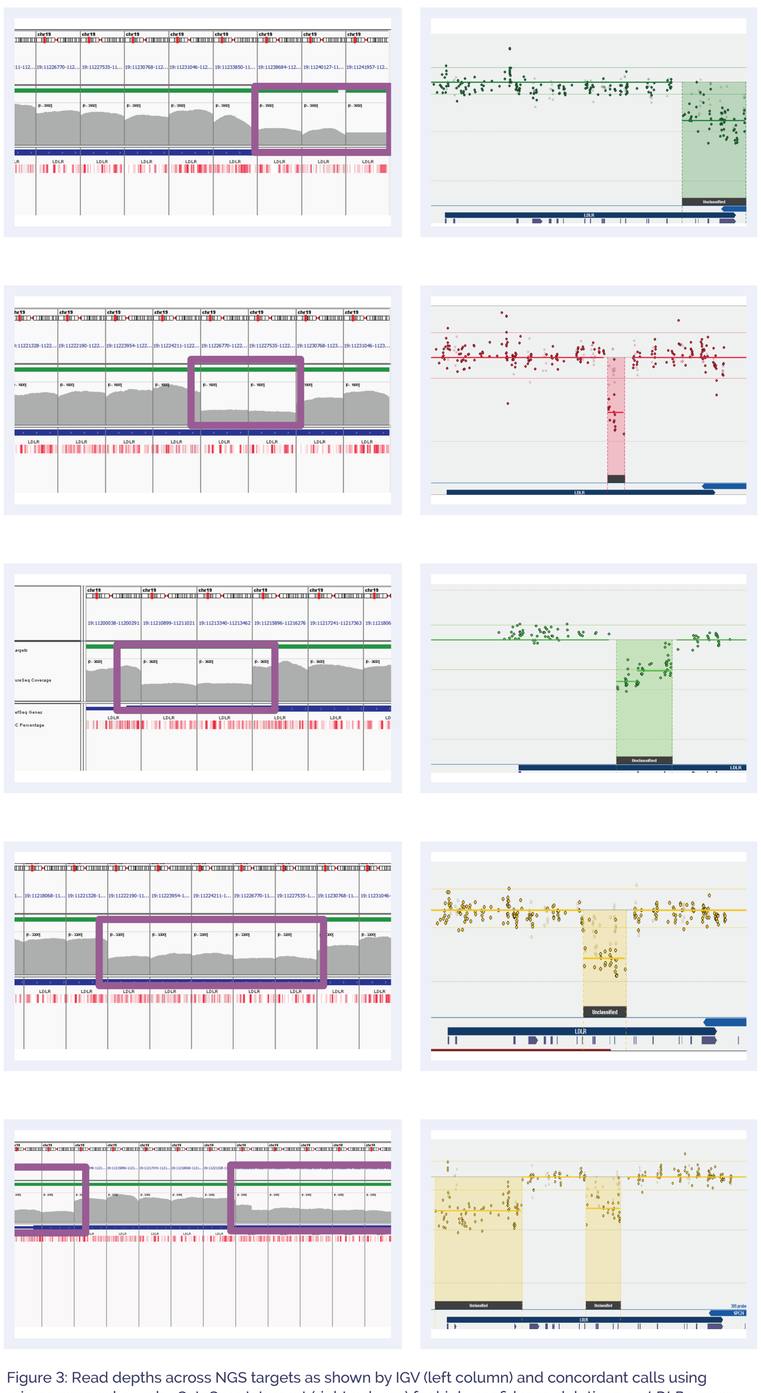

Figure 3: Read depths across NGS targets as shown by IGV (left column) and concordant calls using microarray as shown by CytoSure Interpret (right column) for high-confidence deletions on *LDLR* across 5 separate samples (all heterozygous deletions). Purple boxes highlight regions where read depth is indicative of CNV.

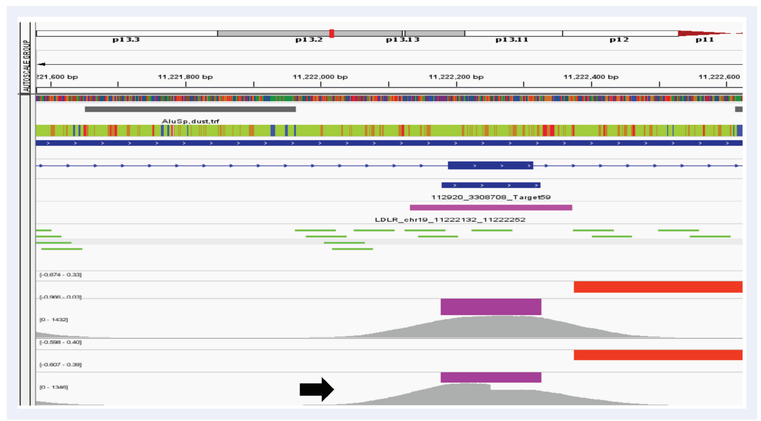

Figure 4: Example of a lower confidence call on *LDLR* due to a mid-exon breakpoint. Tracks as follows: aCGH probes used for confirmation (bright green lines), aCGH CNV calls (red), NGS CNV calls (purple). In one sample (black arrow), the aberration occurs mid-exon but is clearly visible and called by our algorithm. aCGH breakpoint detection occurs only over the next probe already within the CNV.

