Poster # 610

Resolving the “Dark Matter” of the Genome
Complex Structural Variations and True Contiguity of *de novo* Assembly with Next-Generation Mapping

H Cao[1], A Hastie[1], E Lam[1], A Pang[1], Mak A[3], W Andrews[1], T Anantharaman[1], T Chan[1], T Liang[1], M Saghibini[1], H Sadowski[1], ZY Zhu[1], M Austin[1], Z Dzakula[1], E Holmlin[1], Ali Bashir[2], P Kwok[3]

[1]BioNano Genomics, San Diego, CA, USA; [2]Mt Sinai School of Medicine, NY, NY, USA; [3]UCSF, San Francisco, CA, USA

Abstract

Large genomic structural variations (SV > 1 kbp) known to be associated with complex traits and diseases are found in more human genomes than ever before. Even though high-throughput, short read next-generation sequencing (NGS) has experienced rapid advancement and reduced cost this past decade, improvements in NGS’s genome assembly quality and integrity continue to lag behind. While hundreds of different new genome assembly and analysis tools exit and continue to be developed, a fair portion of the human genome architecture remains unresolved or ambiguously characterized. Fast, low-cost methods to reveal difficult heterogeneous structural information such as long range repeats, also known as the “dark matter” of the genome, is much needed.

BioNano Genomics Irys® next-generation mapping (NGM) solution represents a new standard of a single-molecule platform independent of, yet complementary to, DNA sequencing for accurate whole genome assembly and structural variation analysis. Long intact DNA molecules, hundreds kilobase pairs to multiple megabase pairs in length, are fluorescently labeled at sequence motifs and linearized in true nanoscale fluidic channels (~ 40 nm), and labeling patterns are imaged enabling the direct interrogation of the whole genome architecture at high resolution. The single-molecule based sequence motif labeling patterns are de novo assembled yielding multi-megabase long, contiguous consensus genome maps capable of spanning over highly repetitive regions and complex structures, such as large indels, segmental duplications, and scaffolding sequence contigs, synthetic linked reads, and large gaps. Using long and short-read NGS and BioNano NGM, we achieved reference-quality hybrid assembly of a diploid human genome resulting in a 34X improvement in scaffold N50 over long-read NGS alone, delivering what is considered a step towards the standard “medical grade” genome. We also present here results from using BioNano’s Irys System for the well-studied CEU trio from the 1000 Genomes Project, enabled efficient identification and validation of structural variants, including insertions, deletions, and inversions, greater than five kilobase pairs (kbp) in size. Using BioNano’s Irys System, 909 insertions and 661 deletions, including 800 novel insertions that were unidentified in the 1000 Genomes Project were detected for a remarkable seven-fold improvement in sensitivity in detecting structural variations compared to NGS. Based on our study, we were able to determine the novel insertions and deletions using maps assembled from the native long DNA molecules in only one experiment. The detection of numerous novel insertions and deletions in this study demonstrates that long-range genome analysis using the Irys System is a very powerful approach.

Whole genome next-generation mapping provides extremely valuable structural information hard or otherwise impossible to decipher with short-read sequencing data alone, and paves the road for generating true medical grade personalized genome information.

Methods

Cloneless Reference Quality Genome Assembly using Complementary NGS and BioNano Genome mapping

NATURE METHODS | ADVANCE ONLINE PUBLICATION

Assembly and diploid architecture of an individual human genome via single-molecule technologies

Assembly & Scaffold statistics	# Contigs	N50	Max contig/ scaffold length
Sequence assembly (PacBio)	22,433	906 Kb	6.5Mb
Genome Maps (BioNano)	1,039	4.6 Mb	26.6Mb
Hybrid Scaffolds (V2)	202	31.1 Mb	81.4 Mb

N 50 Hybrid-assembly results **34x** improvement over PacBio alone; **7x** better than BioNano alone.

Complementarity of Sequence Contigs and Genome Maps between PacBio and BioNano supports a stronger merged assembly.

