Optimizations in target enrichment and bioinformatics enable sensitive detection of copy number variations in targeted NGS

Jeroen Van Den Akker, Gilad Mishne, Robert O'Connor, Ziga Mahkovec, Lawrence Hon, Taylor Sittler
Color Genomics, Burlingame, California, USA

Introduction

Copy number variations (CNVs) account for a significant proportion of variation in the human genome[1]. CNVs are also known to play a significant role in a variety of diseases, contributing up to 20% of mutations associated with hereditary cancer in certain populations[2,3]. However, their detection in targeted next generation sequencing (NGS) assays has been historically challenging[4,5]. Here, we report on our multifaceted approach for the detection of CNVs (deletions and duplications) in our 30-gene NGS-based test for hereditary cancer risk (the Color Hereditary Cancer Test). We were able to boost detection sensitivity by performing optimizations at multiple levels, which resulted in identification of novel CNVs in virtually all of the genes in our breast/ovarian and hereditary cancer panel.

Conclusions

- NGS has the potential of highly sensitive CNV detection, and allows the structural variant to be accurately characterized.
- CNV detection based on read depth requires coverage normalization correction for systematic biases, as well as multiple segmentation methods targeting different CNV sizes.
- The highest sensitivity is achieved by targeting CNV breakpoints during sequencing, combined with dedicated algorithms relying on split read and paired read signals. This approach allows detection of deletions and duplications based on a minimum of 3 supporting reads.
- While the clinical relevance of CNVs in *BRCA1*, *BRCA2* and the Lynch syndrome genes[3] has been well established, data on other genes associated with hereditary cancer is scarce. Here, we identified CNVs in *CHEK2* (n=42, 13 distinct CNVs), *ATM* (n=31, 12 distinct), *RAD51C* (n=22, 10 distinct) and *PALB2* (n=16, 10 distinct).

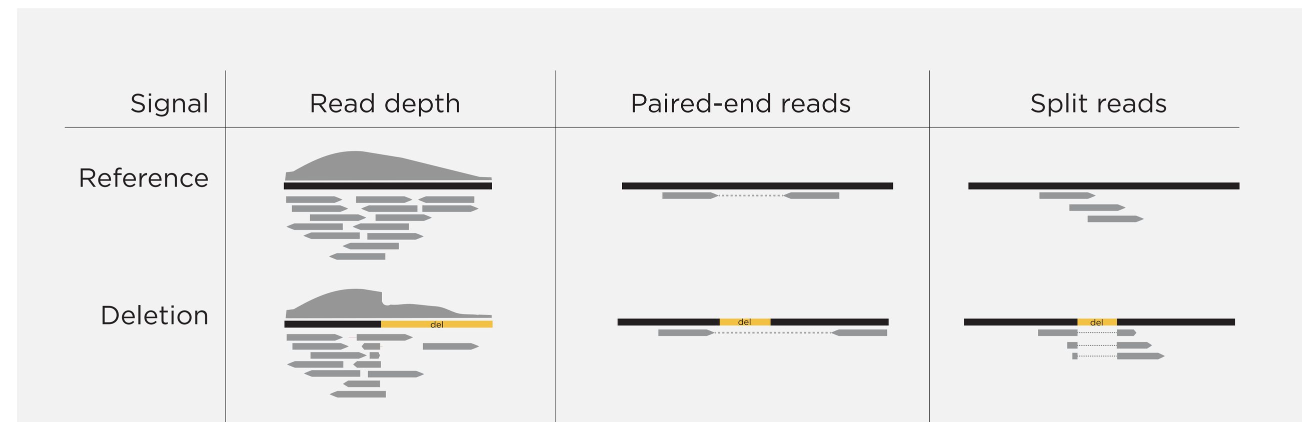

Figure 1A. NGS features used for CNV calling

In this example, a deletion (yellow) reduces the read depth, increases the distance between paired-end reads (when mapped to the reference), and causes portions of reads to map to different locations of the reference, a phenomenon referred to as split reads.

