Mapping the "Dark Matter" of Genome – Long Repeats, Structural Variations in Complex Cancer Genome with Nanochannel Technology

A Hastie[1], E Lam[1], A Pang[1], W Andrews[1], T Anantharaman[1], T Chan[1], M Saghibini[1], H Sadowski[1], M Austin[1], Z Dzakula[1], E Holmlin[1], X Xun[2] , P Kwok[3], M Rossi[4], Han Cao[1]

[1]BioNano Genomics, San Diego, CA, USA; [2]BGI, Shenzhen, China; [3]UCSF, San Francisco, CA, USA; [4]Emory University School of Medicine, Athens, GA, USA

Abstract

In spite of advancement in high-throughput short read next-generation sequencing in the past decade, majority of genome analysis today were limited in exome or transcriptome sequencing, a fair portion of the human genome remains unresolved or ambiguously characterized. Especially, large genomic structural variations (SV, > 1 kb), known to be associated with complex traits and diseases such as cancer, are found more prevalent than we previously thought. During assembly, they leave gaps and unknown structural or heterozygous information as the "dark matter" of the genome, often challenging to detect for short read NGS and conventional labor intensive low resolution cytogenetic techniques. Rapid whole genome mapping in nanochannel arrays represents a new standard of single-molecule platform independent of yet complementary to DNA sequencing for accurate genome assembly and structural variation analysis in complex cancer genome. Extremely long intact DNA molecules of hundreds to thousands of kilobases fluorescently labeled at sequence motifs and linearized in true nanofluidic channels enable direct image interrogation of comprehensive genome architecture at a high resolution. De novo assembly of these single molecules yields unprecedented long contiguous genome maps, advantageous in spanning over highly repetitive regions and complex structures in their native form. The first hybrid assembly of a diploid human genome combines NGS sequencing with Bionano genome maps, resulting in excellent scaffold N50s exceeding 34 Mb, a step forward towards "medical grade" genome and dramatic improvement upon the current assembly contiguity typically observed in using shotgun sequencing approaches alone.

We present here results from analysis on human cancer genome, non-model and large complex genomes. We detected hundreds of large structural variants per genome and haplotype differences in these genomes, revealing the locations, orientations and copy numbers of these complex structural variants often elusive from NGS detection. Using Bionano Irys® platform, we were able to find and characterize SV and repeat regions in the complex cancer genome that are biologically and clinically relevant, especially able to precisely map viral (oncogene) component integration sites within human genome, believed to be linked to oncogenesis.

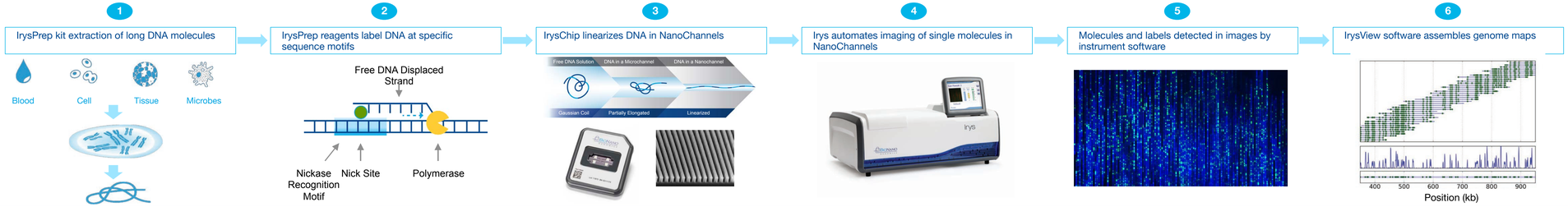
Comprehensive genome mapping provides extreme valuable structural information otherwise hard or impossible to decipher with short read sequencing data alone, and paves the road for generating true golden standard in ultimate medical grade and breeding grade genome information. For the first time, population scale comparative genome study to identify comprehensive genomic structural variation on a single platform with a large patient cohorts is feasible due to the standardized high quality data, automated efficiency of data generation and low cost.

