Detection, Characterization, and Biological Analysis of Long Tandem Repeats Using Nanochannel Technology

S Chan[1], W Wang[1], B ten Hallers[1], SA Peters[2], P Gaiero[3], H de Jong[3], GS Perez[2], A Hastie[1], H Cao[1]
[1]BioNano Genomics, San Diego, California, USA
[2]Plant Research International – Applied Bioinformatics, Wageningen, The Netherlands
[3]Wageningen University – Department of Genetics, The Netherlands

Abstract

Many plant and animal genomes are notoriously difficult to assemble because of the vast amount of long tandem repeats which can span several hundred kilobases to multiple megabases. Although the repeat motifs can be sequenced and the amount of repeat material can be approximated by conventional sequencing technologies, it is often difficult or impossible to assemble them into long contigs, so the exact locations and copy numbers of these repeats often remain elusive, especially when the unit length exceeds the read lengths. Without knowing the genomic context of these repeats or the amount of repeat material, it is difficult to attach any biological relevance to them. Thanks to the recent developments in nanochannel technology which allow us to image intact megabase-scale molecules of DNA, repeat regions can be easily spanned and properly placed in the genome assembly. Using Irys® technology and novel algorithms designed specifically to investigate long repeat arrays, we were able to find, quantitate, and characterize previously enigmatic repeat regions in several plant and animal genomes. The genomic positions of these repeats were found by aligning non-repetitive portions of repeat-containing molecules or consensus *de novo* assemblies to the reference genome. The genomic context then provides insight into the biological significance of the repeats.

We find that repeat arrays are often much longer than predicted (or not detected at all) by short-read sequencing; for example, in a tapeworm genome we observe a tandem repeat array spanning at least 350 kb, composed of a 40 kb motif repeat unit, which is not visible by any conventional sequencing methods. We also see trends in both plants and animals that may be of interest from an evolutionary standpoint, such as a prominent 9 kb repeat that is found in many plant genomes but not in animals. Using Irys genome maps, we can scaffold data generated from other sequencing methods, leading to a more complete genome assembly that is much better than either method alone. Furthermore, integration with sequence data allows us to infer the biological function of the repeats. We demonstrate this strategy with an *Arabidopsis* dataset, in which we find a 10 kb tandem repeat that corresponds to the production of rRNA. Our investigation shows that Irys technology is an effective tool for both completing genome assemblies and elucidating the biological significance of previously problematic repeat arrays.

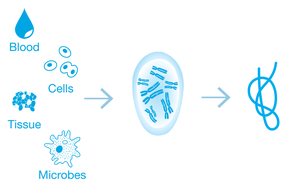
Background

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. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, avoiding the statistical assumptions that are normally used to force sequence alignments of low uniqueness elements. The resulting order and orientation of sequence elements are demonstrated in anchoring NGS contigs and structural variation detection.

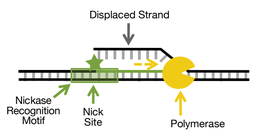
Methods

(1) Long molecules of DNA is 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™ nanochannels 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.

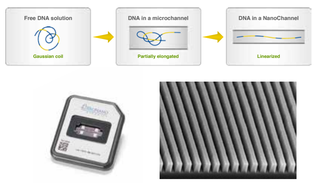
1) IrysPrep Kit extraction of long DNA molecules



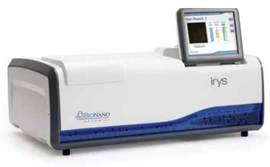
2) IrysPrep reagents label DNA at specific sequence motifs



