De Novo Assembly and Structural Variation Discovery in in Human Disease and Non-Disease State Genomes Using Extremely Long Single-Molecule Imaging

A. Hastie, E. Lam, M. Imielinski[1], C.-Z. Zhang[1], J. Wala[1], A. Pang, S. Chan, W. Andrews, H. Dai, Ž. Džakula, H. Cao
BioNano Genomics, San Diego, CA
[1] Broad Institute, Boston, MA

Abstract

Structural variation analysis (SVA) of human genomes is usually a reference based process and therefor biased and incomplete. In order to have a comprehensive analysis of structural variation, a *de novo* approach is needed. *De novo* genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in the human genome. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. As a result of the remaining limitations of DNA sequencing and analysis technologies, it is not feasible to create high quality assemblies of individuals to detect and interpret the many types of structural variation that are refractory to high throughput or short-read technologies.

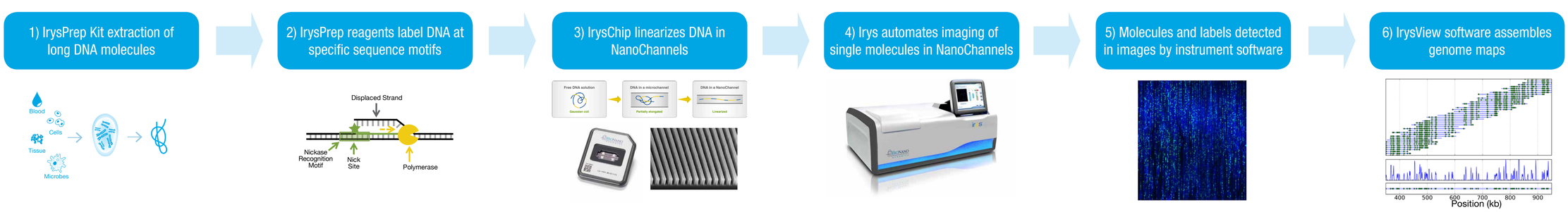
We present a single molecule genome analysis system (Irys®) based on NanoChannel Array technology that linearizes extremely long DNA molecules for direct observation. This high-throughput platform automates the imaging of single molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness for unambiguous assembly of complex genomes. High resolution genome maps assembled *de novo* preserve long-range structural information necessary for structural variation detection and assembly applications. Dozens of human genomes have been *de novo* assembled by Irys to date, including cancer genomes. Structural variation analysis reveals insertions, deletions, inversions and translocations. Each genome shows dramatic structural variation even when considering only normal (non-disease state) individuals, including many megabases of variation within genomic regions not included in the public reference genome assembly (GRCh38), underscoring the need for more *de novo* approaches to genome analysis. In some cases, genome maps can identify translocations partners whose path pass through hundreds of kilobases of the genome which is absent from the reference. Extremely long single molecules can also be used to phase rearrangement breakpoints on the same derivative chromosome; something that can only be inferred by short read and copy number technologies.

Background

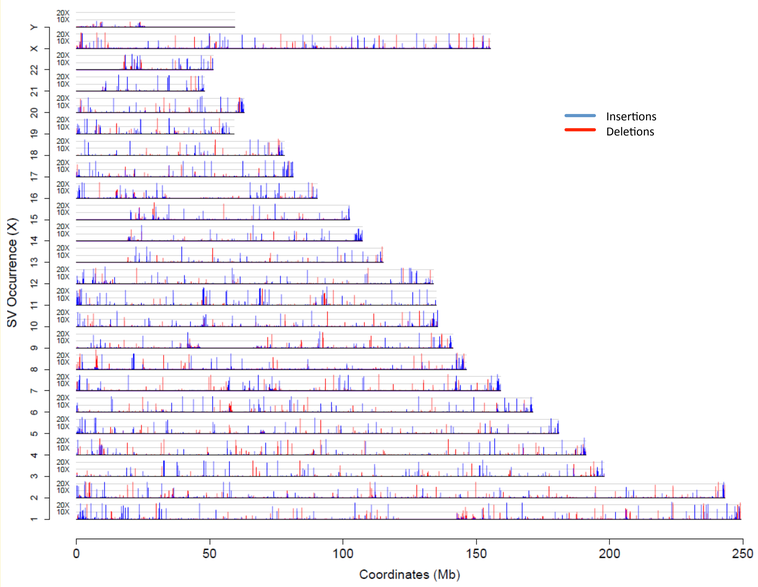
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing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

Methods

(1) Extremely long DNA is extracted from the source sample and (2) labeled with IrysPrep™ reagents by 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) Irys performs automated data collection and image processing. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable. (6) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite.