BioNano’s NGM Finds Seven Times more Large SVs Than NGS

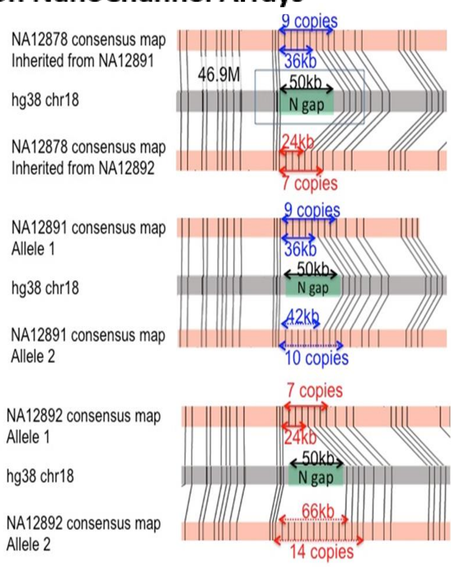
HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

Genome-Wide Structural Variation Detection by Genome Mapping on NanoChannel Arrays

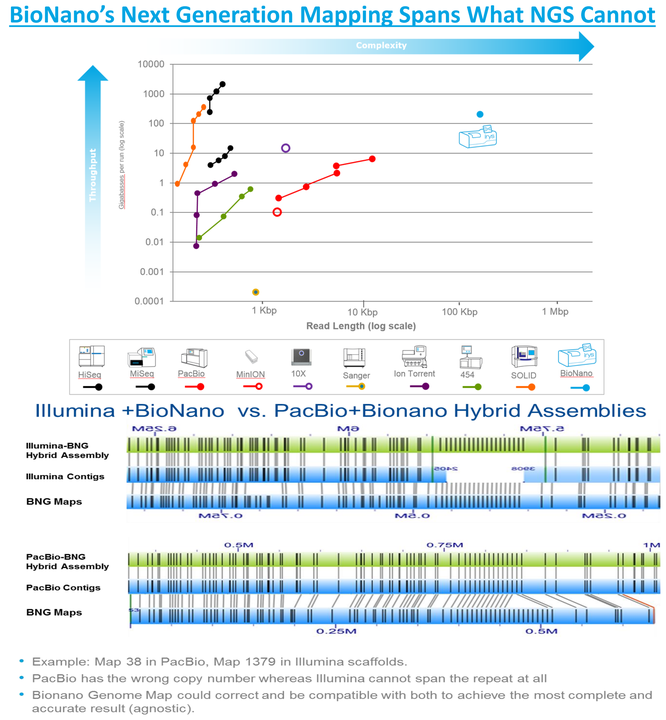
Number of validated insertions and deletions (>5 kb) detected by single-molecule maps and genome maps.

	Insertion	Deletion	Tot.
a. By samples:			
NA12878	769	522	
NA12891	743	496	
NA12892	748	456	
b. By novelty: based on 1000G pilot and phase 1 insertions and deletions >5 kb			
Known Supported (Total)	39 (59)	125 (156)	215
Novel	870	536	
c. By Mendelian inheritance			
Mendelian	879	631	
Non-Mendelian	4	4	
No call*	26	26	
d. Total	909	661	1570

*Unable to generate a call for Mendelian inheritance due to insufficient data to determine an insertion/deletion call to any of the samples.



BioNano’s Next Generation Mapping Spans What NGS Cannot



- Example: Map 38 in PacBio, Map 1379 in Illumina scaffolds.
- PacBio has the wrong copy number whereas Illumina cannot span the repeat at all
- Bionano Genome Map could correct and be compatible with both to achieve the most complete and accurate result (agnostic).

Conclusions

BioNano Genomics Irys System’s Next Generation Mapping (NGM) is a very powerful and versatile genome analysis tool useful for providing native, long-range information of complex genomic architecture at the single-molecule level. NGM can also be used for:

- Generating *de novo* whole genome assemblies used for evolution and comparative genomic analysis
- Genome Assembly and Hybrid Scaffolding
  - Enhancing and refining genome references and draft genomes
  - Orienting and aligning sequencing reads using a hybrid scaffold
  - Closing the contig gaps (Ex.: Reducing #contigs of an assembly from 2500 down to 300)
  - Combining NGM with sequencing reads for de novo assembly in species without a reference genome
- Long-range SV discovery (2 kbp to > 1 Mbp)
  - Imbalanced - In/Dels, CNVs
  - Balanced - Translocations, Inversions
  - Validating novel SVs in large population and patient cohorts

Reference

- Pendleton, M., Sebra, R., et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods (2015); e3454
- Mak AC et al., Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays Genetics (2015)
- Zook, J., et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. BioRxiv (2015)
- Cao, H., et al. Rapid Detection of Structural Variation in a Human Genome using Nanochannel-based Genome Mapping Technology. Giga Science (2014); 3(1):34
- Lam, E.T., et al. Genome mapping on Nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 30(8):7713) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
- Tegenfeldt, J.O., et al. The dynamics of genomic-length DNA molecules in 100-nm channels. PNAS (2004); 101:10979-83
- Cao, H., et al. Fabrication of 10nm Enclosed Nanofluidic Channels. Applied Physics Letters (2002); 81:174