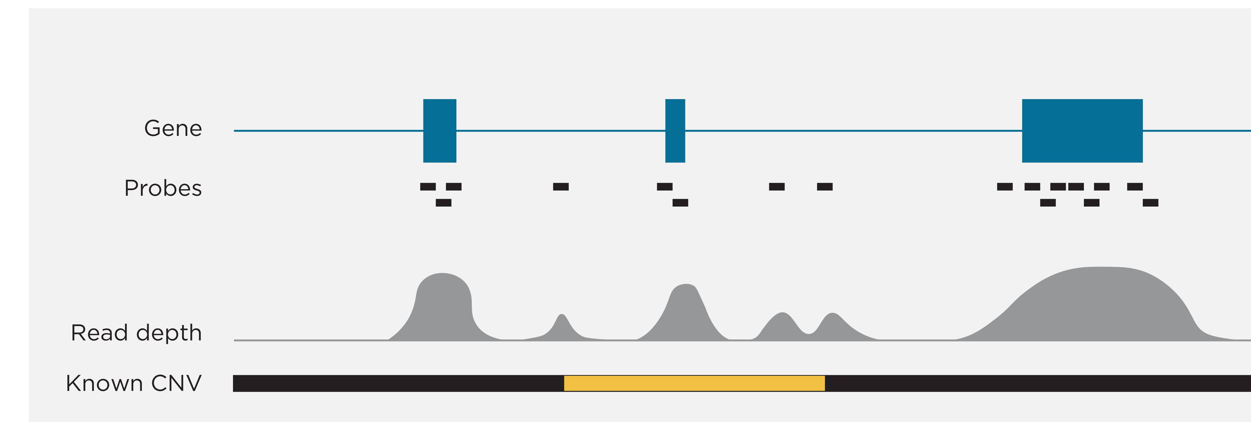

Figure 1B. Optimization of Probe Design

In order to maximize detection sensitivity, the assay is designed to use capture probes that target both coding exons as well as intronic regions known to harbor structural rearrangements. In this way, we can capture the signals that can be derived from CNV breakpoints (Figure 1A).

