Accessing the Genomic Dark Matter: Using Next-Generation Mapping to Identify Large-Scale Repeat Structures in Maize

Jinghua Shi[1], Xiang Zhou[1] , R. Kelly Dawe[2], Ahmed Naguib[1], Zhanyang Zhu[1], Zeljko Dzakula[1], Weiping Wang[1], Saki Chan[1], Alex Hastie[1], Han Cao[#1]
[1] BioNano Genomics, Inc., San Diego, CA
[2] Department of Plant Biology, Athens, University of Georgia, GA

Abstract

Complex genome assembly has improved dramatically with advent of new long-range sequencing and assembly technologies. Among these methods, the next-generation mapping (NGM) with BioNano Irys® system offers distinct advantages by generating optical maps using NanoChannel arrays technology, particularly when combined with long range sequencing approaches such as PacBio.

The recently released maize B73 reference genome was assembled using a hybrid scaffolding approach that combined PacBio sequence and BioNano maps. While this is an excellent assembly with long contig N50, there are areas where the BioNano map is continuous but matching sequence data could not be identified. We analyzed BioNano maps corresponding to the largest of these sequence gaps (>100 kbp) and found that many are associated with either large duplicated regions or short tandem repeat arrays. These include the 45S and 5S rDNA arrays and several centromeric repeat arrays. In addition, we show that the P1 locus, which has 11 nearly-perfect copies in the B73 reference genome, has 10 copies in another maize inbred W22, demonstrating the utility of using NGM to assess copy number variations (CNV).

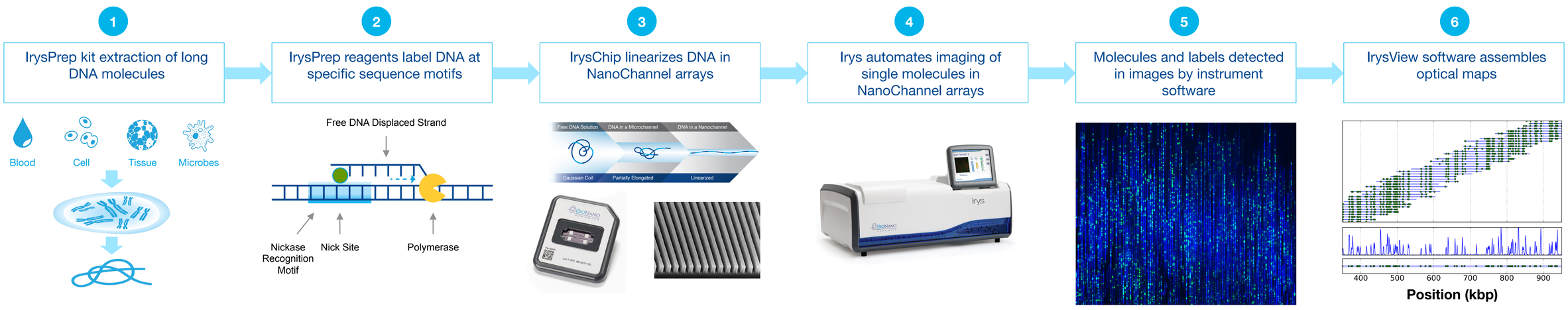
The data also reveal multiple previously unknown large-scale CNVs, many of which likely include genes. These results point to limitation of current sequencing methodologies and highlight the value of using NGM generated ultra-long 150 kbp -2 Mbp molecules for the accurate assembly of the maize genome, a process that is both cost effective and fast.

Background

The maize genome is an excellent example of a large complex genome that is replete with transposable elements as well as large numbers of simple and complex tandem repeats, such as “knob” repeats, centromeric repeats and rDNA repeats. Maize also contains lineage-specific repeated regions referred to as copy number variation. Studies show strong associations of CNV with important phenotypes such as biotic stress and aluminum tolerance.

