Mapping the "Dark Matter" of Genome – Long Repeats and Complex SVs Leading To True Contiguity of *De Novo* Assembly with NanoChannel Technology

F Amini[1], A Pang[1], H Dai[1], A Hastie[1], E Lam[1], T Anantharaman[1], V Dergachev[1], W Andrews[1], S Chan[1], M Sagbini[1], K Haden[1], M Austin[1], R Sebra[2], A Bashir[2], Z Dzakula[1], H Cao[1]

[1]BioNano Genomics, San Diego, California, USA, [2]Mount Sinai Icahn Medical Institute, New York, NY USA

Abstract

In spite of advancement in high-throughput NGS in the past decade, a large 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 diseases, are found more prevalent than we previously thought. During assembly, they leave gaps and unknown structural information as the "dark matter" of the genome, challenging for short read NGS and conventional low resolution cytogenetic techniques. Rapid comprehensive genome mapping in nanochannel arrays represents a new single-molecule platform independent of yet complementary to DNA sequencing for accurate genome assembly and structural variation analysis. Extremely long molecules of hundreds of kilobases fluorescently labeled at sequence motifs and elongated in 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.

Here, we present results from analysis on human, cancer and other complex genomes, uncovering hundreds of large SVs and haplotype differences in these genomes. Without considering SVs that overlap with N-base gaps in hg19, 90% of these SVs are supported by orthogonal experimental methods or historical evidence in public databases. Using Bionao Irys® platform, repeat regions can be more accurately characterized and put into biological context and novel algorithms designed specifically to investigate long repeat arrays, we were able to find and characterize previously enigmatic clinically relevant repeat regions in the human or complex genome.

For the first time, population scale cross-sample genome comparison to identify comprehensive genome structural variation is feasible on a single platform due to the quick turnaround time. Lastly, we present the first comprehensive analysis of a diploid human genome from the CEPH Trio family, that combines long read NGS sequencing with genome maps, resulting in hybrid assembly dramatically improves upon the assembly contiguity typically observed in shotgun sequencing approaches, with N50s exceeding 30 Mb.

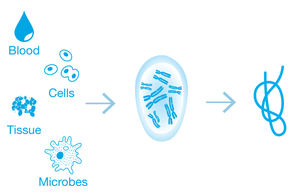
Background

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.

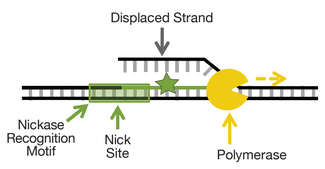
Methods

(1) Extremely long DNA is extracted from the source sample and (2) labeled with IrysPrep™ reagents by 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) Irys performs automated data collection and image processing. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable. (6) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite.

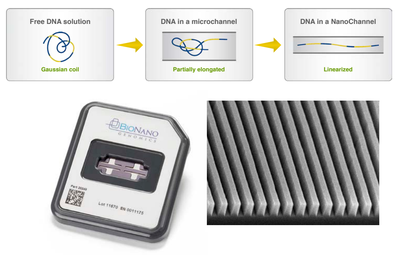
1) Isolate high molecular weight DNA



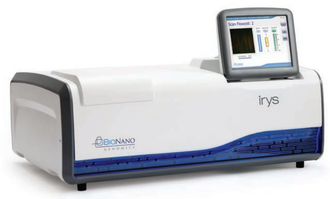
2) Label specific sequences across the entire genome



3) Transfer labeled DNA into cartridge for scanning

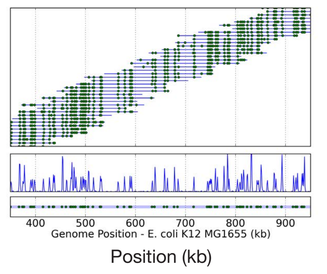


4) Load, linearize and image labeled DNA in repeated cycling

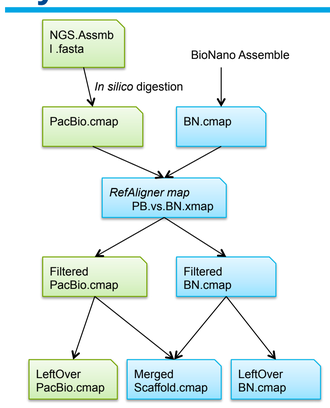


5) Algorithms convert images into molecules

6) Assembly algorithms align molecules for constructing consensus genome maps



Hybrid Scaffolding Flow Chart



1. Align NGS derived map with BioNano derived maps

2. Identify potential chimerical contigs and flag them. (For now, we don't use flagged contigs)
   2b. Alignment Visualization, QC, Manual Inspection

