Isoform Fraction Distribution

The isoform fraction distribution is determined by the box plots for the 5’ and 3’ ends. The data suggests that the values are mainly skewed with a concentration of values between 0.6 to 1.0. Thus, for the majority of the chimeric transcripts, a dominant isoform contributes disproportionately at each of the breakpoint termini. This specific dominance may arise from biologically relevant transcript preferences such as preferential exon usage, transcript stability, or artifacts from uneven sequence coverage and biased read mapping. The median for the isoform fraction at both ends is 0.78, which implies that there is a high representation of a single isoform that occurs within each event. Although the outliers may reveal a potential presence of the true isoforms' diversity or, likewise, a partial or chimeric degradation. The distribution evaluates the credibility of the chimeric junctions since lower isoform fractions can correlate with transcriptional noises or alternative splicing complexities.

Descriptive Statistics: Isoform Metrics and Fragmentation

The descriptive statistics for the isoform metrics and fragmentation analysis provide the essential characteristics of the structure. The calculated mean for the isoform fractions for the 5’ and 3’ ends is relatively high, around 0.75 to 0.80, which emphasizes the dominance of single isoforms for each event. However, the standard deviation is quite small, which indicates that the tight distribution limits the variability may align with either sequencing or annotation bias. Total fragments and spanning read counts reveal the depth of sequencing and the transcript integrity. Since the 25th percentile of the spanning fragments is nearly zero, it suggests that most junctions are poorly supported at the read level. A weak read support indicates that the fusion events are lowly expressed or template switching has occurred.

Correlation Matrix Analysis

The correlation matrix was computed for all the numeric variables for the chimera score, distance between breakpoints, total and spanning fragments, and isoform fractions. A strong positive correlation was determined between the total fragments and the spanning fragment, with > 0.9, emphasizing the sequencing depth and transcript coverage. Furthermore, there is a negative correlation between the genomic distance and both the isoform fractions as well as the spanning fragments. This may indicate that a longer distanced fusion is more inclined to have lower isoform dominance and a reduced fragmentary support from mapping difficulties or a lower production of stable transcripts. Additionally, at the isoform fractions between the 5’ and 3’ ends $r \approx 0.83$, thus indicating that they are highly correlated since there is a stable isoform pattern across both ends of the fusion events. In all accounts, the chimera score demonstrated a relatively weak correlation with the tested variables; this may be because there are complex non-linear features that do not suggest much to the read or isoform counts.

Genomic Distance Between Breakpoints

The distribution for the genomic distances between fusion breakpoints is highly skewed to the right, where most events occur within a close genomic proximity, and a long tail extends towards inter-chromosomal/megabase-scale rearrangements. A sharp peak at zero suggests that there is a preponderance of intra-genic or tandem duplications, and the long tail indicates that there may have been translocations or trans-splicing, however, further validation would be necessary to determine. Events with large genomic distances must be carefully evaluated, as they have the potential to indicate chimeric artifacts from sequencing or alignment errors with low-complexity or repetitive genomic regions. The resulting distance distributions assist in classifying fusion types, such as cis-splicing vs. translocations, and selecting candidates for the next validation.

Chimera Cluster Metrics: Cluster Size vs. Confidence Score

The chimera cluster metrics analyze the profound relationship between the chimera cluster size and the average confidence score. There is a narrow variability of the clusters, where they are relatively uniform in their size and normalized around 1. Regardless of this, the confidence score is quite variable; some of the single-read clusters have achieved high scores, suggesting that the score model that assigns the chimera confidence integrates qualitative metrics such as breakpoint consistency, isoform coherence, or base quality. Analyzing the vertical spread in confidence across such a fixed cluster size emphasizes that cluster size is a poor predictor for biological validity. In regard to breast cancer transcriptomics, it is an indication for necessary multi-parametric evaluation of clusters to ensure that, other than statistical support, there is also oncogenic relevance of transcript structure.

Score vs. Number of Spanning Reads

The scatter plot visualizes the relationship between the chimera confidence score and the number of breakpoint-spanning reads. The data reveals $r \approx 0.04$, this weak positive correlation suggests that the confidence score is slightly influenced by the raw read support. Vindictively, more supporting reads and higher confidence may be expected, however, this occurs when the scoring model integrates complexities in sequence homology, junction entropy, or breakpoint annotation. Though some chimeras that have few spanning reads exhibit surprisingly high scores, whilst others containing more read support have been penalized. The disparity conveys the variable limitations of relying on the sole read counts to determine transcriptomic fusion credibility and undermines the need for reliability in detection algorithms.

Analysis of Unique Chimera Clusters and Partner Genes

The dataset contains 318 unique chimera clusters, indicative of high-level transcript diversity of the breast cancer samples. The 5’ genes and 3’ genes reveal that a large proportion of genes are involved in fusion events; however, not all equally participate as donors and acceptors. This slight difference suggests that certain genes are inclined to either be the 5’ or 3’ partner in chimeras, which may be a cause of sequence features, genomic positioning, or transcriptional activity. The imbalance may induce selective pressures in cancer cells, whereas specific fusion combinations confer growth advantages.

Analysis of Chimera Type Distribution

Interchromosomal chimeras are the most dominan,t which may suggest frequent trans-chromosomal rearrangements in breast cancer. The pattern parallels known cancer hallmarks of genomic instability and chromothripsis, where chromosomes experience massive breaks and erroneous repairs. A high prevalence of read-through chimeras reveals transcriptional read-through events, a cause of dysregulated termination mechanisms in cancer cells.

Intrachromosomal complexes and standard intrachromosomal chimeras appear to be frequent, indicating localized genomic rearrangements. The presence of orientation-specific subtypes such as converging or diverging, likely suggests that fusion formation is influenced by the spatial organization of genes and transcriptional directionality.

Overlapping and adjacent chimera types may arise from any tandem duplications or complex local rearrangements, as their lower frequency implies that the events are disfavored and unlikely to occur under selective pressures in breast cancer.