Detection, Characterization, and Biological Functions of Long Tandem Repeats Detected in Human Genomes Using Nanochannel Technology

S. Chan, X Zhou, Ž Džakula, A Hastie, H Cao
BioNano Genomics, San Diego, CA

Abstract

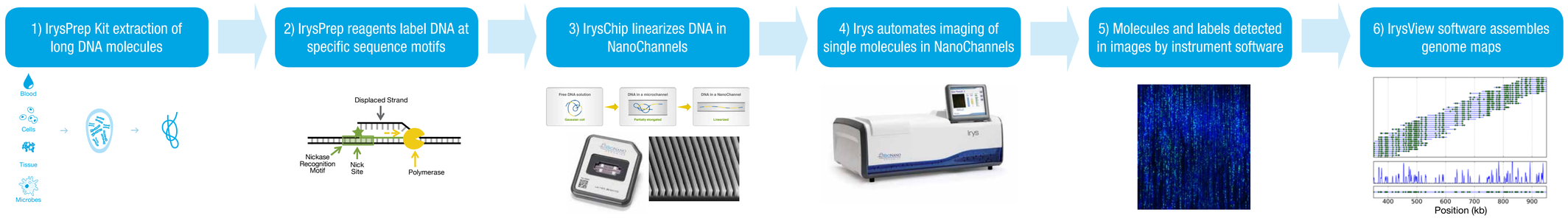
A large portion of the human genome is known to be composed of long tandem repeats that can span several hundred kilobases to multiple megabases. Although the repeat motifs can be sequenced and the amount of repeat material can be approximated by conventional sequencing technologies, it is often difficult 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 more accurately characterized and put into biological context. Using Irys® technology and novel algorithms designed specifically to investigate long repeat arrays, we were able to find and characterize previously enigmatic repeat regions in the human genome. 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. Here, we have found that molecules containing a prominent 5.6-kilobase tandem repeat are derived from the Lipoprotein(a) (Lp(a)) coding region on chromosome 6, which is linked to atherosclerotic diseases such as heart disease and stroke. Furthermore, since the array is contained on single molecules, we can measure allele array lengths, showing that Irys technology has potential for aiding in quick, accurate, and cost-efficient prognosis of these and other genetic diseases which are influenced by copy number variation.

Background

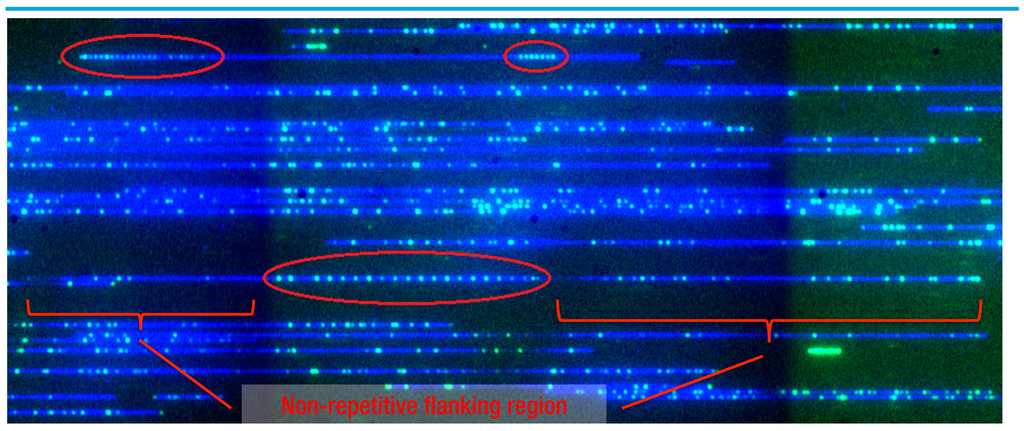
Generating high quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, 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.



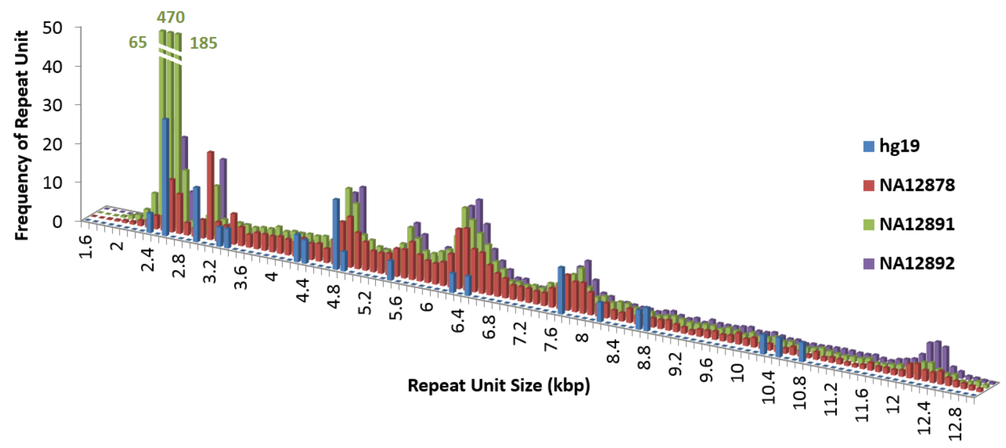
Detection of Unbroken Long Tandem Repeat Arrays



An algorithm was designed to identify tandem repeats with one to two 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. In the example data above, detected repeats are circled in red.