3. Run Merge Pipeline to generate "super-scaffolds".

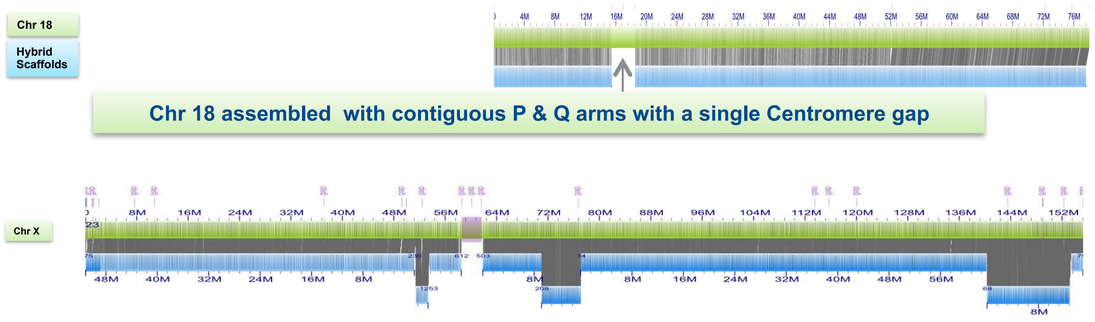
4. Identify PacBio – SuperContig Alignments and generate AGP output etc.

Co-Assembly: Combining NGS Contigs and Genome Maps

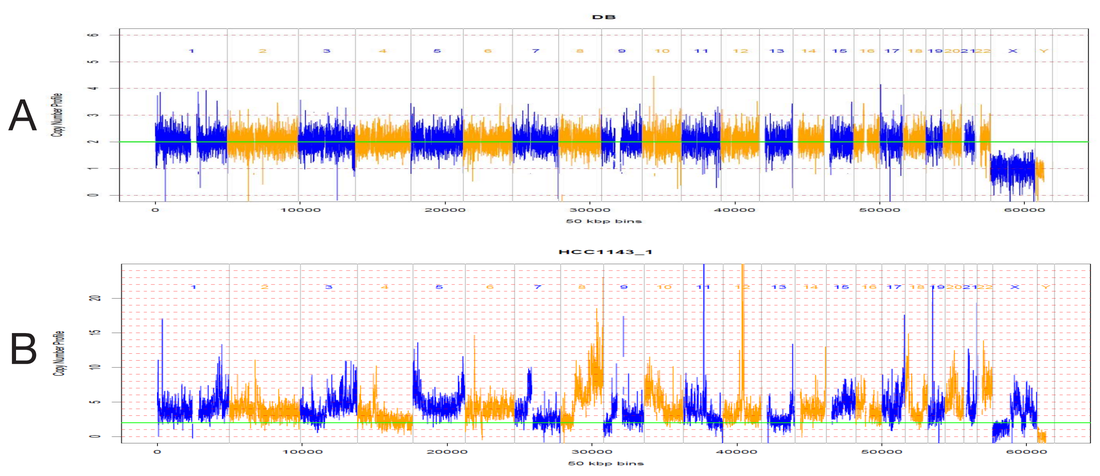
	# Contigs	N50
NGS Contigs	17,199	930 Kb
Selected NGS contigs	3,895	**1.1 Mb**
BioNano	1,003	4.6 Mb
Hybrid Scaffolds	182	**34.5 Mb**

We collaborated with NGS partners to establish a method for the first comprehensive analysis of a diploid human genome - NA12878 (daughter) of CEU Trio that combines single-molecule sequencing (PacBio RSII) with single molecule genome maps (BioNano Genomics' NanoChannel Arrays).

Using novel algorithms to integrate genome maps and NGS, one can generate genomes that rival, or surpass, the best human reference assemblies available today efficiently.



Copy Number Profiles in Cancer Samples



The motif pattern on long single molecules from BioNano's Irys System were aligned to digitally nick labeled reference genome 19 to build a histogram map illustrating the depth profiles. Graph A depicts a "normal" healthy male Caucasian genome profile while graph B shows the cancer genome full of amplified regions (peaks), deletions and break points, polyploidy and aneuploidy. The green line is the normalized diploid level.

IrysSolve Computational Solutions

IrysSolve data analysis pipeline performs automated *de novo* assembly and structural variation detection in an integrated pipeline.

Options include:
- IrysSolve for Intel® Xeon Phi™, which features a specially optimized version of IrysView's proven *de novo* assembly and structural variation detection pipeline, implemented on an open-source Linux platform, featuring full compatibility with IrysView.
- IrysSolve Cloud
- IrysSolve for your Existing Hardware, which allows you to integrate the IrysSolve data analysis pipeline on most pre-existing Linux compute hardware.

Conclusions

Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. Genome mapping in NanoChannel arrays is shown to be a rapid, accurate, powerful and robust method for detection of structural variation and the study of complex regions in the human genome.

References

1) 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.
2) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
3) 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:
4) Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
5) http://dgv.tcag.ca/dgv/app/home