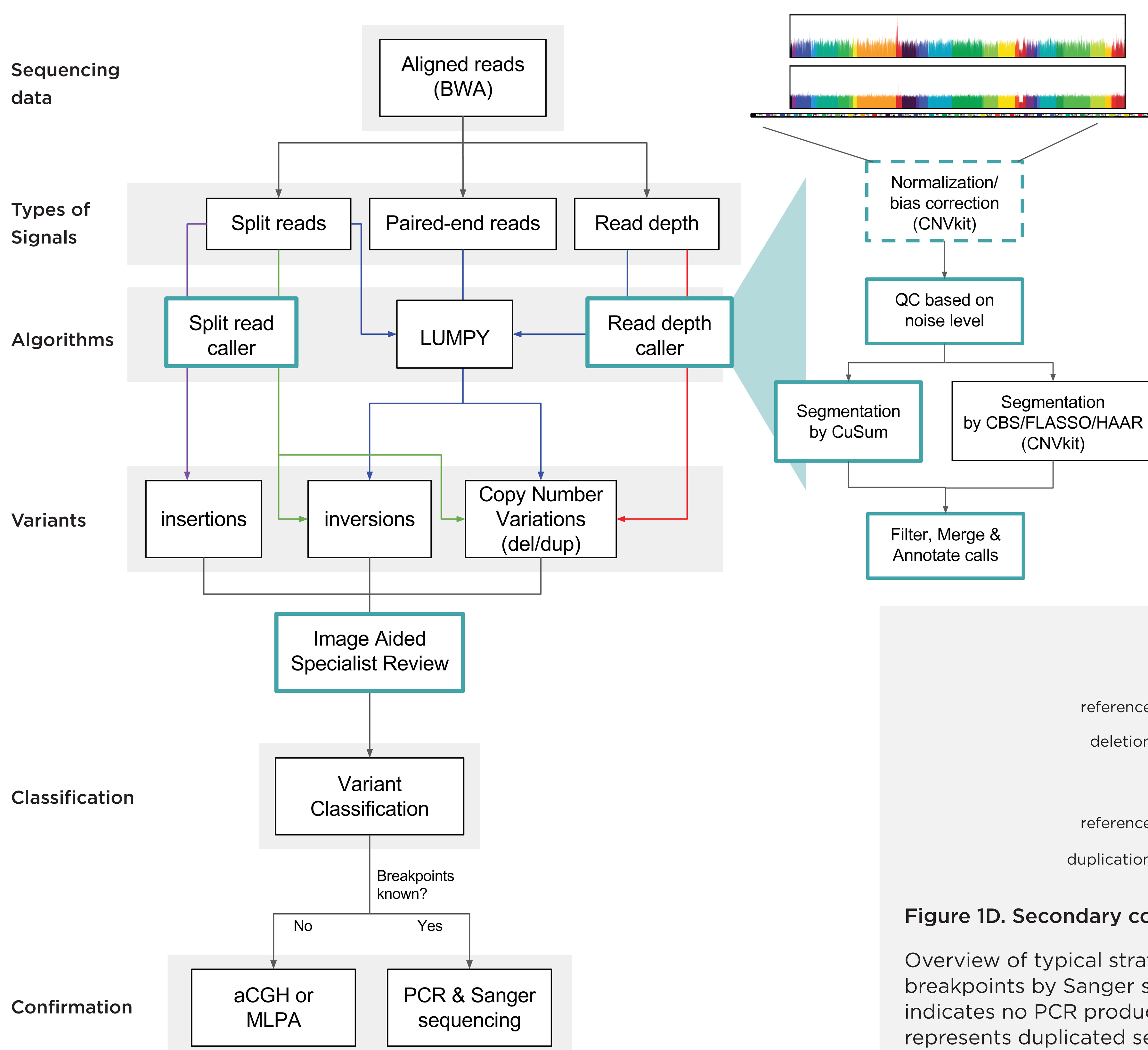

Figure 1C. Bioinformatics Pipeline

Left: Sequential steps from alignment to confirmation of structural variants. Right: Optimizations for CNV detection based on read depth, including a normalization step comprised of in-house and CNVkit derived elements. Top right: Example of noise reduction; top track shows raw data and bottom track shows normalized signal, emphasizing a 5-kb deletion of *BRCA1* exon 13. Both publicly available algorithms (BWA[6], LUMPY[7] and CNVkit[8]), and in-house developed algorithms (outlined in teal) are used in our pipeline.

PCR Strategy | Sanger Strategy
reference
deletion
reference
duplication

Figure 1D. Secondary confirmation strategies

Overview of typical strategies for confirmation of CNVs with known breakpoints by Sanger sequencing. Broken arrows represent primers, 'X' indicates no PCR product, yellow indicates deleted sequence, and blue represents duplicated sequence.

Results

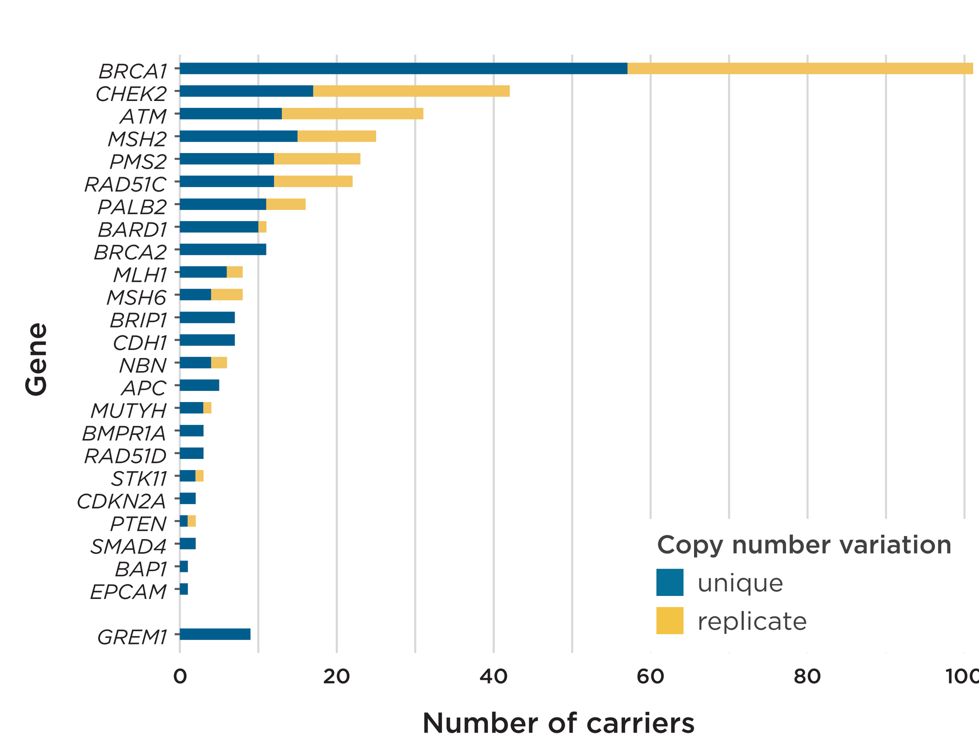

Figure 2. CNVs by gene

Color has identified CNVs in >350 clinical samples to date that were classified as P, LP or VUS, representing approximately 220 distinct variations. In accordance with previous findings[2,9], the frequency of CNVs is much lower in *BRCA2* compared to *BRCA1*, which accounts for almost 30% of all CNVs. *GREM1* was only analyzed for duplications overlapping its enhancer region, revealing 9 distinct events.

