bionano GENOMICS

Next-Generation Mapping: Application to Clinically Relevant Structural Variation Analysis

A Hastie, A WC Pang, J Lee, E T Lam, T Anantharaman, W Andrews, M Saghbini, H Cao

Bionano Genomics, San Diego, California, United States of America

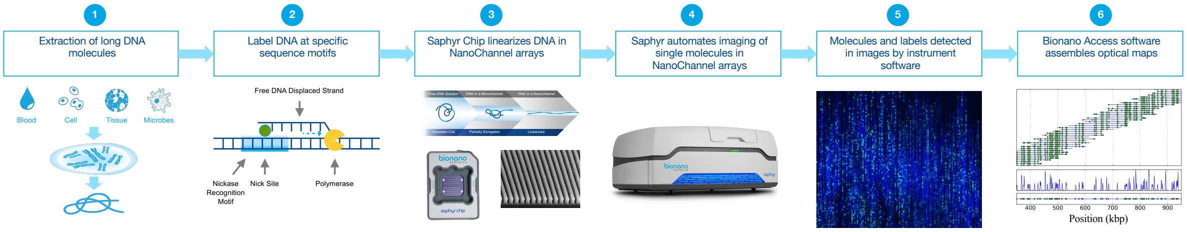
Abstract

Next-generation mapping (NGM) from Bionano Genomics® allows researchers to interrogate genomic structural variations (SVs) in the range of one kilobase pairs and above. It uses extremely long range information to span interspersed and long tandem repeats making it suitable for elucidating the structure and copy number of complex regions of the human genome, such as loci with complex pseudogene and paralogous gene families. Because NGM is a *de novo* process and because molecules analyzed are longer than almost all genomic repeats, NGM is able to detect a wide range of SVs including insertions of novel sequence, tandem duplications, interspersed duplications, deletions, inversions and translocations, a range of SV types detectable by NGM alone. Because of the high speed and comprehensiveness of the SV types detected, NGM is increasingly being applied to the analysis of clinical genomes for the detection of SVs potentially involved in disease pathogenesis. We present several *in silico* and biological validation experiments to demonstrate the sensitivity and specificity of NGM for insertion, deletion and translocation SVs and compare it to benchmark studies using short read and long read sequencing. We also show the application of NGM to studying somatic variation in a breast cancer cell line, finding hundreds of somatic structural variations. Finally, we applied NGM to several leukemia patient samples to find more than 50 cancer related SVs in each patient. NGM is a fast and cost effective method for detection of a broad range of traditionally refractory SVs across the 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 Saphyr™ 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. These long labeled molecules are *de novo* assembled into physical maps spanning the entire diploid genome. The resulting provides the ability to detect a large range of homozygous and heterozygous structural variation with very high efficiency.

Methods


1 Extraction of long DNA molecules

2 Label DNA at specific sequence motifs

3 Saphyr Chip linearizes DNA in NanoChannel arrays

4 Saphyr automates imaging of single molecules in NanoChannel arrays

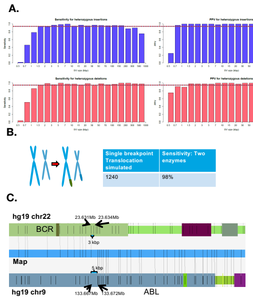
5 Molecules and labels detected in images by instrument software

6 Bionano Access software assembles optical maps

(1) Long molecules of DNA are labeled with Bionano 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 Saphyr Chip using NanoChannel arrays (4) Single molecule are imaged and then digitized by the Saphyr instrument. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Bionano maps may be used in a variety of downstream analysis using Bionano Access software.

Next Generation Mapping Detects Structural Variation With High Sensitivity and Specificity

High Accuracy and Sensitivity in In Silico Simulation Data



Single breakpoint Translocation simulated	Sensitivity: Two enzymes
1240	98%

A. Simulations of human molecules with homozygous and heterozygous insertions and deletions were produced, assembled and SV calling was performed.

- Homozygous SV sensitivity is >99%.
- Heterozygous SV sensitivity is >95% in most size bins.
- Additional large insertions (>200 kbp) are found but classified as “end” calls.
- PPV averages >95% overall.

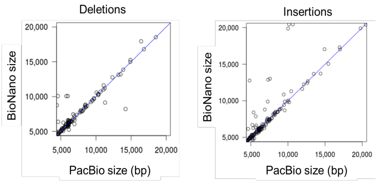
B. Simulations of human molecules with heterozygous translocations were produced, assembled and SV calling was performed. Sensitivity for single breakpoint translocaiton was determined to be 98%.

