Bionano Genomics' Next-Generation Mapping Identifies Large Structural Variants in Plant and Animal Genomes

All types of large structural events are detected as heterozygous or homozygous variants with unrivaled sensitivity and specificity.

With a rapidly growing population, the demands on global breeding programs expand. Tools to screen for genetic regions with high biological potential become increasingly important.

Current tools are designed to associate single-nucleotide polymorphisms (SNPs) with specific phenotypes. **Recent studies show that the heritability of many complex phenotypes is the result of large structural variation (SVs) rather than SNPs**[1]. For instance, it was recently demonstrated that a nematode resistance phenotype in soybean is the result of a large tandem repeat containing a gene block at the Rhg1 locus[2]. Also, structural variation regulates the expression of genes controlling flavor compounds in wild Brazilian strawberries, which might hold the secret to revitalizing commercial varieties[3].

A robust tool to analyze genomes for structural variants is key to unlocking new sources of crop-enhancing traits.

Eukaryotic genomes contain large quantities of repetitive DNA sequences and Whole Genome Sequencing (WGS) does not always align correctly with the repetitive parts of the genome. In plants and animals, species differ in genome size by as much as several orders of magnitude, even between closely related species[4]. Among repetitive sequences, transposable elements are the most responsible for these pronounced differences, reaching up to 85% of some species of large genomes such as that of maize[5]. The short-read sequences provided by Next-Generation Sequencing (NGS) map with poor accuracy to these repeats. Alignment algorithms typically fail to identify the exact genomic location to align these short-reads to. When they do align, the limited 100-150 bp read length and spacing of paired-end reads does not allow for a correct sizing of larger repeats.

Structural variants make up the majority of genomic variation, but NGS can't correctly identify them. NGS reliably identifies single nucleotide variants and small insertions and deletions. However, NGS has limited power to identify most large insertions, deletions, or copy-number variations. Short reads are forced to map an incorrect or too divergent reference, and mismatched reads are often excluded from the alignment. Various NGS based SV calling algorithms routinely disagree and have limited power to detect other structural variations (SVs) such as inversions and translocations.

Bionano Next-Generation Mapping (NGM) is the only technology that can show you all SV types, homozygous and heterozygous, starting at 1000 bp up to millions of bp. Megabase size molecules of genomic DNA are labeled, linearized and uniformly stretched in high density NanoChannel arrays, and imaged on the Bionano system. Using a nicking endonuclease, a specific 6 or 7 basepair sequence is labeled approximately 10 times per 100 kbp. The label patterns allow each long molecule to be uniquely identified, and aligned. Using pairwise alignment of the single molecules, consensus genome maps are constructed, refined, extended and merged. Molecules are then clustered into two alleles, and a diploid assembly is created to allow for heterozygous SV detection. Genome maps can be created using different endonucleases to generate broader coverage and higher label density.

Bionano's SVs are observed, not inferred as with NGS. NGS algorithms piece together sequence fragments in an attempt to rebuild the actual structure of the genome. SVs are **inferred** from the fragmented data, with mixed success. With NGM, megabase-size native DNA molecules are imaged, and most large SVs or their breakpoints can be **observed** directly in the label pattern on the molecules. A native-state DNA molecule with a specific SV is direct proof of the variant's existence, unlike an algorithmically determined variant reconstruction.

© Copyright 2017, Bionano Genomics Inc. M 858.888.7600 F 858.888.7601 | info@bionanogenomics.com | bionanogenomics.com
For research use only. Not for use in diagnostic procedures.

Bionano Next-Generation Mapping has successfully identified large structural variants of biological significance in plant and animal genomes

Bionano maps tandem gene repeats in maize: The *P1* gene in maize is of particular interest to researchers. It is involved in parasite resistance, and in the synthesis of red flavonoid color – and thus influences kernel color. The *P1-wr* allele is a duplicated gene cluster, and the maize reference genome contains a gap spanning this hard to assemble tandem repeat (Figure 1). Characterizing the structure and repeat copy number of this repeat in individual lines would be tedious and labor intensive using subcloning and sequencing of each unit. NGM however easily detects copy number variation in each line. The *P1-wr* locus is made up of 11 copies of 12.7 kbp unit in reference B73 line, while only 10 copies are detected in W22 (Figure 1).

