Tandem Repeat Detection Using Next-Generation Mapping and its Application in Detecting rDNA Copy Number Variation in Plant and Human Genomes

X Zhou[1], S Chan[1], J Shi[1], W Wang[1], Z Zhu[1], A Hastie[1], Ž Džakula[1], H Cao[1]

[1]BioNano Genomics, San Diego, CA, USA

Abstract

Eukaryotic genomes harbor large tandem repeats, such as rDNA repeats. Copy number variation of rDNA has been shown to correlate with global changes in gene expression. These repeat elements, as special genetic markers, are essential for genomic functions, disease diagnosis, and mapping studies. However, rDNA regions normally span several megabase, which is far beyond the accessibility of current NGS technologies.

The BioNano Genomics Irys® System images fluorescently labeled long DNA molecules of 150 kb to 1 Mb. The exceptional integrity of DNA provides us a method to accurately determine repeat copy numbers and identify boundaries for alignment to physical maps.

We present a robust tool for the detection of both simple and compound tandem repeats with False Positive (FP) and False Negative (FN) support in complex genomes. In *Arabidopsis* and maize, 43S rDNA unit was detected as a 9-10 kb simple repeat, while in tomato genome, rDNA unit was detected as a two-label compound repeat. We further analyzed rDNA copy number and unit size variances between wild and cultivated tomato samples. Human 45S rDNA was detected as a 5-label compound repeat. Molecules containing these repeats play vital roles in precisely locating the rDNA regions on the physical map.

Based on the genome annotations, we validated 83% and 68% of the compound repeats detected in *Arabidopsis* and maize at sequence level, respectively. We concluded that the tool was fast (8 gigabases per minute), accurate, and robust using the default parameters, single threaded.

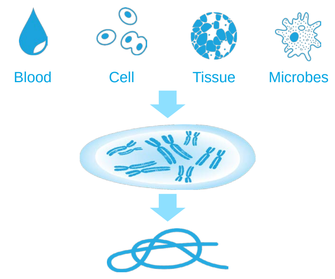
Genes under selection for high levels of transcript often occur as tandemly repeated clusters. The best known examples are the ribosomal DNA repeats, which are present in thousands of copies as tandem repeats and form Nucleolus Organizer Region (NOR).

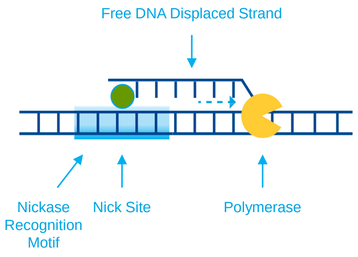
Many copies are needed because these genes make the RNA structural components of ribosomes and there is no further translation step. One gene could never make enough rRNA for all the ribosomes required for protein translation.

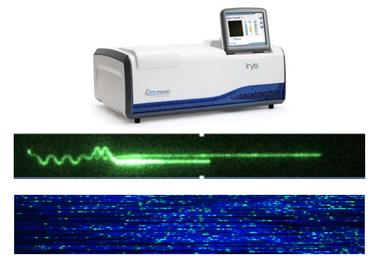
Due to the highly repetitive feature, it has been extremely challenging to accurately estimate copy number of NOR repeat and precisely locate the region in a physical map, whereas such repeat arrays are easily identifiable using BioNano's Irys® System.

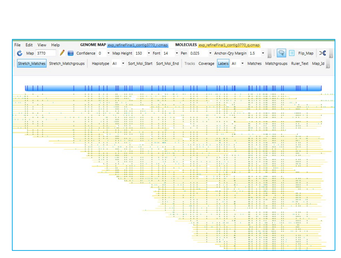
Methods

1. IrysPrep® Isolate High Molecular Weight DNA Molecules
2. IrysPrep® Label DNA Molecules at Motif-Specific Locations
3. IrysChip® Linearize DNA Molecules in NanoChannels
4. Irys® Automatically Image Linearized DNA Molecules
5. IrysView® and IrysSolve® Assemble, Visualize, and Analyze Genome Data





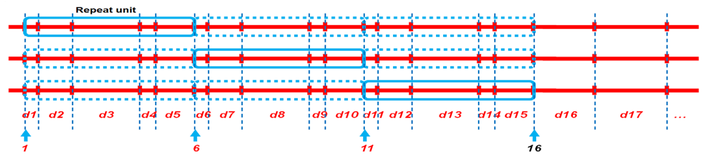




Algorithm Design

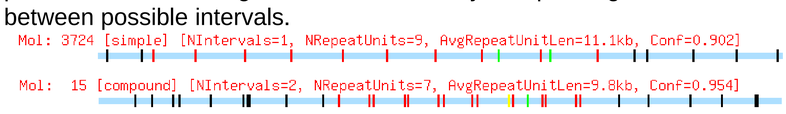
The tool uses a pattern-matching algorithm to detect similar patterns of the labels on the molecules. Specifically, the algorithm scans from the leftmost label to the right. For each label, the algorithm then carves out a segment, constrained by the user-defined length limit.

Then, for this particular segment, the algorithm iterates the number of intervals in each repeat unit, from 1 to a user-defined value. It then calculates the length of the first repeat unit and continues to find the closest labels at locations of the integer multiples of the average repeat unit length until it reaches the end of the current segment or the similarity score between the current repeat unit and the other repeat units fall below the pre-defined cutoff value.



To find the best alignment at a specific position, the algorithm finds all potential repeats with different number of intervals and number of repeat units. The repeat with the largest product of the two values will be considered as the final repeat at this position.