DMD

For processing samples with the DMD panel, analysis requires additional care to account for gender and X chromosome number. Breakpoints of CNVs are normally better resolved on aCGH due to coverage across intronic regions. However, within the 50 samples processed for this study, the OGT DMD panel successfully detected all affected exons overlapping the aberrations within the *DMD* gene, ranging from single exons to large genomic regions covering multiple exons. This was concordant with all CNVs described by the Coriell Institute for Medical Research.

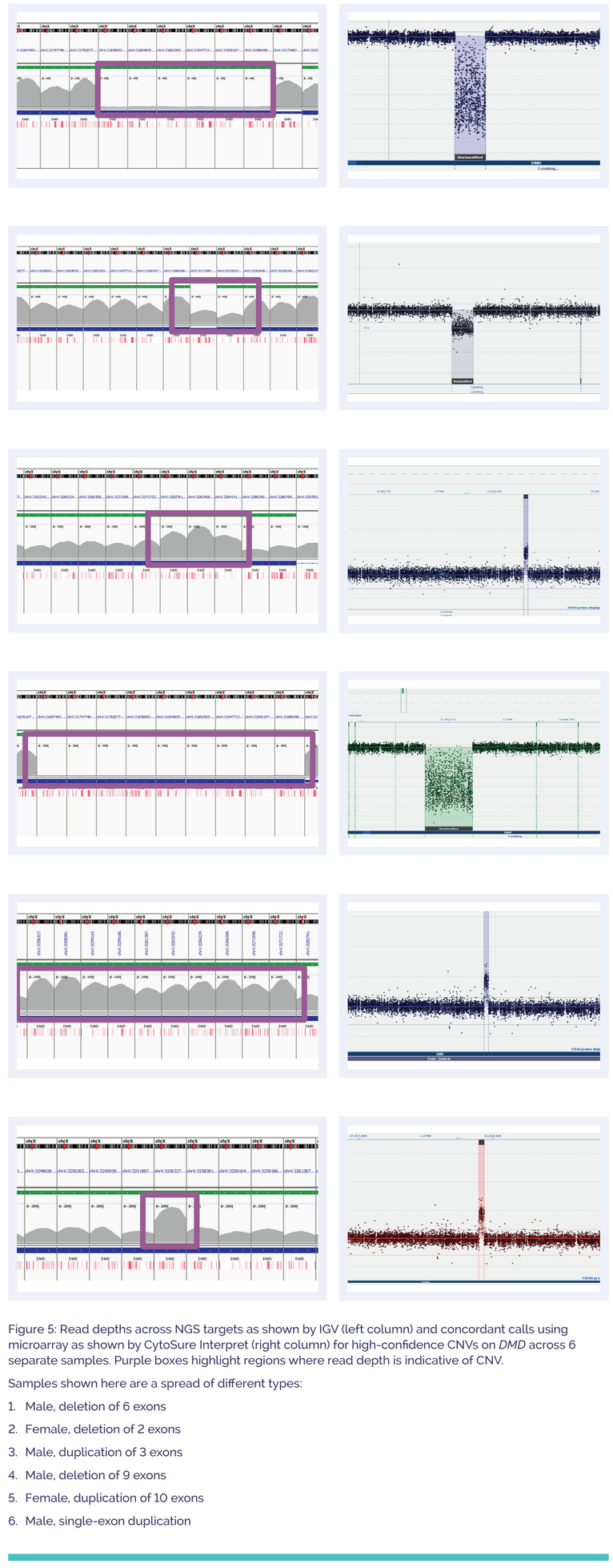

Figure 5: Read depths across NGS targets as shown by IGV (left column) and concordant calls using microarray as shown by CytoSure Interpret (right column) for high-confidence CNVs on *DMD* across 6 separate samples. Purple boxes highlight regions where read depth is indicative of CNV.

Samples shown here are a spread of different types:

1. Male, deletion of 6 exons
2. Female, deletion of 2 exons
3. Male, duplication of 3 exons
4. Male, deletion of 9 exons
5. Female, duplication of 10 exons
6. Male, single-exon duplication

Conclusions

We have shown that intragenic CNVs can be detected using the OGT SureSeq NGS assay and confirmed with aCGH. The concordance was 100% over the targeted exons on the NGS panel against other techniques. These results indicate that a combined NGS and bioinformatics approach can be reliably used to determine CNVs in *LDLR* and *DMD* with potential for use in other applications.

References

1. Brice, P; Burton, H; Edwards, CW; Humphries, SE; Aitman, TJ; (2013) Familial hypercholesterolaemia: A pressing issue for European health care. Atherosclerosis, 231(2), pp. 223-226.
2. Emery, AE; (1991) Population frequencies of inherited neuromuscular diseases - a world survey. Neuromuscular Disorders 1(1), pp. 19-29.
3. Thorvaldsdóttir, H; Robinson, JT; and Mesirov, JP; (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14(2), pp.178-192.

Acknowledgements

a. Oxford Gene Technology (OGT), Oxford, UK.
b. Unidade de I&D, Grupo de Investigação Cardiovascular, Departamento de Promoção da Saúde e Prevenção de Doenças Não Transmissíveis, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisbon, Portugal.
c. BioISI-Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisbon, Lisbon, Portugal.

Further information **contact@ogt.com | ogt.com | Oxford Gene Technology Ltd., Begbroke Science Park, Woodstock Road, Begbroke, Oxfordshire, OX5 1PF, UK.**

990314 01/21