The maize B73 version 4 assembly took a hybrid scaffolding approach that integrated PacBio and BioNano technologies. The BioNano assembly, with a longer N50, contains regions that are not identified in PacBio assembly. In these cases the hybrid scaffold frequently contains large sequence gaps separating PacBio contigs. Here we utilized the motif distribution patterns in the BioNano maps to interpret whether the regions of poor PacBio assembly contain CNVs. The data show that over 50% of the longest gaps contain either CNVs that are either large segmental duplications with low copy numbers, or short tandem repeats with high copy numbers. We also provide new BioNano data for a second inbred (W22) to demonstrate that large scale CNV can be readily assayed by BioNano data alone.

Methods



(1) Long molecules of DNA are labeled with IrysPrep® 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 IrysChip® using NanoChannel arrays and single molecules are imaged by Irys. (4) Single molecule data are collected and detected automatically. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Maps may be used in a variety of downstream analysis using IrysView® software.

Results

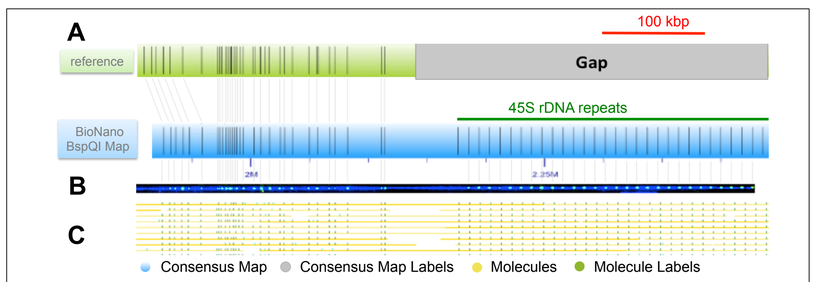


Figure 1. BioNano map and hybrid scaffold of the 45S rDNA region. **A.** alignment between reference in silico sequence motif map and BioNano Map. Grey area represents the N gaps in the reference. **B.** Raw image of a singe molecule captured in a nanochannel. **C.** alignment of digitized single molecules.

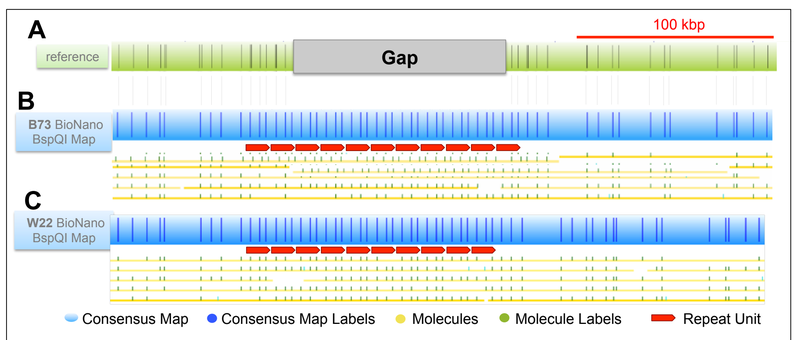


Figure 3. BioNano map readily detects copy number variation in different inbreds. P1 locus is made up of 11 X 12.7 kbp units in the reference B73, while there were only 10 copies in W22. **A. reference *in silico* sequence motif map**, **B, C.** assembled BioNano map containing P1 region in different inbred lines. **B.** B73, **C.** W22.

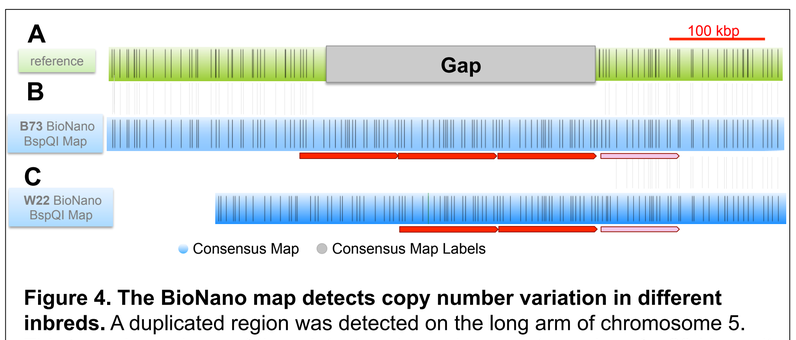


Figure 4. The BioNano map detects copy number variation in different inbreds. A duplicated region was detected on the long arm of chromosome 5. This locus is made up of three intact and one degenerate copies of a 99-kbp unit in B73, while there were two intact and one degenerate copies in W22. **A. reference *in silico* sequence motif map**, **B, C.** assembled BioNano maps containing the region that differed between inbred lines. **B.** B73, **C.** W22. Red arrows, intact duplicated copies; pink arrows, degenerated duplicated copies.