Bionano characterizes large repeat arrays in crow: Matthias Weissensteiner at Uppsala University in Sweden used a combination of short-read, long-read sequencing and NGM to identify 36 previously unidentified large repetitive regions anchored to assembly breakpoints. The majority of these anchored repeats contain complex arrays of a 14 kbp satellite repeat or its 1.2 kbp subunit, and were largely missed by sequencing efforts (Figure 2).

Anchoring these large heterochromatic regions using NGM allows him to localize chromosomal features such as centromeres in the assembly, something that is not trivial in non-model organisms. Since these repeats are typically associated with centromeric and telomeric regions they likely influence meiotic recombination, and provide a possible novel explanation for areas of increased genetic differentiation.

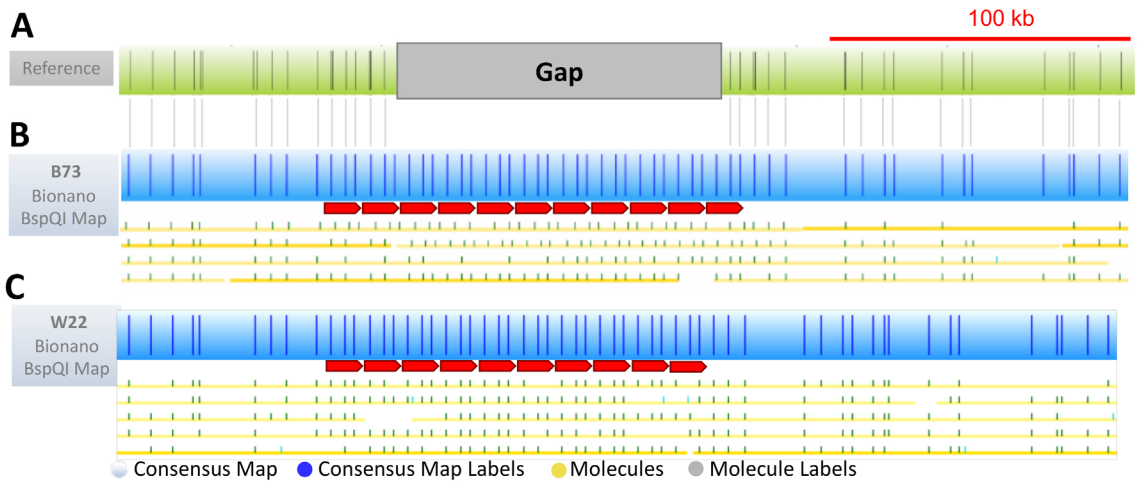

Figure 1:
Bionano genome maps of the P1-wr allele in two individual maize lines. Vertical lines represent Nt.BspQI enzyme recognition sites. The reference shows a gap, while NGM identifies 11 tandem gene copies (red arrows) in B73 and 10 copies in W22.

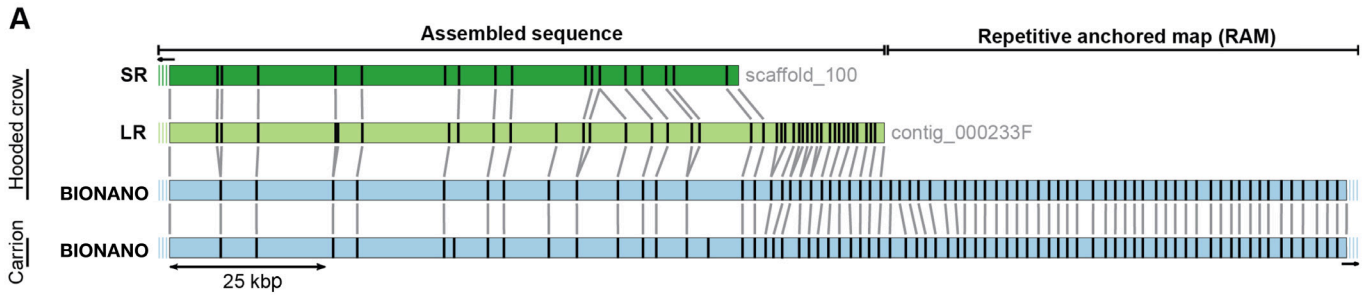

Figure 2:
Alignment of short-read assembly (SR), long-read assembly (LR) and Bionano genome maps (Bionano) for the hooded crow and carrion. The large anchored repeats are absent from the short-read assembly, partially found in the long-read assembly, and fully mapped by NGM. Vertical lines represent Nt.BspQI label sites.

