Identifying Polymorphic Variants at Complex Regions Using Optical Mapping

Joyce Lee[1], Feyza Yilmaz[2], Andy Wing Chun Pang[1], Alex Hastie[1], Human Genome Structural Variation Consortium
[1]Bionano Genomics, San Diego, California, United States of America
[2]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, United States of America

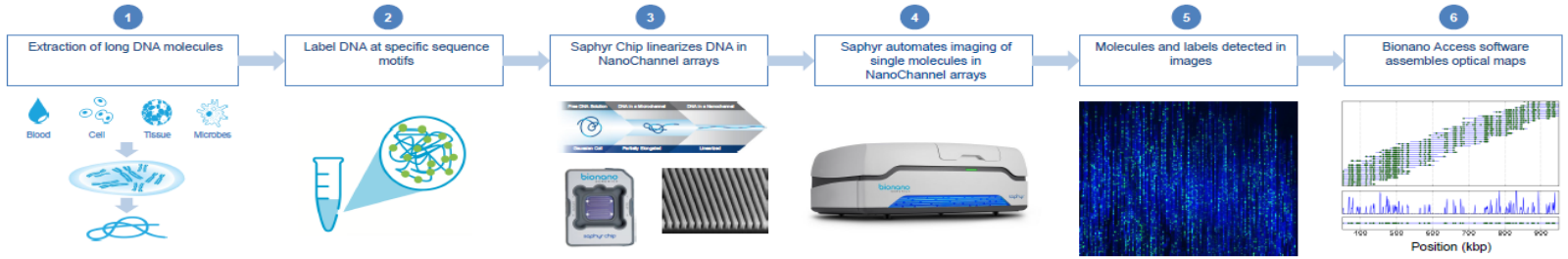
Abstract

The human reference (GRCh38) released in 2013 is by far the best annotated human assembly, and it has been widely used in clinical studies. However, the reference is a haploid representation of multiple donors, and it does not include all alleles of the human species. There are regions in the human genome that are polymorphic between individuals, and previous studies[1,2] have showed that polymorphic regions can be predisposed to segmental duplications (SD). SDs are regions with ≥ 1000 bp and have at least one highly identical (>90%) copy in the genome, making them very challenging to traditional short-read or even long-read sequencing technologies. One of the methods to solve these SDs is by using high-molecular weight DNA that can span the flanking of these SDs, and Bionano optical mapping uses such methodology and demonstrates superior performance in assembling and identifying large polymorphic variants in such challenging regions.

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. The Saphyr™ system 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 entire diploid genome. The resulting provides the ability to correctly position and orient sequence contigs into chromosome-scale scaffolds and detect a large range of homozygous and heterozygous structural variation[3] with very high efficiency.

Method



(1) Long molecules of DNA are labeled with Bionano reagents by (2) incorporation of fluorophores at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the Saphyr Chip using NanoChannel arrays (4) Single molecules are imaged by Saphyr and then digitized. (5) Molecules are uniquely identifiable by distinct distribution of sequence motif labels (6) and with pairwise alignments, molecules can be assembled into local maps or whole genome *de novo* assemblies.

Result

Using Bionano's DLE-1 optical mapping data of 30 samples from five ethnicities, which are also part of the Human Genome Structural Variation Consortium samples, we detected a total of 19821 insertions and deletion (≥ 5kbp) at clean regions where no more than two maps are aligned. After removing redundant calls between samples, 3995 variants remained, of which 516 insertions and 331 deletions were localized at regions with at least five different variants identified in at least seven different samples. About 75% of these polymorphic variants overlapped SDs, and long read sequencing detected about 40% of the 847 variants at polymorphic regions. Aside from insertions and deletions, we also identified 162 non-redundant inversions and 192 duplications across all samples, of which 54% and 58% of them overlapped SDs respectively.

Manual curation of variants at SDs revealed highly complex regions where Bionano assemblies outperformed long read sequencing and clearly characterized the structures and identified variants with high accuracy. For example, at a SD region on chr1 (108.25 – 108.45 Mbp) which overlaps the neuroblastoma breakpoint family (*NBPF*) (Fig.1), a 75 kbp inversion was found in 22 out of the 30 samples. While Bionano showed contiguous molecules across the SD region revealing the inversion that was confirmed by Strand-seq, long read sequencing assembly is broke up at the region.

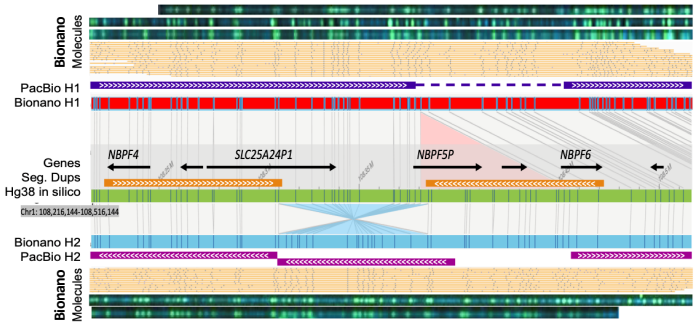


Fig.1 HG00733 SV locus flanked by SDs. Bionano maps spanned the SD region and detected a 75 kbp inversion in one allele and a 74 kbp deletion in the alternate allele.

In the alternate allele in two of the samples, a 74 kbp deletion was found and is confirmed by long read sequencing. Another 74 kbp deletion and 130 kbp inversion were found in two other samples. Overall, all 30 samples have contiguous optical maps spanning the SD region and revealing the polymorphic variants (Fig.2).

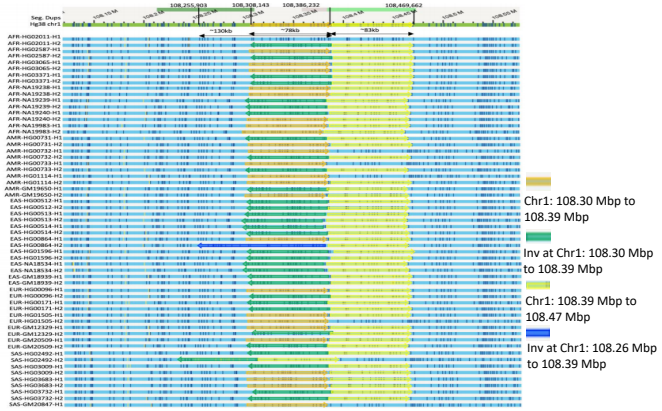


Fig.2. In all 30 samples, Bionano optical maps contiguously spanned the SDs and revealed various structures of the samples.

Conclusion

In conclusion, with high molecular weight DNA, Bionano assemblies contiguously span complex regions such as SDs which are hard to access by other technologies. The contiguity allows Bionano to efficiently reveal the complete structure and detect polymorphic variants with high sensitivity in populations.

Reference

1. Bailey, Jeffrey A., et al. "Segmental duplications: organization and impact within the current human genome project assembly." *Genome research* 11.6 (2001): 1005-1017.
2. Mehan, Michael R., Nelson B. Freimer, and Roel A. Ophoff. "A genome-wide survey of segmental duplications that mediate common human genetic variation of chromosomal architecture." *Human genomics* 1.5 (2004): 1-10.
3. Bionano Genomics, Inc. (2020). Bionano Solve Theory of Operation: Structural Variant Calling. Document Number: 30110, Revision J.