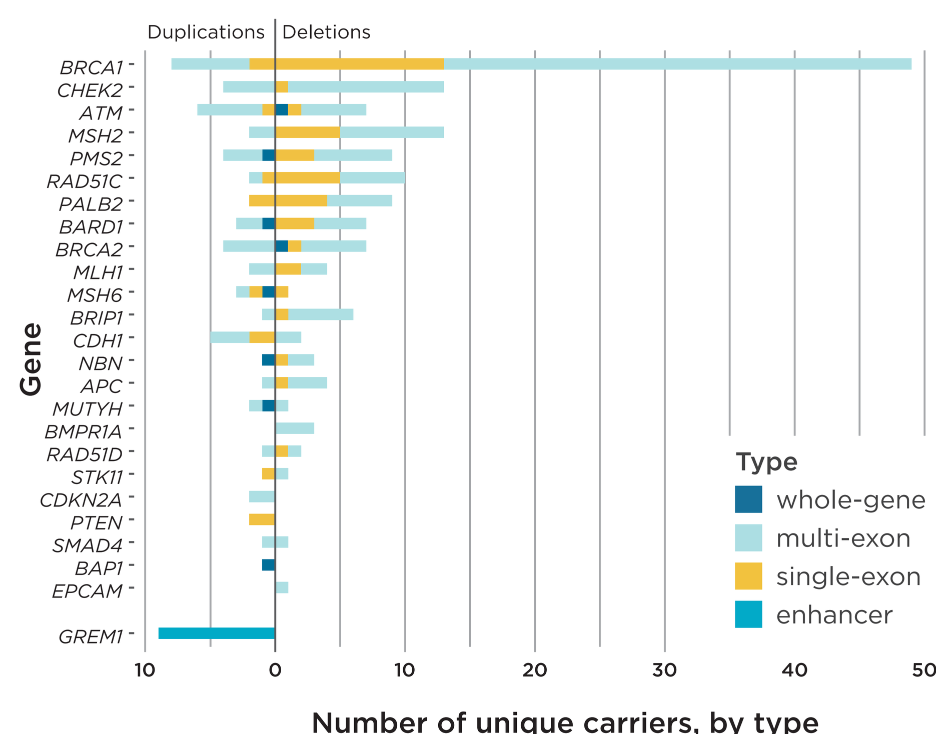

Figure 3. CNVs by type

Due to the optimizations outlined in figure 1, our assay can detect CNVs of all sizes. Almost 30% of our reported CNVs impacted only a single exon, including 13 in *BRCA1*. Our split-read detection algorithm has detected 17 variants between 50-250 base pairs, a size range known to be difficult in most NGS assays.