Background

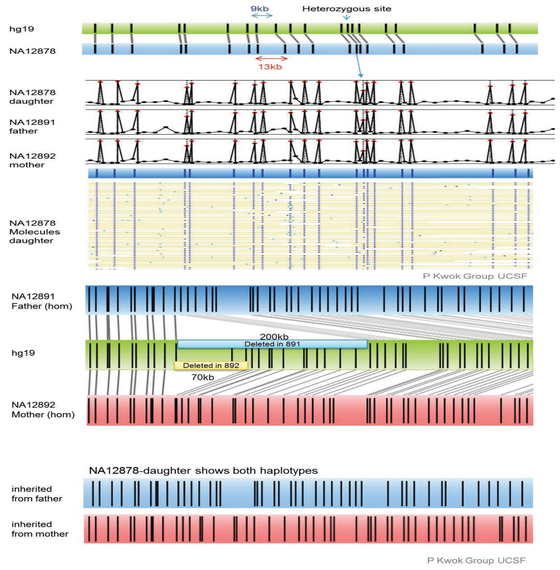
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short-read sequencing technologies alone. The Irys System provides direct visualization of long DNA molecules in their native state, bypassing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

Methods

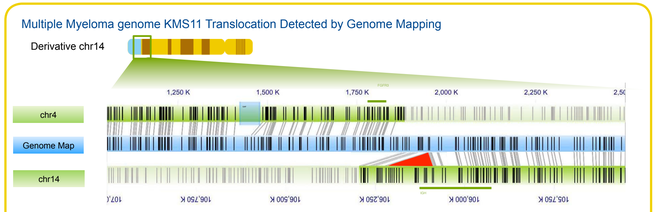
(1) Long molecules of DNA are labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® nanochannels and single molecules are imaged by Irys. (4) Single molecule 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.



Structural Variations in CEPH Trio Analysis- large Insertion haplotypes on chr19

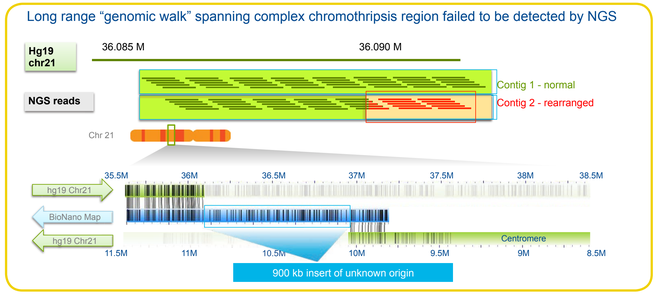


Multiple Myeloma genome KMS11 Translocation Detected by Genome Mapping

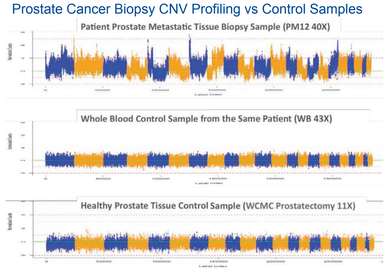


Our data shows that KMS11 is positive for t(4;14), a known variant in this cell line. The FGFR3 can be dysregulated in myeloma by the translocation, as it brings the gene into the vicinity of IgH enhancers. The deletion in the IGH locus could be incidental to KMS11 or simply a common SV in human population.

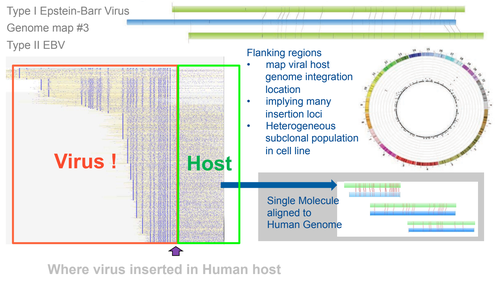
Long range "genomic walk" spanning complex chromothripsis region failed to be detected by NGS



Prostate Cancer Biopsy CNV Profiling vs Control Samples



Identification of Virus Integration Sites in Genome could potentially leads to neoplasia



Conclusions

We present here results from analysis on human and cancer genome, non-model and large complex genomes. We detected hundreds to thousands of large structural variants per genome and haplotype differences in variable regions of these genomes, revealing the locations, orientations and copy numbers of these complex structural variants often elusive from NGS detection. Using Bionano Irys® platform, we were able to find and characterize SV size, types and repeat regions in the complex cancer genome that are biologically and clinically relevant, especially able to precisely map viral (oncogene) component integration sites within human genome, believed to be linked to oncogenesis by disrupting host genome functions.

Multiple methods were shown to be employed, by *de novo* genome mapping process, or via direct alignment of the long raw molecules against digitally digested and "barcoded" reference genome for detecting chromosomal abnormalities such as translocations, arm breakpoints or other lesions. Ultra long molecules used in the Irys System could be used to bridge unknown long distance SV event, such as chromothripsis that often incorporating fragmented chromosomes arms pulverized during cancer transformation

Reference

1 Cao, H., et al. Rapid Detection of Structural Variation in a Human Genome using Nanochannel based Genome Mapping Technology. Giga Science (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
4 Das, S. K., et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Research (2010); 38: 8
5 Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.