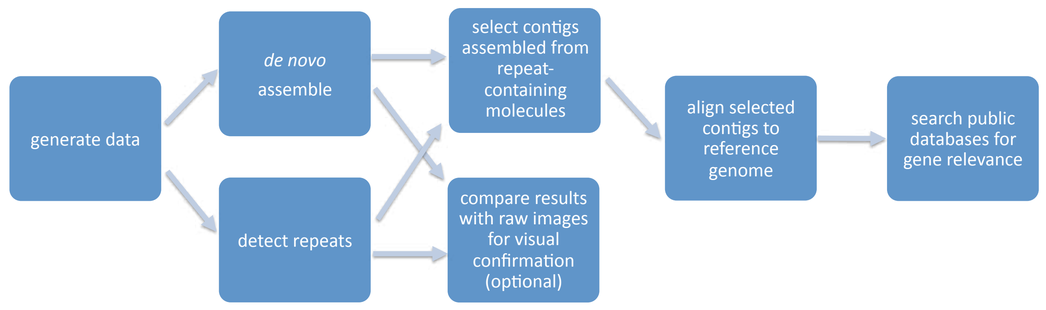
Molecules often capture the entire unbroken repeat array. Molecules that span the repeat or extend past it on either side can be mapped to a reference or assembly. The flanking regions provides genomic context of the array. Unlike other sequencing methods which fail to span the array, Irys technology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

Prevalence of Repeats in Human Datasets



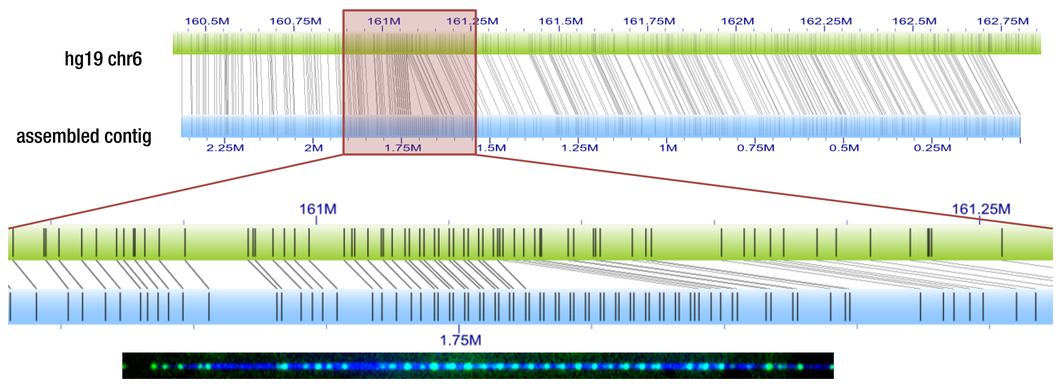
Comparison of frequency of repeat motifs found in hg19 reference and trio assemblies (normalized to 1X coverage). Only repeats with 5 or more tandem repeat units were considered. Peaks are smaller and less defined for larger repeat motifs due to increased likelihood of variation (and therefore extra or missing nick sites) among longer repeat units. Prominent peak at 2.5 kb in NA12891 (father) has been seen in other human data and appears to be male-specific.

Relating Biological Function to Detected Repeats



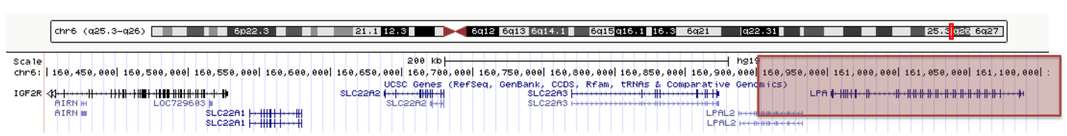
Generalized strategy for finding biologically interesting areas of the genome that contain repeats.

Copy Number Variation in Lp(a) Coding Region



A 5.6-kb tandem repeat containing 14 repeat units was found in a human dataset. The data was *de novo* assembled and then aligned to the hg19 reference. The contig assembled from the repeat-containing molecules aligned to chromosome 6. Maps are displayed here as bars with vertical stripes at nick sites. A raw image of a representative molecule is shown beneath the contig map for visual confirmation of the assembled region.

The coordinates of this repeat array were found to correspond to the Lp(a) coding region, by using the UCSC genome browser (shown below).



Conclusions

Irys technology is a powerful tool for rapid large-scale investigation of the human genome. 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. If a correlation can be drawn between copy number variation and particular diseases, the Irys platform has potential to provide cost-effective and quick diagnosis of such diseases.

References

1) 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.
2) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303