Structural variant detection using image-based machine learning and on-flow cell proximity information

Gavin Parnaby[1], Mitchell A Bekritsky[1,*], Jeff Gau[1], Vitor Onuchic[1], Drew Ellershaw[2], Mohammed Hashemi[1], Rishi Verma[1], Pascal Grobecker[2], Gery Vessere[1], Arun Subramaniyan[1], Rami Mehio[1], James Han[1]

[1] Illumina Inc., San Diego, California, USA; [2] Illumina Inc., Cambridge, Cambridgeshire, UK

SCAN ME

INTRODUCTION

Constellation mapped read technology is a novel approach leveraging on-flow-cell library preparation that utilizes proximity information from neighboring nanowells to generate long-range genomic insights from standard SBS sequencing.

Constellation mapped reads identify complex structural rearrangements by counting proximal clusters between any pair of genomic regions, termed ‘colocation’.

HOW IT WORKS

Constellation mapped read technology uses flow cell-bound transposomes to eliminate library prep by capturing and tagmenting long DNA molecules as they flow across the flow cell surface. All downstream clustering and SBS steps are maintained. Due to on-flow-cell tagmentation, nearby clusters contain reads from the same original input molecule, enabling recovery of long-range information (Figure 1, see poster 569 438).

High-resolution visual representations of genome structure, termed colocation maps, can then be generated by extracting information about reads from proximal clusters between any pair of genomic regions. These maps divide the genome into bins and count reads in neighboring clusters for each possible pair of genomic bins. Large numbers of reads from neighboring clusters occur almost exclusively when those bins are in close genomic proximity.

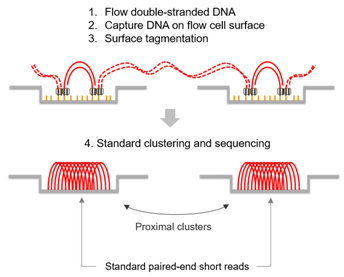

Figure 1. Double stranded DNA flows onto the flow cell. DNA is captured on the surface and tagmented. Clusters that originate from the same DNA template molecule are nearby on the flow cell surface.

In regions with no structural variants, bins that are nearby in the reference genome are nearby in the sample and appear as a diagonal line in a colocation plot. In regions with structural variants, nearby bins in the reference genome are no longer nearby in the sample and exhibit off-diagonal signals (Figure 2). Image recognition methods can subsequently be trained on a combination of simulated and real data to automatically classify the rearrangement type and provide additional information about the event.

Figure 2. Simplified diagrams, simulated colocation matrices, and examples from HG002 / NA24385 are shown for various structural rearrangements. The leftmost diagram of each subplot shows the colocation count for each bin-pair, with the boxes underneath the matrix representing the genome bin ordering indicated by the matrix. The center colocation plot shows a simulated homozygous example of a specified structural rearrangement, and the rightmost plot shows an example of the rearrangement in HG002 / NA24385 that is either heterozygous or homozygous. The inversion is not from HG002. A: No SV; B: Heterozygous insertion; C: Homozygous deletion; D: Heterozygous tandem duplication; E: Heterozygous inversion.

IMAGE RECOGNITION-BASED SV CLASSIFICATION

Deep learning object detection algorithms applied to the colocation matrices are used to classify large structural variants. The backbone network uses convolutional layers to detect event signatures - for example inversions generate ‘butterfly’ shapes off-diagonal with a location that identifies the event breakpoint. Output heads classify SV event type, position and size based on the detected motifs and relative orientation (Figure 3). The approach includes an anomaly detector algorithm, identifying regions in the genome with anomalous link information that cannot easily be classified – including complex balanced rearrangements.

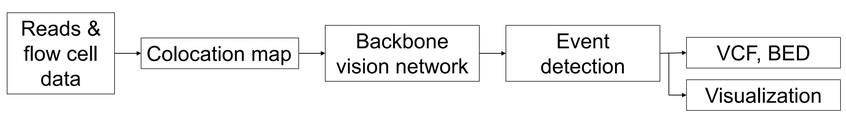

Figure 3. A diagram of the image recognition module, including inputs and outputs, that can be trained to classify structural variants from colocation data.

SIMULATION ADDRESSES TRUTH SET CONSTRAINTS

Figure 4 shows several outputs from the detection pipeline. The figure illustrates simulated events (including a balanced inversion) as well as real examples – a heterozygous deletion from the HG002 truth set, and an inversion. The white box indicates ground truth while the black box shows the output from the AI detection pipeline – in these examples, they overlap closely, indicating close concordance with truth. The network outputs event type, position, dimension and confidence score. The object detector can identify balanced events including inversions and inter-and intra-chromosome translocations.

