Towards understanding the genomic architecture of cancer genomes

Ernest T Lam[1], Alex R Hastie[1], Marcin B Imielinski[2], Cheng-Zhong Zhang[2], Jeremiah Wala[2], Zeljko Dzakula[1], Han Cao[1]

[1]BioNano Genomics, Research and Development, San Diego, CA
[2]Broad Institute of Harvard and MIT, Center for Biomedical Informatics, Cambridge, MA

Abstract

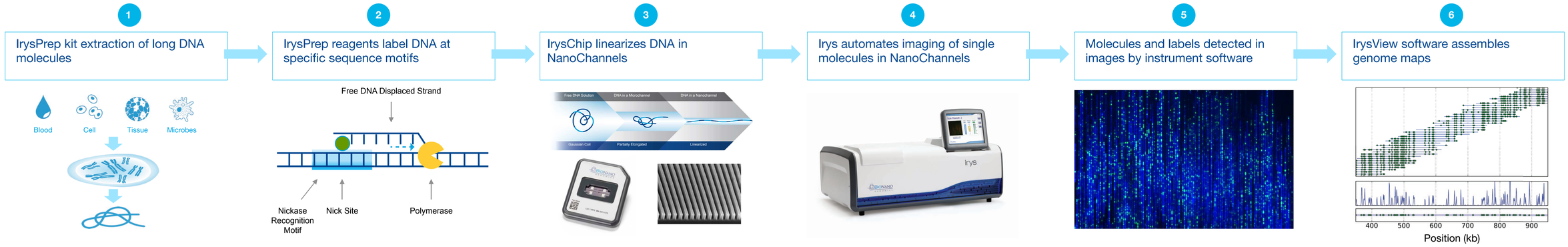
Understanding the genetic architecture of cancer requires whole-genome and integrative approaches. Cancers often feature genomic alterations that range from single-base changes to large-scale structural rearrangements. Having a complete catalogue of mutations in cancer is crucial for identifying key drivers and providing accurate diagnosis, prognosis and targeted therapy. Next-generation sequencing (NGS) platforms have limited power to decipher large, complex structural variants frequently observed in cancer. Genome mapping represents a complementary technology that provides critical structural information. It involves high throughput analysis of single molecules spanning hundreds of kilobases in nanochannels. Long-range information is preserved and direct interrogation of complex structural variants made possible. Therefore, leveraging the strengths of these complementary platforms would give a comprehensive view of a cancer genome.

Here, we present our analysis of well-studied and highly rearranged cancer genomes. We constructed completely de novo genome map assemblies with N50 lengths of more than 1 Mb. We derived multi-sample normalized copy number profiles of matched tumor-control pairs based on genome mapping data. We observed that tumor samples had highly variable copy number profiles, corresponding to focal and chromosome-scale changes. Copy number breakpoints were shown indicative of translocation events. We also present a pipeline to integrate NGS and genome mapping data to validate and refine translocation calls. Genome mapping data helped bridge and phase neighboring translocation events. Finally, we present a computational approach to identify translocations by clustering single molecules with abnormal alignment to the reference and by performing local assemblies of these molecules. Overall, integrating NGS and genome mapping data provides a comprehensive view of a cancer genome.

Background

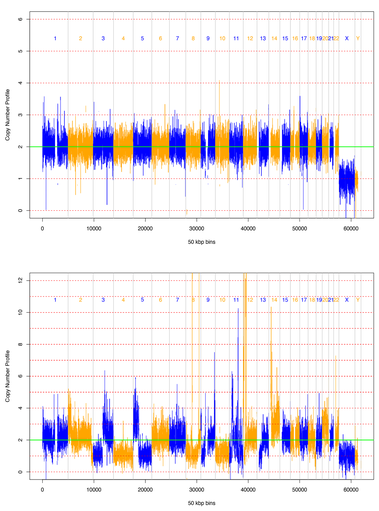
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing the statistical inference needed to align paired-end reads with an uncertain insert size distribution. These long labeled molecules are *de novo* assembled into physical maps spanning the whole genome. The resulting order and orientation of sequence elements in the map can be used for anchoring NGS contigs and structural variation detection.