© Copyright 2017, Bionano Genomics Inc. M 858.888.7600 F 858.888.7601 | info@bionanogenomics.com | bionanogenomics.com
For research use only. Not for use in diagnostic procedures.

Bionano Next-Generation Mapping detects structural variants with a sensitivity and specificity far greater than NGS

NGM detects seven times more SVs larger than 5 kbp compared to NGS. Professor Pui Kwok at the University of California, San Francisco, demonstrated the robustness of NGM for genome-wide discovery of SVs in a trio from the 1000 Genomes Project[7]. Since high quality NGS data on these samples is publicly available, structural variation analysis using short-read data has been performed with over a dozen different algorithms. Using Bionano maps, hundreds of insertions, deletions, and inversions greater than 5 kbp were uncovered, 7 times more than the large SV events previously detected by NGS (Figure 3). Several are located in regions likely leading to disruption of gene function or regulation.

NGM has exceptional sensitivity and specificity to detect insertions and deletions over a wide size range as demonstrated using simulated data. Insertions and deletions were randomly introduced into an in silico map of the human reference genome hg19. The simulated events were at least 500 kbp from each other or N-base gaps. They ranged from 200 bp to 1 Mbp, with smaller SVs more frequent than larger ones.

Based on the edited and the unedited hg19, molecules were simulated to resemble actual molecules collected on a Bionano system and mixed such that all events would be heterozygous. Two sets of molecules were simulated, each labeled with a different nicking endonuclease. Datasets with 70x effective coverage were generated. The simulated molecules were used as input to the Bionano Solve pipeline and SV calls were made by combining the data from both nicking endonucleases using the SV Merge algorithm. SV calls were compared to the ground truth.

Figure 4 shows sensitivity and positive predicted value (PPV) for heterozygous insertions and deletions within a large size range. SV size estimates were typically within 500 bp of the actual SV sizes, while reported breakpoints were typically within 10 kbp of the actual breakpoint coordinates. Additional large insertions (>200 kbp) were found but classified as end-calls.

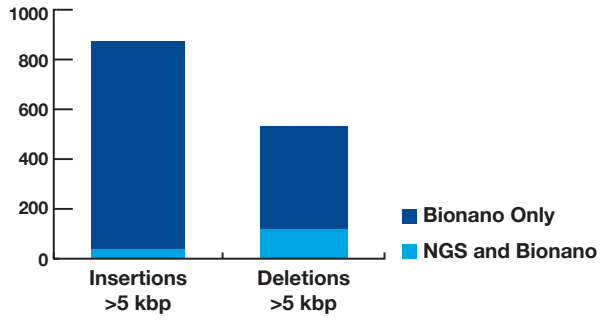

Figure 3:
NGM was used to detect insertions and deletions >5 kbp in the CEPH trio. The total number of large insertions and deletions detected in the three individuals is shown. The number of SVs found using Bionano was 7x larger than those already detected from NGS data by over a dozen algorithms.