C. Example of a Philadelphia translocation t(9;22) was detected in leukemia cancer cells. The map (blue) was aligned to the public reference assembly hg19, and the resulting alignments show a conjoined junction, creating a fusion gene *BCR-ABL*.

	PACBIO				BIONANO			
	CHM1 and CHM13 assemblies	Mixture assembly	Sensitivity	PPV	CHM1 and CHM13 assemblies	Mixture assembly	Sensitivity	PPV
Homozygous Insertions	467	353	75.6%	96.1%	707	700	99.0%	97.9%
Heterozygous Insertions	586	252	43.0%		663	554	83.6%	
Homozygous Deletions	221	183	82.8%	94.9%	269	268	99.6%	97.1%
Heterozygous Deletions	501	337	67.3%		517	477	92.3%	

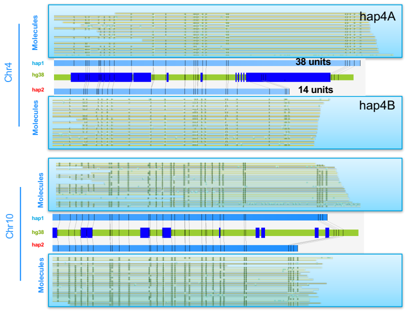
Two Homozygous cell lines, CHM1 and CHM13 were independently *de novo* assembled and SVs called. Raw data was mixed together, assembled and SVs called (mixture column). In the simulated diploid assembly, CHM1 only and CHM13 only SVs are heterozygous and those detected in both are homozygous SVs. Results are compared to a similar experiment using PacBio data for SV calling (Huddleston et al., 2016)

Next Generation Mapping Can be Cross Validated by Orthogonal Technologies



SV size comparison between overlapping Bionano and PacBio SV calls shows good size concordance.

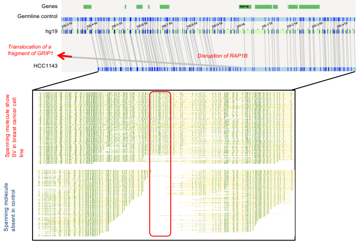
Tandem Repeat Array Length Measurement Haplotype and Paralog Resolution



D4Z4 repeat arrays occur at the subtelomeric region of chromosome 4q and 10q. 4q has two haplotypes, 4qA and 4qB, if the array length of 4qA falls below ~5 copies, Facioscapulohumeral (FSHD) muscular dystrophy can occur. In order to diagnose a pathogenic repeat array, the array for each allele or paralog must be measured and differentiated. The figure shows the differentiation and measurement of each allele on chromosome 4 and paralogs on chromosome 10 (this is a non pathologic measurement) in genome maps (blue bars) and single molecules shown below.

Detection of Somatic Variation in a Breast Cancer Cell Line Compared to a Lymphoblast Cell Line (LCL)

Somatic SVs	HCC1143 Tumor vs LCL	NA12878 LCL vs LCL
Insertion	14	0
Deletion	52	0
Inversion breakpoint	26	0
Translocation	34	0
Total	126	0



(Top) The table shows the number of somatic SVs detected in a breast tumor cell line sample (HCC1143) and absent in the matched blood cell line sample. SVs detected in HCC1143 but not detected in HCC1143BL were retested through molecule alignments to the SV with tumor and normal molecules to confirm somatic nature. As a negative control, the NA12878 sample was run twice, and there were no SV uniquely detected. (Bottom) A translocation is shown that disrupts RAP1B, a member of the RAS Oncogene Family in HCC1143. The spanning molecule pileups show strong molecule support for the translocation in the tumor but no evidence in the control cell line.

Conclusions

We demonstrate that the Saphyr system can be used to accurately detect genetic mutation hallmarks in samples with hematologic and epithelial malignancies. We were able to find known calls from cytogenetic experiments, and also detected novel aberrations. Especially useful for rare disease studies, researchers using this system can uncover rare and *de novo* variants by comparing with our control sample database and the unaffected parents, respectively. Our results shown here indicate that the Saphyr system can identify a broad spectrum of SVs of functional importance, providing researchers a new fast and cost-effective approach to deciphering relevant genomic mutations otherwise missed by other technologies. In summary, Bionano optical mapping provides the highest sensitivity for large (>1.5kb) homozygous and heterozygous SVs, including deletions, duplications, insertions, inversions, translocations, tandem repeat expansions/contractions and other SVs.

Reference

1) Cao, H., et al., Rapid detection of structural variation in a human genome using NanoChannel-based genome mapping technology. Gigascience (2014); 3(1):34
2) Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
3) Xiao, M et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
4) Huddleston J, C. M.-L. (2016, Nov 28). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res, gr.214007.116.