Methods



(1) Long molecules of DNA are labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® nanochannels and single molecules are imaged by Irys. (4) Single molecule data are collected and detected automatically. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Maps may be used in a variety of downstream analysis using IrysView® software.

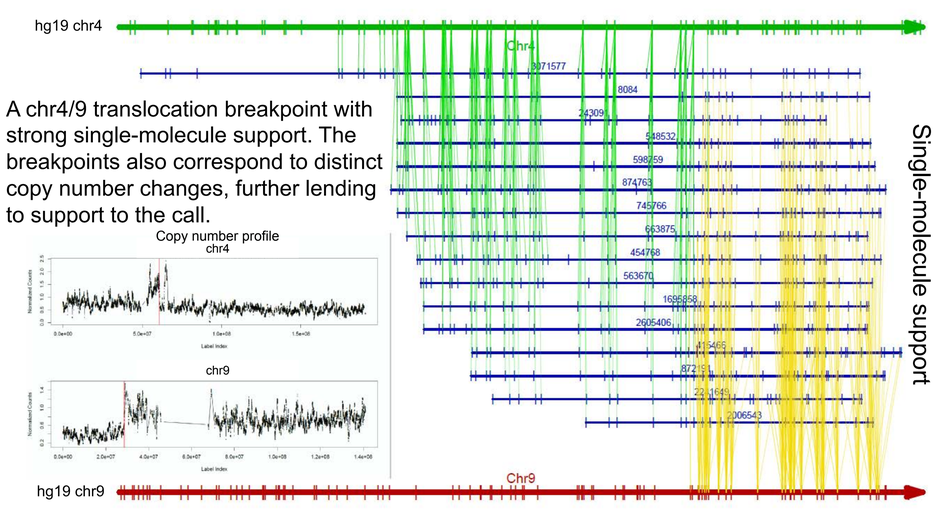
Normalized copy-number profiles



Copy number profiles were derived based on single-molecule alignment to reference. They can reveal large-scale aberrations, localize breakpoints of large SV events, guide SV detection to regions of interest, and detect aneuploidy.

As shown, the copy number centers at around 2 throughout the genome in the diploid control sample. Since it was male, there is half the coverage of chrY. In the cancer sample, the copy number fluctuates dramatically compared to the control sample.

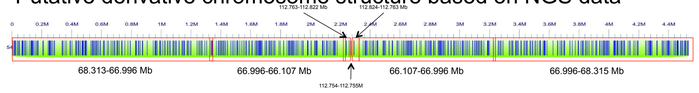
Identification of translocation events



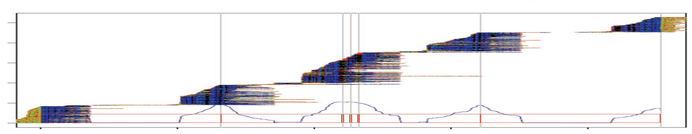
A chr4/9 translocation breakpoint with strong single-molecule support. The breakpoints also correspond to distinct copy number changes, further lending to support to the call.

Phasing of translocation breakpoints

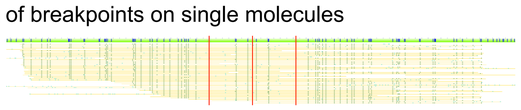
Putative derivative chromosome structure based on NGS data



Validation of breakpoints by single-molecule mapping



Phasing of breakpoints on single molecules



Single molecules map to each of four fragments of chromosome 12, supporting the structure of the derivative chromosome.

Conclusions

Analysis of cancer genomes remains one of the major challenges in genomic analysis due to their genome complexities. Current NGS approaches can detect individual translocation breakpoints but are limited by their short read lengths and insert sizes. Genome mapping provides additional long-range information about the genome structure. Genome mapping of a human individual can currently be accomplished today with three chips or less in a few days. The throughput will improve such that a single chip would generate high-coverage data for analysis of a cancer genome in less than one day, making high-coverage analysis of cancer genomes possible.

Reference

1. Cao, H., et al. Rapid detection of structural variation in a human genome using nanochannel based genome mapping technology. GigaScience (2014); 3(December 2014): 34.
2. Hastie, A.R., et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex aegilops tauschii genome. PLOS ONE (2013); 8(2): e55864.
3. Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303.