3) IrysChip linearizes DNA in NanoChannels

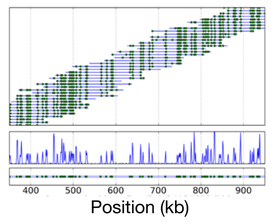


4) Irys automates imaging of single molecules in NanoChannels

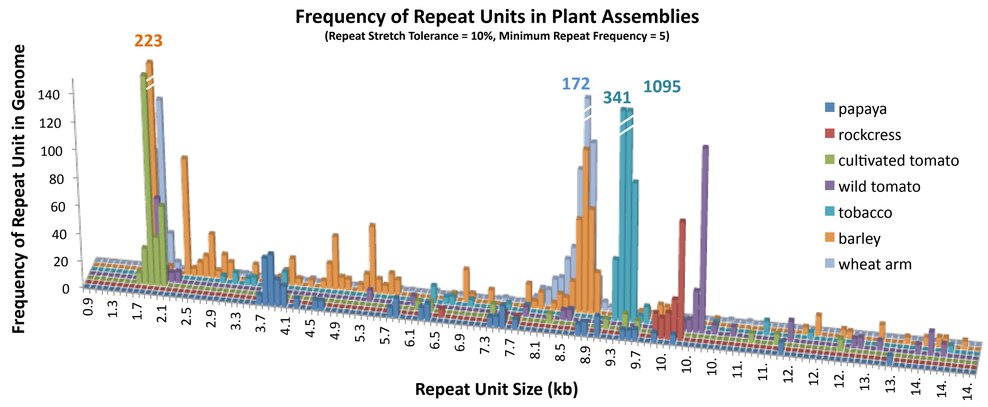


5) Molecules and labels detected in images by instrument software

6) IrysView software assembles genome maps

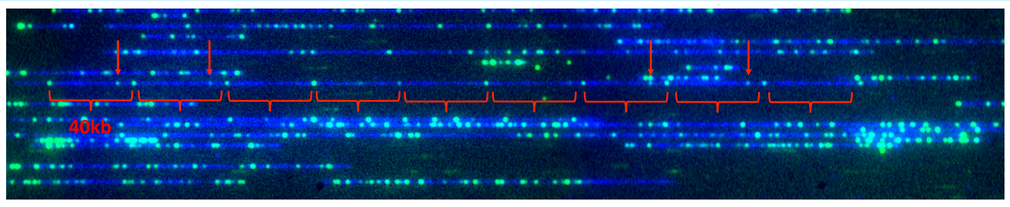


Comparison of Repeats in Plant & Animal Genomes



Comparison of frequency of labeled repeat motifs found in several plant *de novo* assemblies generated from Irys data and software. Only labeled repeats with 5 or more copies were considered. Notice the similarity between wheat arm and barley, both of which are cereal crops, as well as the peak at ~2 kb that is shared among both the grains and the tomatoes. The prominent peak at 10 kb for rockcress (*Arabidopsis*) was shown to correspond to rDNA. Abundance of similar repeats seen in other plants, imply a shared biological function.

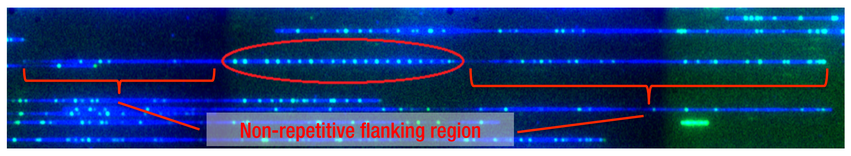
Detection of Intact Long Tandem Repeat Arrays



