Technical Note

Exon-level CNV Detection in BRCA1/BRCA2 genes using Amplicon based CleanPlex NGS kits

1. Overview

CNVs, including heterozygous and homozygous deletions and duplications, are well-established contributors to hereditary breast and ovarian cancer risk. Amplicon-based next-generation sequencing (NGS) has become a powerful tool for detecting exon-level copy number variations (CNVs) in clinically relevant genes such as BRCA1 and BRCA2.

Compared to hybrid capture sequencing approaches, amplicon-based enrichment provides high uniformity of coverage even using reduced input DNA requirements, with cost-effective workflows. The Paragon Genomics CleanPlex BRCA1 & BRCA2 Kit v3 is specifically designed to generate consistent coverage across exons, enabling high-resolution CNV detection at the exon level.

However, raw read counts alone cannot be directly interpreted for CNV detection due to variability introduced by amplification efficiency, GC content, and sequencing depth differences across samples. Therefore, robust normalization strategies are essential to distinguish true biological copy number changes from technical noise. When properly normalized, read depth serves as a quantitative proxy for copy numbers, allowing identification of exon-level deletions and duplications with high sensitivity and specificity.

2. Method

The CNV detection procedure implemented in this study is based on a read-depth normalization approach applied to amplicon sequencing data. The analysis begins with extraction of raw read counts per amplicon for each sample.

The method consists of two primary normalization steps. First, intra-sample normalization is performed to correct for differences in total sequencing depth across samples. This is achieved by scaling each amplicon's read count relative to the median or total read count within the same sample. This step ensures that each sample is normalized to a comparable baseline. Optionally, amplicons can be separated into various categories by combinations of properties such as genomic locations, amplicon length, GC content, pools, etc., so that intra-sample normalization can be performed separately to reduce category-specific bias.

Second, inter-sample normalization is applied across samples for each amplicon. This is accomplished by dividing each intra-sample-normalized value by the median value of that amplicon across all samples. This corrects for amplicon-specific biases that are not considered in the previous step.

The resulting normalized ratios are centered around 1.0 for diploid regions. Deviations from this baseline indicate potential CNVs. Ratios near 0.5 are indicative of heterozygous deletions, while ratios in the range of 1.5–2.0 suggest duplications.

Threshold-based classification considering numbers of contiguous amplicons and between replicates is then applied to assign copy number states. The results are compared against known reference samples to validate accuracy. This approach leverages the high depth and reproducibility of amplicon sequencing to achieve reliable CNV detection.

A study was performed using samples corresponding to well-characterized Coriell reference materials, including NA18507, NA14094, NA14626 and NA18949. These reference samples provide known CNV events that serve as ground truth for validating the detection method. Libraries were constructed using CleanPlex BRCA1 & BRCA2 Kit v3 from Paragon Genomics. They were sequenced as a part of an Illumina MiSeq run with the Micro sequencing kit, which provides sufficient depth for targeted somatic applications.

3. Results

Detailed in the accompanying xlsx file, the dataset used in this analysis consists of 8 sequencing samples, labeled S9 through S16, 2 replicates for each blinded Corielle cell line.

Columns B through I in the dataset represent raw read counts for each amplicon across the eight samples. Each row corresponds to a specific amplicon targeting an exon of BRCA1 or BRCA2. The amplicon identifiers include gene name and amplicon ID, enabling direct mapping of read counts to genomic regions.

Hidden columns O-V are results of the first normalization step - category-factored intra-sample normalization. This step adjusts for differences in total sequencing depth between samples. In this particular exercise, amplicons are separated into 4 categories based on length (longer vs shorter than 140bp) and pools (1 or 2). This step is expressed mathematically as:

Normalized read count values (intra-sample) = raw read count/median (raw read counts for amplicons in the same category for the same sample)

Columns Y-AF show results of the second normalization step, inter-sample row normalization, correcting for remaining amplicon-specific biases. For each amplicon, the intra-sample normalized values are divided by the median value across all samples. This step accounts for systematic differences in amplification efficiency and probe performance.

