Mapping the "Dark Matter" of the Genome with Super Long Molecules - The Unknown Unknown

H Cao[1], A Hastie[1], A Pang[1], E Lam[1], W Andrews[1], T Anantharaman[1], T Chan[1], X Zhou[1], J Reifenberger[1], M Saghbini[1], H Sadoski[1], M Austin[1], P Sheth[1], Z Dzakula[1], T Dickinson[1], X Xun[2], T Graves[3], A Bashir[4], P-Y Kwok[5]

[1]BioNano Genomics, San Diego, CA, 92121, USA; [2]BGI, Shenzhen, China; [3]Washington University, Saint Louis, MI, USA; [4]Mt Sinai School of Medicine, New York, NY, USA; [5]UCSF, San Francisco, CA, USA

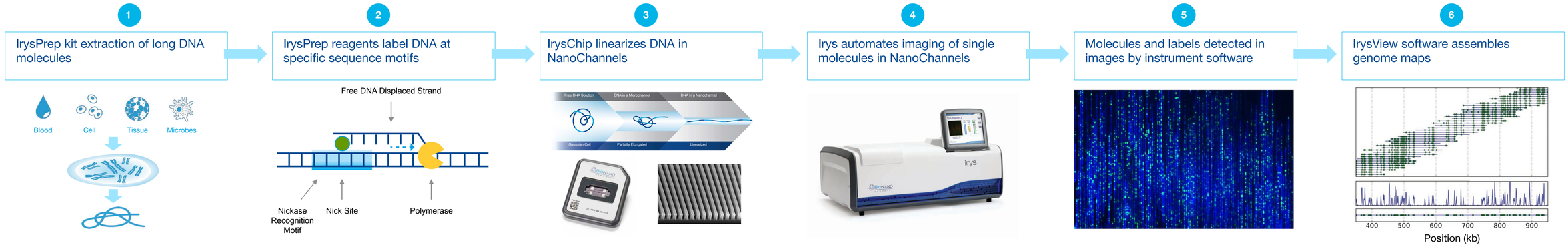
Abstract

In spite of advancements made in high-throughput next-generation sequencing in the past decade, a large portion of the human genome remains unresolved or ambiguously characterized. Especially, large genomic structural variations (SV, > 1 kb) known to be associated with complex traits diseases, are found to be more prevalent than previously thought. Often challenging to detect for short-read NGS and conventional low-resolution cytogenetic techniques, these large SVs – the “dark matter” of the genome – result in gaps and unknown structural information in assemblies. Rapid comprehensive genome mapping in nanochannel arrays provides a new single-molecule imaging platform independent of, yet complementary to DNA sequencing for accurate genome assembly and structural variation analysis. Extremely long molecules of hundreds to thousands of kilobases, fluorescently labeled at sequence motifs and elongated in nanofluidic channels, enable direct image interrogation of comprehensive genome architecture at a high resolution. *De novo* assembly of these single molecules yields unprecedented long contiguous genome maps, advantageous in spanning over highly repetitive regions and complex structures in their native form. Here, we present results from analyses on normal human genomes, cancer genomes, and other complex genomes. We detected hundreds of large structural variants and haplotype differences in these genomes. In one human genome, we obtain hundreds of insertions/deletions and inversions larger than 1 kb. Without considering SVs that overlap with N-base gaps in hg19, 90% of these SVs are supported by orthogonal experimental methods or historical evidence in public databases. A larger portion of the complex genome is composed of previously unknown repeat material spanning tens of kilobases to multiple megabases, whose exact locations and copy numbers remain elusive to NGS. Without knowing the genomic context of these repeats or the amount of repeats, it is difficult to attach any biological significance to them. Using BioNano’s Irys® platform and novel algorithms designed specifically to investigate long repeat arrays, we were able to find and characterize clinically relevant repeat regions linked to human cardiovascular disease risk. For the first time, population-scale cross-sample genome comparison to identify comprehensive genomic structural variation is feasible on a single platform due to the quick turnaround time. Lastly, we present the first assembly of a diploid human genome that combines long-read NGS sequencing with Irys genome maps, resulting in a hybrid assembly with a map N50 exceeding 31 Mb, a dramatic improvement upon the assembly contiguity typically observed in shotgun sequencing approaches. Overall, genome mapping provides valuable structural information otherwise hard or impossible to decipher with short-read sequencing data alone, and paves the road for generating true contiguity in ultimate medical-grade and breeding grade-genome information.

Background

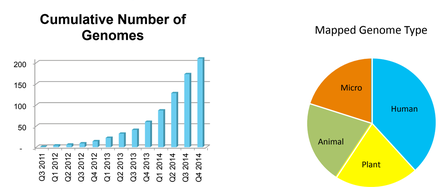
Generating high-quality finished genomes replete with accurate identification of structural variation and highly complete (minimal gaps) remains challenging using short-read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing the statistical inference needed to align paired-end reads with an uncertain insert size distribution. These long labeled molecules are *de novo* assembled into physical maps spanning the whole genome. The resulting order and orientation of sequence elements in the map can be used for anchoring NGS contigs and detecting structural variation.

Methods

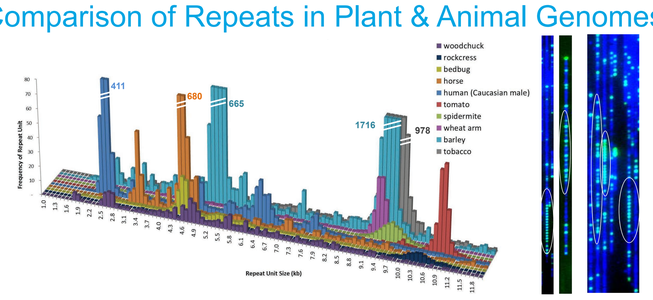


(1) Long molecules of DNA is labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® nanochannels and single molecules are imaged by Irys. (4) Single molecule 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.

