Genome-Wide, Highly Sensitive and Accurate Structural Variation Detection in Plants and Animals by Next-Generation Mapping

A. Hastie, S. Chan, J. Lee, T. Laing, L.X. Zhang, E.T. Lam, T. Anantharaman, M. Saghbini, H. Cao

BioNano Genomics, Inc., San Diego, CA, USA

Abstract

Establishing reference genomes for large, and complex genomes such as maize, rice, barley, wheat and other grass plants is now made possible by the combination of single molecule methodologies from Bionano Genomics next-generation mapping (NGM) and Pacific Biosciences sequencing as well as other long-range enabling technologies. However, to fully reveal the level of structural diversity, one needs to be able to compare multiple cultivars and organism at the whole genome level. This is being accomplished by NGM where genome wide structural variation in the form of insertions and deletions >1 kbp and inversions as well as large scale rearrangements can be identified in a highly efficient manner.

We report on the structural relationship between different maize and barley cultivars, where we identify thousands of insertions and deletions and hundreds of large rearrangements. We show that these rearrangements affect genic regions and predict rearrangements that effect the expression of genes by altering protein coding regions. Taken together, NGM is not only a highly efficient method to compare the structure of plant as well as animal genomes (SV analysis) but it is the most comprehensive as a result of the longest read lengths by available technologies that allow access into the "dark matter" of the genome.

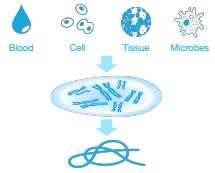
Background

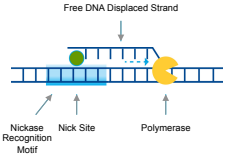
Irys is based on the next-generation mapping technique in which labeled DNA is linearized and imaged in massively parallel NanoChannel arrays. Each imaged DNA molecule is a map of a portion of the sample genome, and these individual reads are *de novo* assembled into highly accurate consensus genome maps that cover the entire genome.

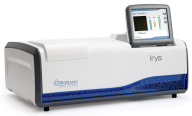
A major advantage of the Irys System is that it provides direct visualization of intact DNA molecules up to 2 Mbp in their native state, reliably preserving the long range genomic information that is usually lost in other, less direct approaches. Since there are no amplification steps, errors and artifacts are extremely rare. For these reasons, NGM using the Irys System is an ideal candidate for projects involving structural variation analysis, and diploid (or polyploid) assembly with proper phasing. Genome maps can also be used to scaffold and validate NGS contigs, resulting in highly contiguous and correct genome assemblies.

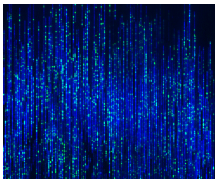
Methods

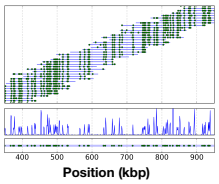
1. Extraction of long DNA molecules using IrysPrep kit
2. Fluorescent labeling of DNA at specific sequence motifs, with IrysPrep reagents
3. Linearization of DNA in NanoChannel arrays on IrysChip
4. Automated imaging of single molecules in NanoChannel arrays
5. Automated detection of molecules and labels in images, by instrument software
6. Assembly of data into optical consensus maps by IrysView software



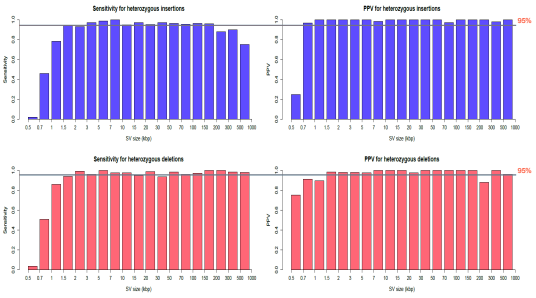




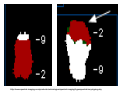




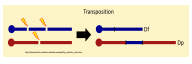
In Silico SV Detection Sensitivity and PPV



Translocations Simulated	Sensitivity for Translocation Detection
1240	98%

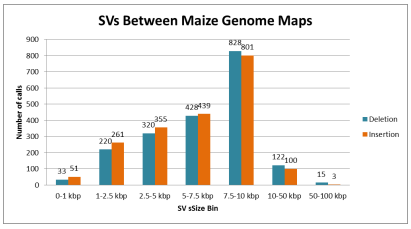


Transposition type rearrangements (number)	50-100 kbp sensitivity	100-300 kbp sensitivity	>300 kbp sensitivity
1836	39%	93%	98%



Human genomes were assembled from simulated molecules containing heterozygous insertions, deletions and translocations (no homozygous SVs). After assembly, genome maps were aligned to hg19 to find variation. Sensitivity to simulated insertions and deletions averages 95% from 1.5 kbp and up and PPV above 95% in almost all size bins (top bar graphs). Sensitivity for translocations and transpositions is 98%, only reducing when shorter transpositions are considered, these transpositions are still identified as insertion/deletion events (lower tables).

Genome Wide SV Analysis Between Maize Cultivars B73 and W22

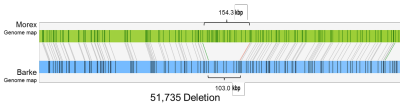


Deletion	1966
Insertion	2010
End	1749
Inversions	55
Rearangements	5

Two different maize genomes were analyzed for structural variation between them, in total 3976 insertions and deletions were detected with the most frequent size being 7.5-10 kbp. In addition, 55 inversions and 5 rearrangements were found as well as 1749 "end" sv's, which are genomic segments that diverge between genomes but cannot be classified as a more specific type of SV, generally they are large insertions.

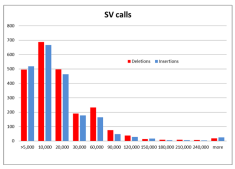
Large Inversion Discovered Between Barley Cultivars

Structural Variation Analysis

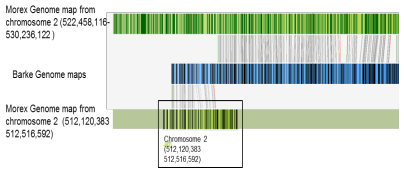


51,735 Deletion

Stats	
Aligned length	3.204 Gbp
Insertions	2121
Deletions	2272
Inversions	34



Above, overview of SVs called between two barley cultivars.



Lower panel shows a large, 20 Mbp, inversion on chromosome 2 detected between cultivars.

Conclusions

De novo detection of insertions, inversions and translocations is inefficient and inconsistent using commercially available technologies. Next-generation mapping is a modern approach that provides an efficient, cost effective, sensitive and specific solution for structural variation detection. Reference free structural variation detection is a valuable method for genomic comparison in plant and animal research where there is a great need for reference genomes.

Related posters: P0033, P0712, P0957, P0958

References

- Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
- 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

Acknowledgments

- Nils Stein & Martin Mascher: IPK
- Doreen Ware: USDA-ARS, CSHL