Normalized ratio (final) = Normalized read counts value (intra-sample) / median (normalized read count values across all samples for that amplicon)

The final normalized ratios were evaluated to identify exon-level CNVs and compared against known reference profiles from Coriell samples. According to the reference information shown in Figure 1, NA14094 contains a heterozygous deletion in BRCA1 exon 10, NA14626 contains a heterozygous duplication affecting the BRCA1 exon 12, and NA18949 contains a heterozygous deletion affecting BRCA1 exons 14 and 15.

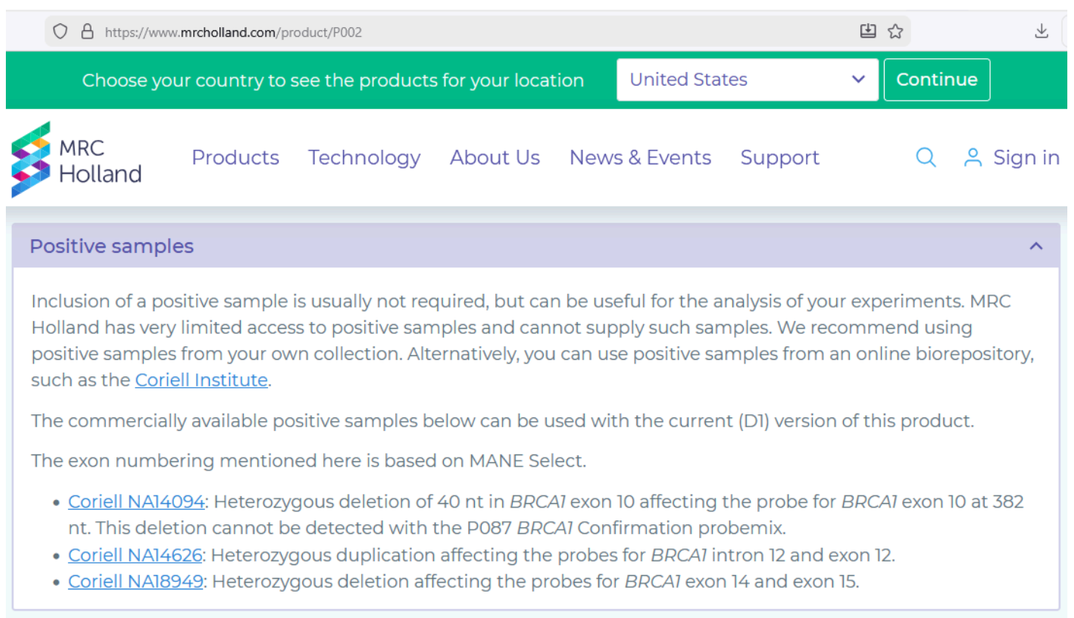

Fig 1. CNV annotations of 3 Coriell cell lines on the MRC Holland website.

The CNV truth was matched to the observed normalized ratios, which revealed the identities of the samples sequenced:

- Both samples corresponding to NA18949, S11 and S12, show reduced normalized ratios ~0.5, across BRCA1 exons 14 and 15.
- Both samples corresponding to NA14626, S13 and S14, show an elevated normalized ratio to ~1.5 in the region, spanning BRCA1 exon 12.
- Both samples corresponding to NA14094, S15 and S16, show a reduction in normalized ratios to approximately 0.5, at BRCA1 exon 10.

The changes in raw read counts and normalized ratios are plotted in the examples below - S11 with the heterozygous deletion in amplicons indexed 16-24, and S13 with the heterozygous duplication in amplicons 29-33. There was no local smoothing applied to the curves.