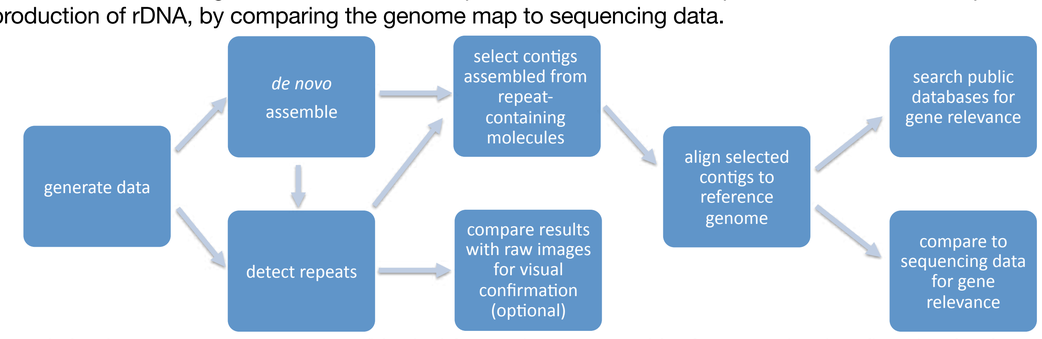
An algorithm was designed to identify tandem repeats with 1 to 2 nick sites per repeat motif. The algorithm allows for stretch variation between labels as well as enzymatic errors that cause extra and/or missing labels. The detected repeats can be traced back to the original image for visual validation. The example above, generated from a tapeworm genome, shows a molecule containing a long tandem repeat array with at least 9 repeat units, each ~40 kb long. The presence of additional labels on 4 of the repeat units (indicated with red arrows) suggests that a variation at the nick site is conserved in some but not all of the repeat units.

Relating Biological Function to Detected Repeats

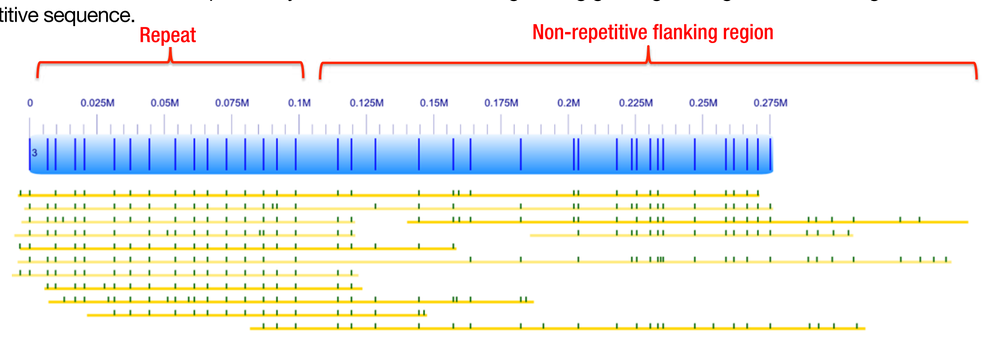
Molecules often capture the entire unbroken repeat array. Molecules that span the repeat or extend past it on either side can be mapped to a reference or assembly. The flanking regions provide genomic context of the array. Unlike other sequencing methods which fail to span the array, Irys technology allows for the detection of both the copy number and the locations of these regions in the genome. The example below is from a human dataset.



The generalized strategy for finding biologically interesting areas of the genome that contain repeats is illustrated below. Using this method, a 10 kb repeat in a strain of *Arabidopsis* was found to correspond to production of rDNA, by comparing the genome map to sequencing data.



Shown below is a consensus genome map (blue bar) from a *de novo* assembly of papaya, showing aligned molecules (yellow bars) used in the assembly. Labels are shown as vertical stripes on the horizontal bars. Several molecules span the entire repeat, and there is enough flanking material to assemble the region. The genome map can be compared to sequence data so the location of the repeat array can be determined. Neighboring genes give insight into the biological function of the repetitive sequence.



Conclusions

Irys technology is a powerful tool for rapid large-scale investigation of plant and animal genomes. As demonstrated here, it can be used to identify large-scale trends in groups of related organisms, which may shed some insight on the effects of evolution and/or breeding. Because a single molecule can span an entire repeat array, the genomic context of the repetitive material is preserved. This allows us to attach biological significance to otherwise rather meaningless data. Another benefit of being able to see molecules that span the array is the ability to find the exact copy number by simply counting label sites, which means that generating artifactual data is highly improbable. This is particularly useful for analyzing samples with previously unquantitated amounts of repetitive material, which are often completely omitted from assemblies, especially with plant samples. As new discoveries reveal the biological functions of what was once considered “junk” or “filler” DNA, Irys technology will become an indispensable tool for tackling the genomes of both well-studied and newly discovered organisms.

References

- 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
- Lam, E.T., et al. Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly. Nature Biotechnology (2012); 10:2303