To quantify the similarity between two repeat units, we use the L1 family similarity function, a variation of Euclidian Distance (i.e., Manhattan Distance) function, due to the nature of the problem. Furthermore, false positives and false negatives are handled by interpolating the extra labels between possible intervals.



Results

We reliably detected NOR repeats in different organisms.

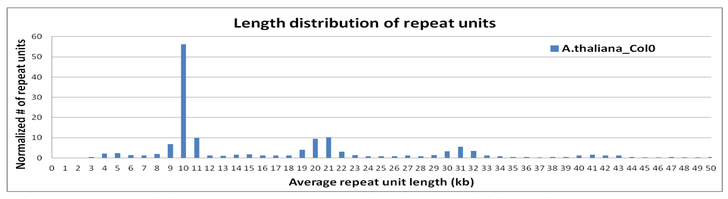


Figure 1: 10 kbp NOR repeat units in *Arabidopsis*.

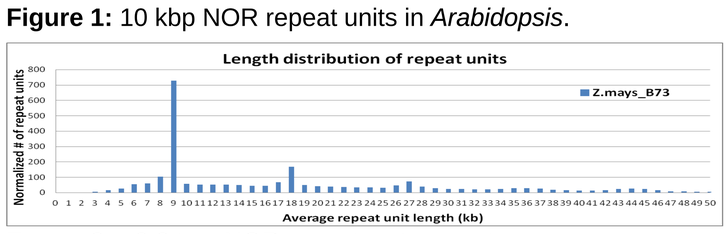


Figure 2: ~8.8 kbp NOR repeat units in maize.

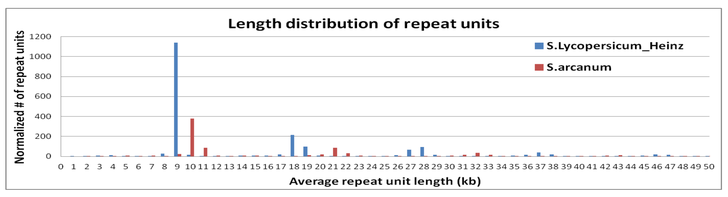


Figure 3: 9 kbp NOR repeat units in cultivated Heinz tomato, whereas 10 kbp repeats in wild *S.arcanum* tomato. Size and copy number differences between cultivated and wild tomato indicates rDNA genes are undergoing positive selection through domestication.

We accurately estimated the copy number of NOR repeats.

By repeat detection tool: As shown in the left figures, smaller genome (120 Mbp) *Arabidopsis* (Columbia0) contains ~150 copies, while larger genome (2.5 Gbp) maize (B73) contains ~4000 copies of NOR repeat units.

By mapping and assembly: BioNano assembled a repetitive region of 26.43 Mbp, at an average coverage of 75.51. Normalizing by the average coverage of the whole genome (57.35x), the normalized length becomes 34.94 Mbp, which represents ~4000 copies of the 8.8 kbp NOR repeat units.

NOR repeats are detected as monomers, dimers and polymers, which may represent a higher level organization of the region.

Precise localization of 43S rDNA repeats in maize (B73) genome.

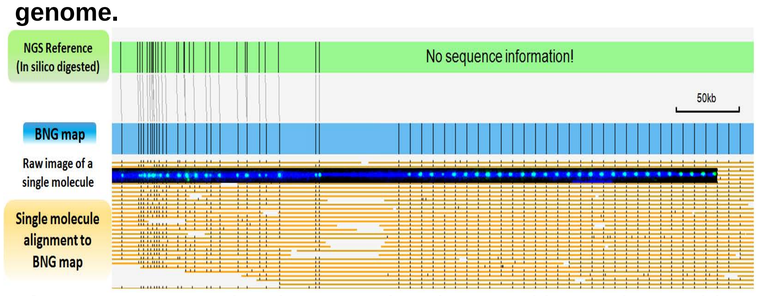


Figure 4: Long molecules spanning the repetitive region and its flanking non-repetitive region can be aligned to specific genome location.

Conclusions

BioNano Genomics Irys® System provides a platform for mapping and visualizing extremely long molecules, which make it possible to analyze long tandem repeat arrays comprehensively, accurately, and without bias. Accompanied by our repeat detection tool, repetitive units of the complex genomes can be profiled, analyzed, and quantified unprecedentedly.

Our results show that Nucleolus Organizer Region repeats can be reliably detected in various species, which further allows us to estimate the copy numbers accurately as well as align the regions to chromosomal locations precisely. See also Posters: P2078, P0702, P0236, and P1272.

Reference

- Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology. 2012 Aug; 30(8):771-6
- Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J, Luo MC, Gu Y, Xiao M. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. PLoS ONE 8, e55864 (2013)
- Angel C. Y. Mak, Yvonne Y. Y. Lai, Ernest T. Lam, Tsz-Piu Kwok, Alden K. Y. Leung, Annie Poon, Yulia Mostovoy, Alex R. Hastie, William Stedman, Thomas Anantharaman, Warren Andrews, Xiang Zhou, Andy W. C. Pang, Heng Dai, Catherine Chu, Chin Lin, Jacob J. K. Wu, Catherine M. L. Li, Jing-Woei Li, Aldrin K. Y. Yim, Saki Chan, Justin Sibert, Željko Džakula, Han Cao, Siu-Ming Yiu, Ting-Fung Chan, Kevin Y. Yip, Ming Xiao and Pui-Yan Kwok, "Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays". Genetics, Jan 2016 (Centennial issue) 115.183483
- S. H. Cha. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International J. Mathematical Models and Methods in Applied Science. vol. 1, no. 4, pp.300 -307, 2007
- Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2010;38:D5-D16