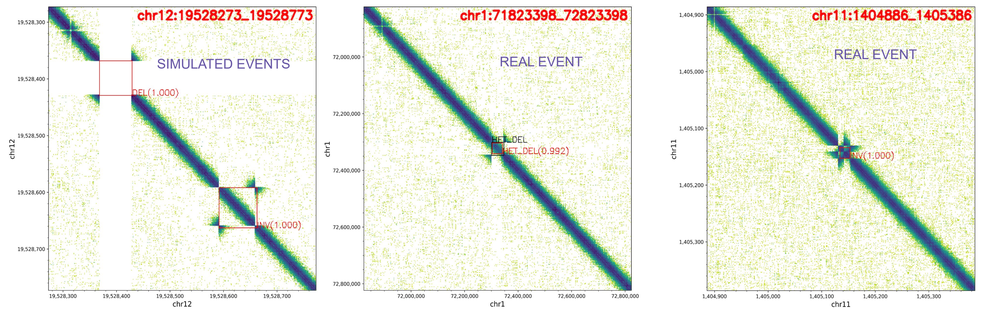

Figure 4. Colocation maps and object detector outputs a) multiple events (hom-del and inv) b) het-del c) het-ins

After identification, large structural rearrangements and anomalous regions are highlighted on the colocation map for easy visual analysis. We detect structural variants in difficult-to-map regions, across the full spectrum of size and position, from 10kbp up to chromosomal scale events. Detected SVs can be refined by read information, allowing base-pair breakpoint resolution.

The pipeline will be integrated into DRAGEN enabling full genome analysis in reasonable time-scales, with standard VCF output format.

COMPLEX REARRANGEMENTS

Colocation matrices provide distinctive signals for complex rearrangements that can have significant clinical impact, including translocations, ring chromosomes, and others. Colocation matrices generated from Coriell samples with well-characterized genomic anomalies exhibit signals for these complex rearrangements that would be undetectable with standard SBS in figures 5 – 7.

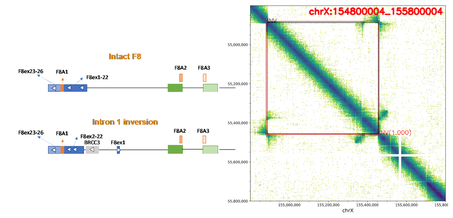

Figure 5. A colocation matrix from an individual with severe hemophilia A, caused by an inversion in intron 22 of *F8*. The inversion is characterized by the butterfly pattern in the matrix, indicating that the first exon and BRCC3 have been inverted. Individuals with severe hemophilia A suffer from spontaneous bleeding episodes due to minimal clotting activity.

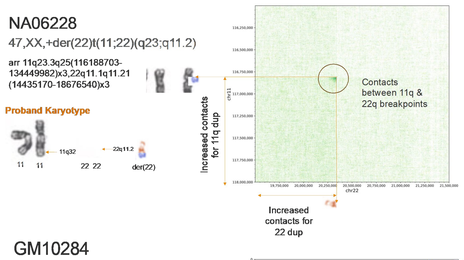

Figure 6. A colocation matrix from an individual with derivative 22 syndrome, also known as Emanuel syndrome, indicating increased colocation signal between 11q and 22q due to an unbalanced translocation of the two chromosomes. Individuals with derivative 22 syndrome suffer from hypotonia, developmental delay, and other congenital abnormalities.

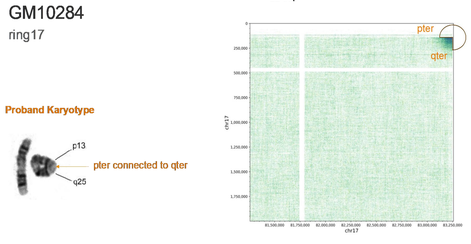

Figure 7. A colocation matrix from an individual with ring chromosome 17, indicated by increased colocation signal between the beginning and end of the chromosome in the colocation matrix. Symptoms in individuals with ring 17 may vary from lissencephaly and severe intellectual disability to short stature, microcephaly, etc.

FUTURE DIRECTIONS

Karyotyping, gene panels, microarray analysis, DNA sequencing, and specialized assays have all been traditionally employed to detect balanced and unbalanced structural rearrangements in the genome. These assays have several limitations, including variable robustness for different classes of rearrangements, targeted assays, long runtimes, and needing multiple tests to fully characterize a sample’s genome structure. Whole genome sequencing using constellation mapped reads with colocation analysis can detect all major structural variant types, including insertions, deletions, inversions, translocations, and other, more complex events, in a single, unbiased assay.

For Research Use Only. Not for use in diagnostic procedures.

illumina®