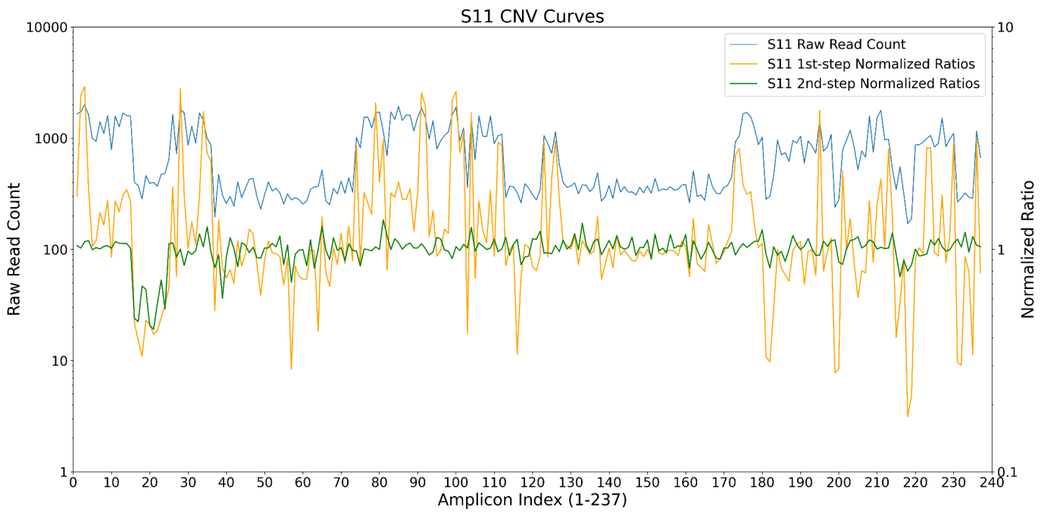

Fig 2. Un-smoothed curves of raw read counts and normalized ratios for CNV detection in Sample 11, NA18949.

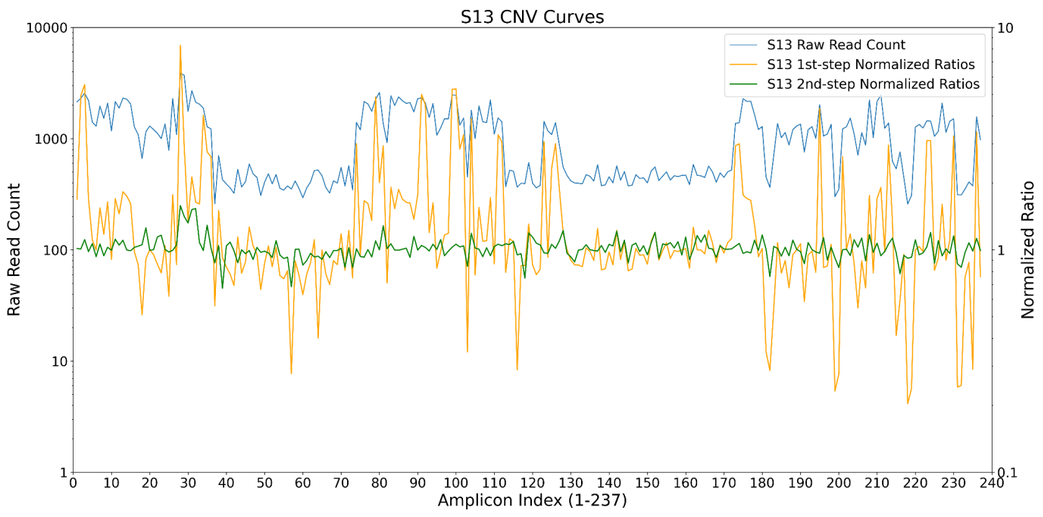

Fig 3. Un-smoothed curves of raw read counts and normalized ratios for CNV detection in Sample 13, NA14626

In addition, sequencing reads revealed multiple lines of orthogonal evidence for CNV, independent of read counts. Please see the IGV screenshots below.