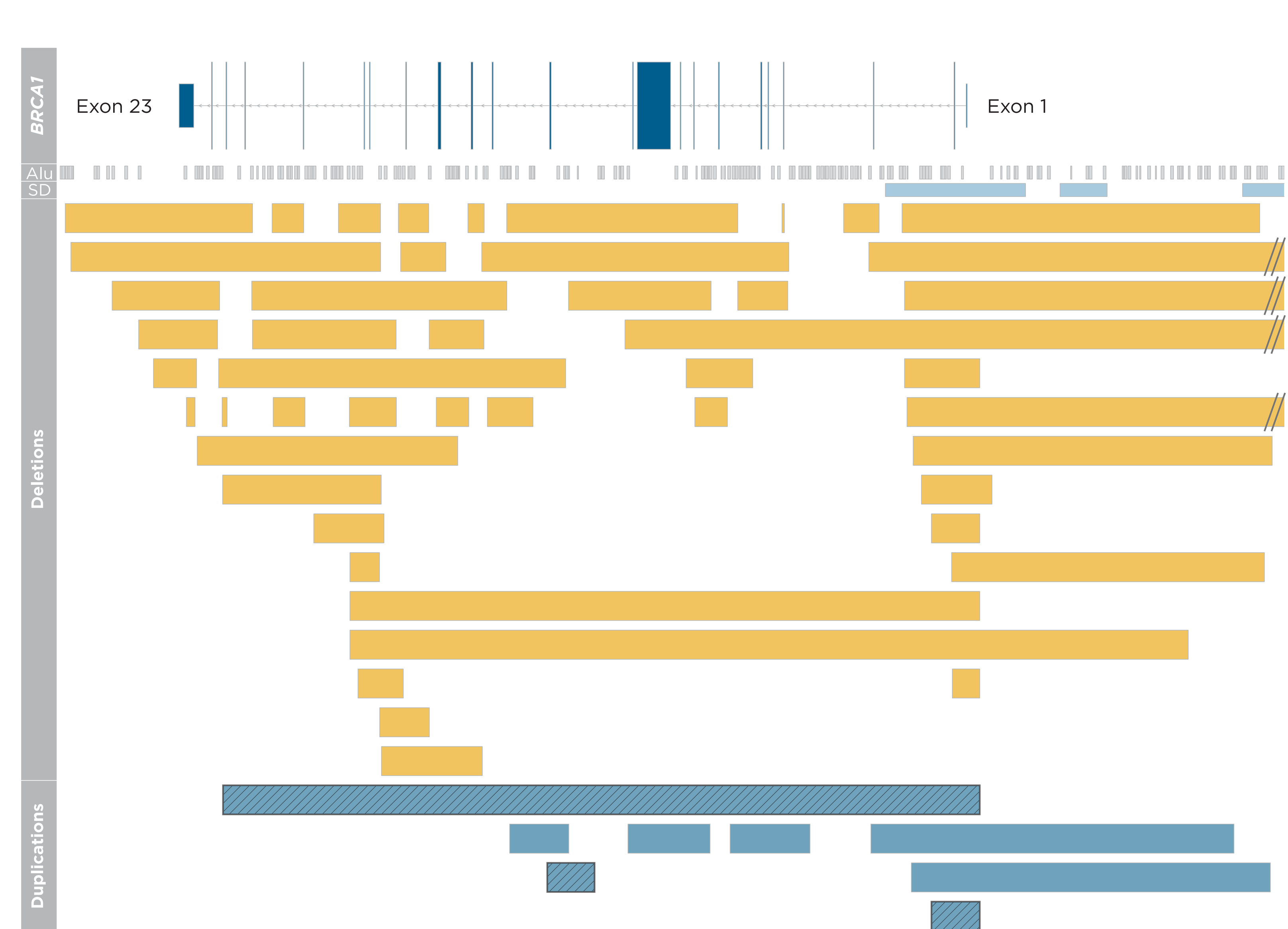

Figure 4. CNVs identified in *BRCA1*

The pluriformity of CNVs was analyzed in detail in the subcohort of 101 CNVs in *BRCA1*. While read depth analysis may call replicates with slightly different boundaries, CNVs with both breakpoints differing by less than 1,000bp were considered identical. This resulted in at least 57 distinct CNVs, for which breakpoints frequently colocalized with Alu repeats[2] (grey track) and segmental duplications in *BRCA1* exons 1-2 and the upstream gene *NBR2* (light blue track). Out of the 15 duplications detected in *BRCA1*, 12 were confirmed to be in tandem (hatched boxes indicate that breakpoints have not been resolved). This knowledge is required to predict the RNA sequence, which guides variant classification.

References

1. Weischenfeldt, J. et al. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
2. Sluiter, M. D. & van Rensburg, E. J. Large genomic rearrangements of the BRCA1 and BRCA2 genes: review of the literature and report of a novel BRCA1 mutation. *Breast Cancer Res. Treat.* **125**, 325–349 (2011).
3. van der Klift, H. et al. Molecular characterization of the spectrum of genomic deletions in the mismatch repair genes MSH2, MLH1, MSH6, and PMS2 responsible for hereditary nonpolyposis colorectal cancer (HNPCC). *Genes Chromosomes Cancer* **44**, 123–138 (2005).
4. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
5. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
6. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
7. Layer, R. M. *et al.* LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
8. Talevich, E. *et al.* CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
9. Lim, Y. K. *et al.* Identification of novel BRCA large genomic rearrangements in Singapore Asian breast and ovarian patients with cancer. *Clin. Genet.* **71**, 331–342 (2007).