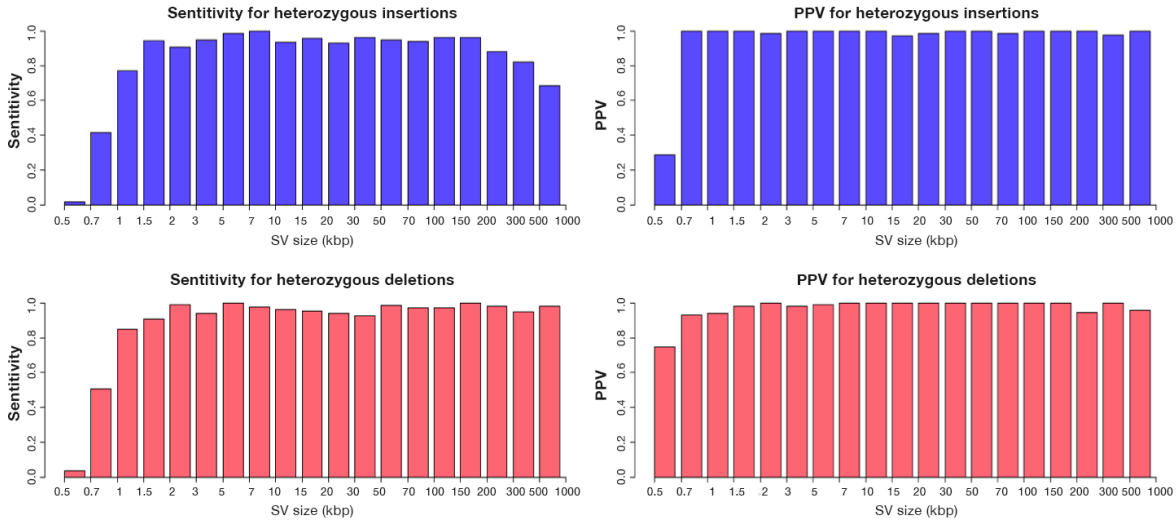

Figure 4:
Heterozygous SV calling performance from a simulated dataset. Molecules were simulated from unedited and edited versions of hg19 (with insertions and deletions of different sizes) and used for assembly and SV calling.

© Copyright 2017, Bionano Genomics Inc.

NGM has exceptional sensitivity and specificity to detect heterozygous insertions and deletions over a wide size range as demonstrated using experimental data. Since there is no perfectly characterized human genome that can be considered the ground truth, a diploid human genome was simulated by combining data from two hydatidiform mole derived cell lines. These moles occur when an oocyte without nuclear DNA gets fertilized by a sperm. The haploid genome in the sperm gets duplicated, and the cell lines resulting from this tissue (CHM1 and CHM13) are therefore entirely homozygous.

Structural variants detected in the homozygous cell lines were considered the (conditional) ground truth. An equal mixture of single molecule data from two such cell lines was assembled to simulate a diploid genome, and SV calls made from this mixture were used to calculate the sensitivity to detect heterozygous SVs.

Table 1 shows the number of insertions and deletions larger than 1.5 kbp detected in the CHM1 and CHM13 homozygous cell lines relative to the reference, and the in silico CHM1/13 mixture. SVs detected in CHM1 only or CHM13 only are heterozygous and those detected in both are homozygous. NGM has a sensitivity of 92% for heterozygous deletions and 84% for heterozygous insertions larger than 1.5 kbp. The largest detected deletion was 4.28 Mbp in size and the largest insertion 412 kbp.

A similar experiment on PacBio long-read sequencing was described recently[8]. Structural variants were called with the SMRT-SV algorithm in CHM1 and CHM13, and compared to those called in an equal mixture of both. The sensitivity to detect homozygous SVs using PacBio was 87%, compared to 99.2% using Bionano. The sensitivity to detect heterozygous SVs using PacBio was only 41%, which is less than half the 86% sensitivity for heterozygous SV detection using Bionano. Even when the PacBio SV calls were limited to insertions and deletions larger than 1.5 kbp, the sensitivity for homozygous SVs was only 78%, and for heterozygous SVs 54% (Table 1).

Conclusion

Bionano Next-Generation Mapping Identifies Large Structural Variants in Plant and Animal Genomes. All types of large structural events are detected as heterozygous or homozygous variants with unrivaled sensitivity and specificity.

Learn More

You can download detailed technical information about the Irys System and SV calling at the Irys Technology page on the Bionano Genomics website: http://www.bionanogenomics.com/technology/irys-technology/

	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%	

Table 1:
Two homozygous cell lines, CHM1 and CHM13 were independently de novo assembled and insertions and deletions >1.5 kbp called. Raw data was mixed together, assembled and SVs called (Mixture assemblies column). The sensitivity and specificity (PPV) to detect heterozygous relative to homozygous SVs is shown.

References: 1.Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 11:446–450. 2. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, et al. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 2012;338(6111):1206–1209. 3. Chambers et al.: Identification of a strawberry flavor gene candidate using an integrated genetic-genomic-analytical chemistry approach. BMC Genomics 2014 15:217. 4. Gregory TR (ed): The Evolution of the Genome, pp 89-162 (Elsevier, Burlington 2005). 5. Schnable et al., 2009. 6 Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW. in revision. Combination of short-read, long-read and optical mapping assemblies reveals heterochromatic tandem repeat arrays with population genetic implications. Genome Res. 7. Mak AC, 2016. 8. https://www.ncbi.nlm.nih.gov/pubmed/27895111 Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin CS, Korlach J, Wilson RK, Eichler EE. Genome Res. 2016 Nov 28. pii: gr.214007.116.

For general information about the Irys® System, please contact info@bionanogenomics.com or visit bionanogenomics.com

Bionano Genomics®, Irys®, IrysView®, IrysChip®, IrysPrep® and IrysSolve® are trademarks of Bionano Genomics Inc. All other trademarks are the sole property of their respective owners.