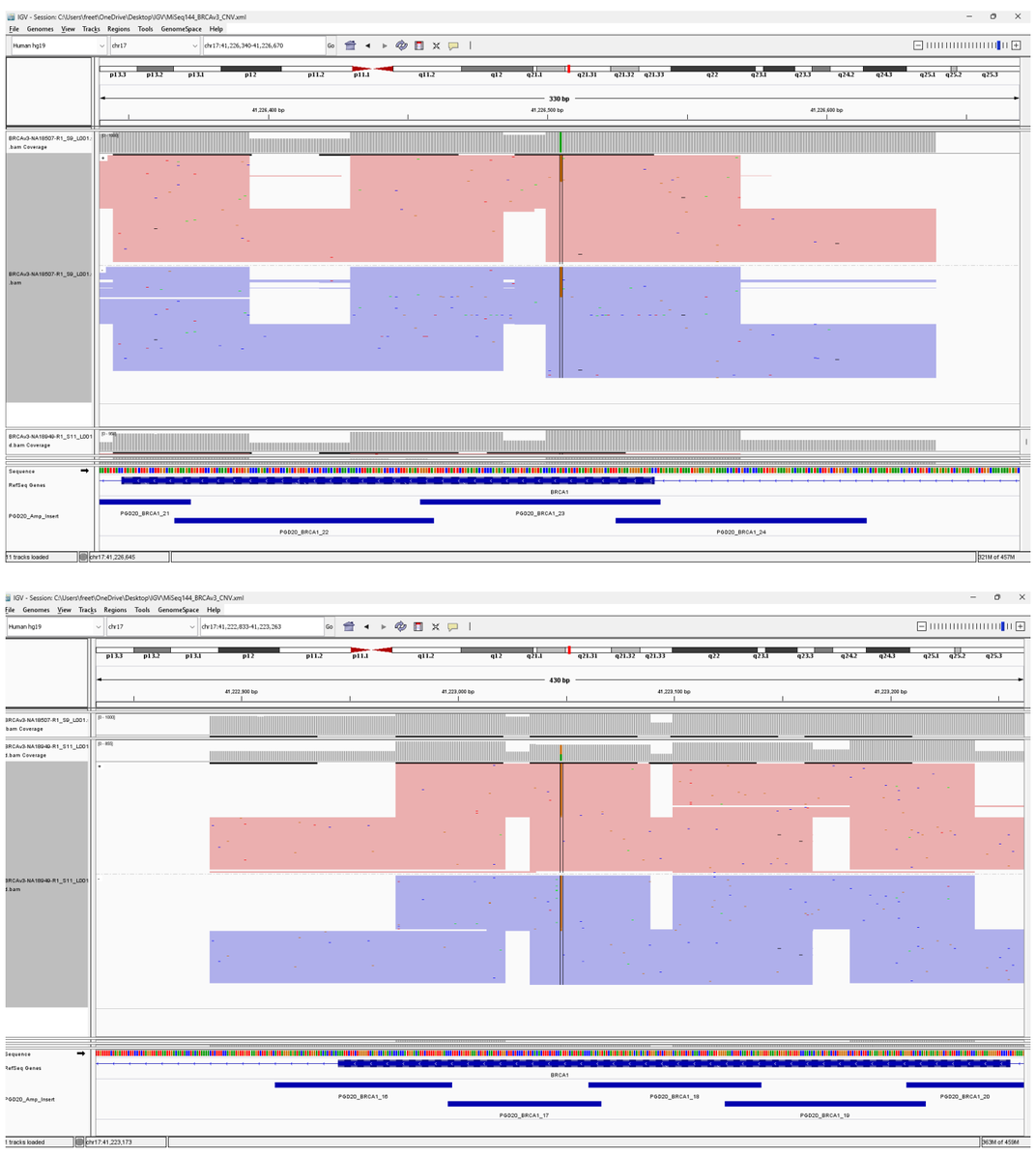

Fig 4. Two IGV (Integrated Genome Viewer) screenshots showing a heterozygous SNP in BRCA1 exon 14 (top), and a homozygous SNP in exon 15 (bottom).