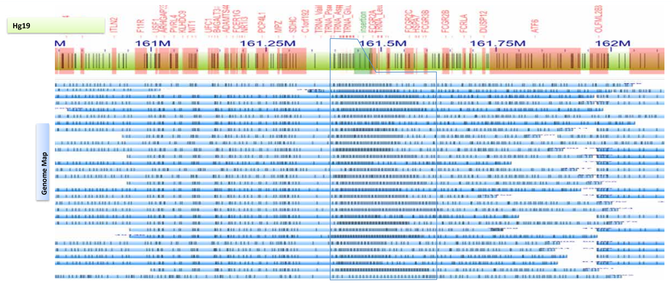


SVs Across 22 Diploid Individuals



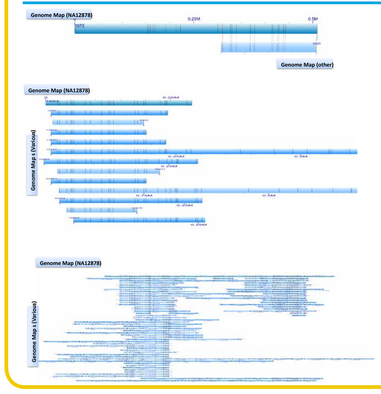
SV positions in 22 Euploid individuals. SVs > 1.5kb are plotted across the genome. Of a total of 24,360 SVs, 3419 SVs are only identified in a single individual, while 5616 SVs were common to at

Common CNV in a tRNA region in Human Individuals



Copy number polymorphism in a tRNA gene cluster in Chromosome 1. The length of the tRNA cluster in the reference is 50 kb while the length ranges from 85- 210 kb in the individuals that were mapped.

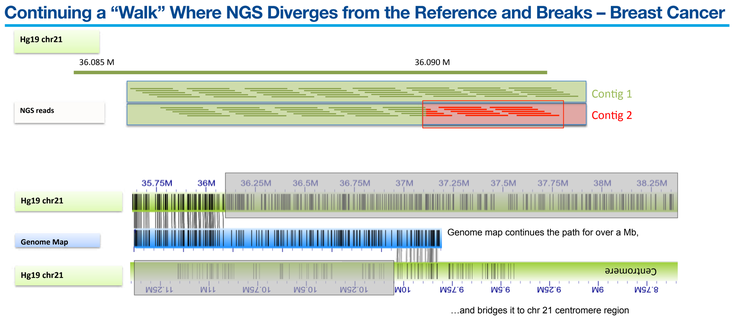
Unreferenced DNA in Genome Maps



NA12878 Map ID	Length	Occurrence in other genome maps from other individuals
1672	509,474	1
9536	197,386	1
1522	274,787	2
7550	169,684	2
1576	280,393	3
2592	100,713	3
4723	123,662	3
1352	286,682	4
1793	136,730	4
1011	697,855	5
2752	121,323	5
7553	129,557	5
1702	366,784	6
14469	479,914	6
1019	593,042	8
1471	706,429	8
5274	143,866	8
1071	449,186	9
1219	310,568	9
1330	152,780	10
1409	304,891	10
1180	450,435	11
1601	277,510	11
5489	120,344	11
1172	440,745	12
1292	715,096	12
1693	292,896	12
1406	292,504	13
1459	639,244	13
1475	425,298	14
1279	687,797	15
1336	576,921	15
1017	529,133	16
1080	857,117	18
1569	613,189	26
1103	963,954	29
9698	548,212	29
972	607,283	31
572	1,118,409	43
1061	1,264,805	44
1245	529,630	44
920	2,493,496	61
713	2,979,866	98
99	946,866	214

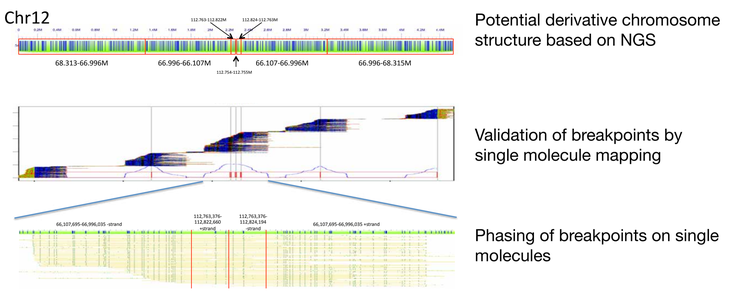
In the NA12878 genome map, 26Mb from 51 maps do not align to the GRCh38 reference. When comparing the 51 unmapped maps to other genome maps, most align to genome maps assembled from multiple different individuals, examples are shown on the left. Unplaced NGS sequence can be anchored to the unmapped genome maps.

Continuing a "Walk" Where NGS Diverges from the Reference and Breaks – Breast Cancer



NGS *de novo* assembly at a somatic SV is broken by repeat sequences (cartoon shown on top, mapped reads in green and divergent reads in red). The genome map was assembled into a 1.75Mb contig that spans repetitive DNA and almost 1Mb of unreferenced DNA to identify the rearrangement partner.

Rearranged Chromosome with Phased Breakpoints – Breast Cancer



The top green map is a model of a derivative chromosome based on NGS and local *de novo* assembly of breakpoints, each fragment is in a red box. Single molecules were aligned to the breakpoints to validate. Single molecules map to each of four fragments of chromosome 12 on the bottom, thereby phasing the fragments onto a single DNA strand.

Conclusions

BioNano Genomics genome mapping is a powerful tool for detection of structural variation in human individuals. Many SVs are common in our dataset suggesting that the hg19 reference contains mistakes or uncommon variants. Genome mapping and single molecule reads are useful for identifying translocation partners where repeat sequences break sequence assemblies and even through unreferenced DNA. Single molecule reads are able to phase multiple fragments of a rearranged derivative chromosome in a breast cancer sample.

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
3) 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
4) Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.