Evolution of Structural Variation in Cancer Revealed by Read Clouds

Noah Spies, Ziming Weng, Alex Bishara, Justin M Zook,
Robert B West, Marc Salit, Arend Sidow

Background Structural variants, particularly distant translocations, are difficult to identify despite their fundamental importance in cancer and other diseases. Because any two genomic loci can be connected through a genomic rearrangement or translocation, the search space for structural variation is proportional to the square of the genome size, resulting in a massive multiple-testing problem for mammalian genomes. Even though current short-read technologies have very low rates of chimeric molecules and mismapping to the genome, these types of experimental and computational errors compound to result in high rates of false positives when searching genome-wide for structural variation. Furthermore, standard sequencing reads derive from short genomic fragments typically only several hundred base pairs in length, and thus cannot map uniquely to translocation breakpoints occurring in even moderately long repeat sequences.

Results The 10X Genomics platform generates barcoded short-reads from large genomic DNA fragments, which can then be clustered in silico to generate read clouds identifying the original large DNA fragments. We size-selected large (50–100kb) genomic DNA fragments from 7 spatially distinct tumor samples from a single sarcoma, as well as matched normal tissue, then applied the 10X platform to generate read clouds.

We have implemented new methods to identify structural variants from these read cloud data. We use the read cloud barcodes to identify candidate events where the similarity in barcode patterns between two loci is higher than expected given the distance between the loci. We then perform breakpoint refinement using the patterns of dropoff in observed long fragment density at the structural variant breakpoints.

Using this new method, we find structural variants that differ between sectors of the sarcoma, although most somatic structural variants (and

single-nucleotide variants) are shared across all samples in the tumor. Multiple, independent, ancestral chromothripsis events occurred in our sarcoma case, totaling hundreds of individual breakpoints shared between sectors.

To better understand these bursts of genome rearrangement, we have implemented a novel approach using patterns of read clouds to automatically reconstruct the order and orientation of complex structural variants involving many breakpoints. Furthermore, using the read cloud barcodes, we are able to identify all reads supporting a structural variant and assemble the full sequence of many of these complex structural variants (although this is still dependent on the local sequence complexity). This approach reveals that many of the complex structural variants involve the rearrangement of many short (several kb) genomic segments derived from distant locations on the same chromosome, forming new chromosomes. In the process of creating these neochromosomes, large intervening genomic segments are lost, resulting in a loss of heterozygosity.

Conclusions By harnessing the barcoded sequencing platform, we are able to phase and assemble complex genomic rearrangements, illuminating larger patterns of genome evolution in cancer. Because the read clouds derive from long DNA fragments, physical coverage of each breakpoint is substantially higher than for standard short-read data, resulting in a much higher signal-to-background. This approach is also able to identify structural variant breakpoints occurring in repetitive genomic regions, and can actually assemble the nucleotide sequences of these events. Finally, our results demonstrate that even very large (in this case, over 20 cm in length) tumors need not show substantial subclonal diversity, and that rather a series of extreme genomic rearrangements occurred early in tumor development.