In figure 4, a variant locus for NA18507 in BRCA1 exon 14 is shown as homozygous on the IGV coverage track. It is corrected as heterozygous after primer bases of the amplicon PGD20_BRCA1_24 are trimmed. In contrast, a locus in exon 15 shown as heterozygous on NA18949 coverage track is actually homozygous on the figure. The two loci in combination corroborate loss of heterozygosity between NA18507 and NA18949, across BRCA1 exons 14 and 15. The longer a heterozygous deletion, the stronger evidence LOH will become.

b) Reads soft-clipping and gapped alignment in NA14094:

Fig 5. An IGV screenshot showing reads soft-clipping due to the presence of a 40-bp deletion (left-justified as in the purple rectangle).

The soft-clipping above in figure 5 helps determine the exact start and end of the 40-bp deletion, as marked by the purple rectangle with left-justification. Reads with the soft-clipping come from super-amplicons shown in the zoom-out view below in Figure 6.

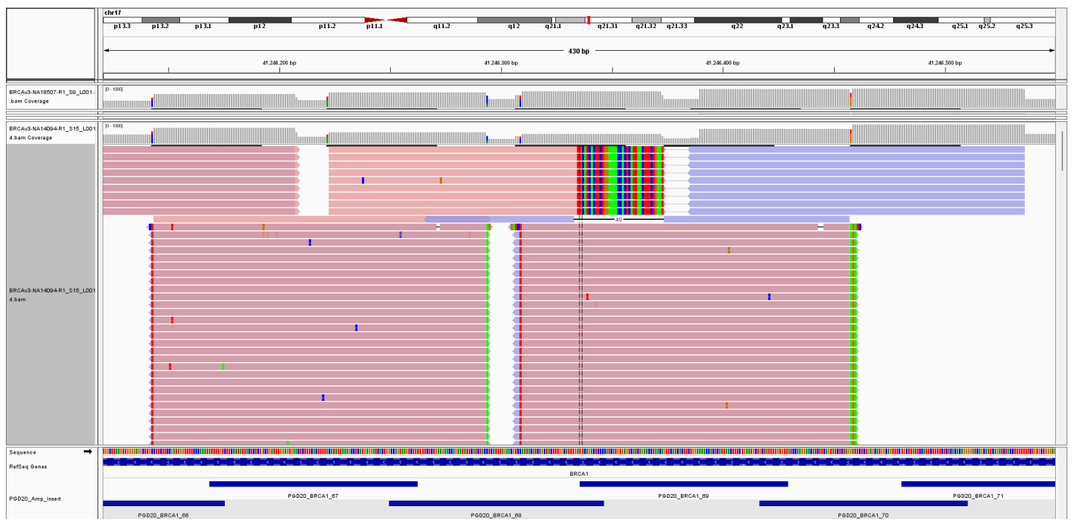

Fig 6. An IGV screenshot showing paired reads in the region zoomed out from Fig. 5.

A super-amplicon is formed in Pool 1 by the forward primer of PGD20_BRCA1_68 and the reverse primer of PGD20_BRCA1_70, which produced R1 reads with 39bp soft clipping. In Pool 2, another super-amplicon formed by the forward primer of PGD20_BRCA1_67 and the reverse primer of PGD20_BRCA1_69 produced an R2 read with a 40bp gap in alignment. Such evidence is observed specifically in the NA14094 samples and particularly useful in detecting relatively short deletions.

4. Conclusions

This study demonstrates the feasibility of detecting exon-level CNVs using amplicon-based NGS data from Paragon Genomics CleanPlex BRCA1 & BRCA2 kit v3. The combination of high sequencing depth, robust normalization, orthogonal loss of heterozygosity and alignment evidence from super-amplicons enables reliable identification of deletions and duplications at single-exon level. This method provides a scalable and cost-effective solution for integrating CNV detection into routine clinical sequencing pipelines.