200+ Unique Genomes Analyzed on Irys

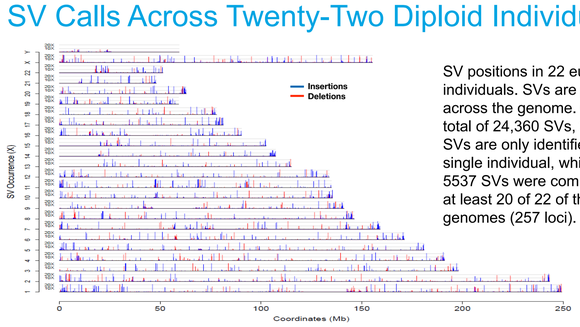


Comparison of Repeats in Plant & Animal Genomes



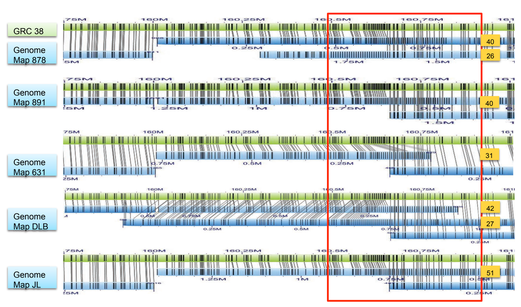
Comparison of the frequency of labeled repeat motifs found in several datasets generated from Irys data and software. Only labeled repeats with 5 or more copies were considered. Notice the similarity between barley and wheat, both of which are cereal crops. The prominent peak at 10 kb for rockcress (*Arabidopsis*) was shown to correspond to rDNA. The 2.4kb repeat in a human male is not as prominent in females and is believed to lie primarily on the Y chromosome.

SV Calls Across Twenty-Two Diploid Individuals



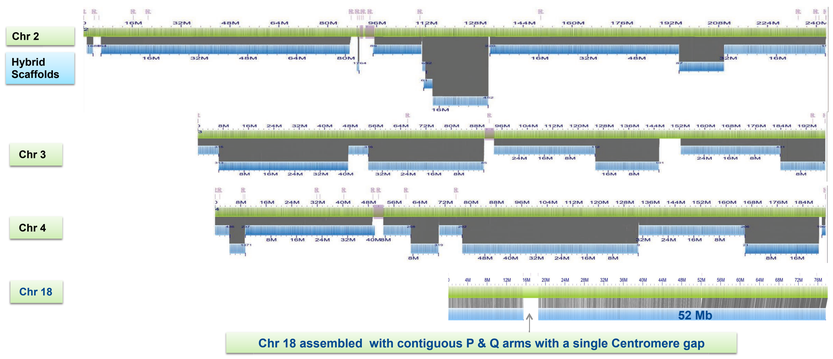
SV positions in 22 euploid individuals. SVs are plotted across the genome. Of a total of 24,360 SVs, 2655 SVs are only identified in a single individual, while 5537 SVs were common to at least 20 of 22 of the genomes (257 loci).

Lp(a) Locus: Copy Number Variation



5.6 kb Lp(a) domain copy numbers are highly variable and stably inherited within a family. Its copy number is directly associated with plasma concentration of Lp(a) lipoprotein and inversely proportional to Coronary Heart Diseases (CHD) risk.

Toward True Contiguity – Examples of Chromosomal-Level Hybrid Scaffold



Chr 18 assembled with contiguous P & Q arms with a single Centromere gap

	# Contigs	NGS length**	N50
NGS input*	4,528	n/a	1.1 Mb
BioNano	1,003	n/a	4.6 Mb
Hybrid Scaffolds	202	**94%**	**31.3 Mb**

* All contigs with at least 10 NtBspQI sites
** Length of NGS input (2.707 Gb)

Overview of the results of the co-assembly. N50 values are **30 times** better than PacBio *and Illumina combined*; **6 times** better than Genome Map alone.

NGS contigs were converted to *in silico* maps and those with >10 nick sites were selected for hybrid scaffolding. The hybrid scaffold was produced by overlap merging on the map level at pattern matches to produce the most contiguous map possible. Genome maps and NGS contigs were excluded from the analysis if they had conflicting alignments. A more aggressive NGS assembly was used in a second round of merging of maps, this could be done since the risk of producing chimeric maps was low since conflicting alignments were removed. The final hybrid scaffold was used to anchor NGS contigs.

Conclusions

Here we have presented several examples in order to demonstrate the significant value added to genomic research which is only possible by directly interrogating extremely long reads. Long arrays of tandem repeats containing individual units, each of which are longer than nearly any NGS sequence read and full array lengths far beyond the longest sequence reads available today can be seen in single molecule Irys reads. Thousands of large SVs are identified in Irys data, some of which are invisible or unphasable by other methods. As demonstrated with the analysis on the Lp(a) gene, the Irys platform has the potential to provide quick, cost-effective diagnosis of genetic diseases linked to copy number and structural variation. These data underlie the importance of complete *de novo* assembly by long read technology for personalized genomics.

Reference

1) Cao, H., et al., Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience (2014); 3(1):34
2) 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.
3) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
4) Das, S. K., et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Research (2010); 38: 8
5) Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.