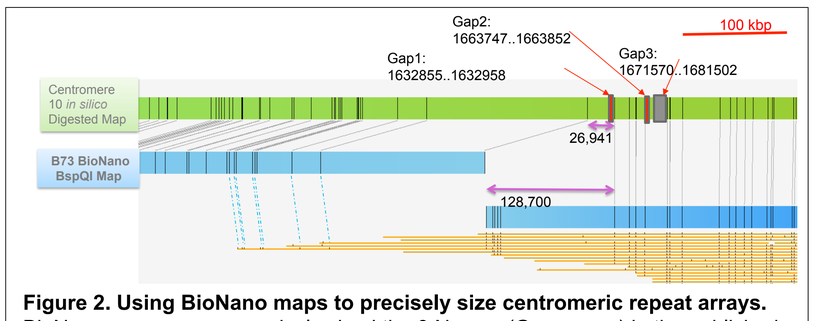


Figure 2. Using BioNano maps to precisely size centromeric repeat arrays. BioNano genome map precisely sized the 3 N gaps (Grey areas) in the published B73 centromere 10 full sequence. Gap 1 is shown to be 100 kbp instead of 100 bp.

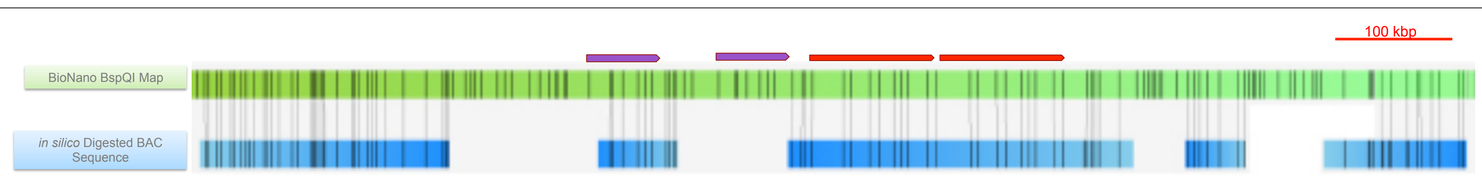


Figure 5. Use of BioNano maps to identify large segmental duplications. The region shown in this image is found on maize abnormal chromosome 10. It contains a cluster of kinesin genes that are responsible for meiotic drive in maize strains. Since the region is absent in the reference strains, BioNano genome mapping was used to orient PacBio-sequenced BACs and interpret copy number. Red arrows, duplicated region 1. Purple arrows, duplicated region 2. Note that one copy is not covered by selected BAC clones.

Conclusions

The BioNano Genomics Irys System provides a platform for mapping and visualizing extremely long molecules, which makes it possible to analyze long tandem repeat arrays comprehensively, accurately and without bias due to selection or amplification. Accompanied by our repeat detection tool (X. Zhou et al. PAG XXIV poster), repetitive units of the complex genomes can be profiled, analyzed and quantified in an unprecedented fashion. Our results show that tandem repeat arrays can be reliably detected as regularly spaced labeling patterns (as is the case for the P1 locus), or as long no label regions (as is the case for centromere 10) provided there are diagnostic patterns on either side of the array.

See also Posters: P0957, P0961, P0033, and P0958.

References

Wolfgruber, T. K., M. M. Nakashima, K. L. Schneider, A. Sharma, Z. Xie et al., 2016 Frontiers in Plant Science 7.
Springer, N. M., K. Ying, Y. Fu, T. Ji, C. T. Yeh et al., 2009 PLoS Genet 5: e1000734.
Cocciolone, S. M., D. Nettleton, M. E. Snook and T. Peterson, 2005 Plant Biotechnol J 3: 225-235.
Jiao, Y, P. Peluso, J. Shi, T. Liang, et al. in revision.
X. Zhou, S. Chan, J. Shi, et al. PAG XXIV poster.