Rapid Detection of Large Structural Variations in a Human Genome Using NanoChannel Genome Mapping Technology

A Hastie[1,5], HZ Cao[2,3,5], DD Cao[2,5], E Lam[1,5], Y Sun[2,4], H Huang[2,4], W Andrews[1], M Requa[1], T Ananthara-man[1], M Austin[1], M Sagbini[1], H Vansteenhouse[1], S Chan[1], A Krogh[3], H Cao[1], X Xu[2]

[1]BioNano Genomics, San Diego, California, USA, [2]BGI-Shenzhen, Shenzhen, China, [3]Department of Biology, University of Copenhagen, Copenhagen, Denmark, [4]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China, [5]These authors contributed equally to this work

Abstract

Large structural variations (SVs) are less common than SNPs and indels in the population but collectively account for a significant fraction of genetic polymorphism and diseases. Base pair differences arising from SVs are on a much higher order (>100 fold) than point mutations; however, none of the existing prevailing methods can comprehensively and effectively detect them. To address these challenges, we first applied a high-throughput, cost-effective genome mapping technology using long single molecules (>150kb) to discover genome wide SVs and structure differences in the YH genome. We detected 278 large SVs (>10 kb), of which 251 of 278 (90%) are retrospectively supported by multiple orthogonal methods such as whole genome or fosmid end sequencing (200/278) and historical evidence contained in the DGV database (51/78). To further investigate the SVs that couldn't be validated by sequencing based tests, we found that 71 out 78 (91%) intersected with repeat elements, often the blind spot of re-sequencing and *de novo* assembly methods. More than 70% of detected SVs are insertion events, known to be difficult to detect by sequencing. In this study, genome mapping also provides valuable information for complex regions (MHC, KIR, TRB/TRA, IGH/IGL et.al) with haplotypes.

In addition, for the first time, with long single molecule labeling patterns, inserted exogenous viral sequence and locations can be mapped on a whole genome scale important for understanding virus induced oncogenesis.

NanoChannel based genome mapping make it now feasible and cost effective to conduct large population-based comprehensive SV studies efficiently on a single platform.

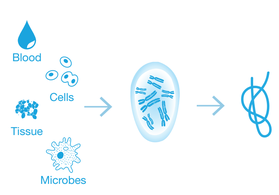
Background

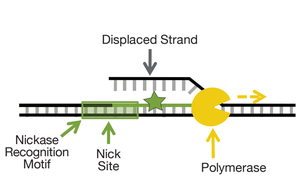
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing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

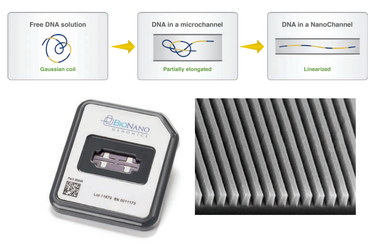
Methods

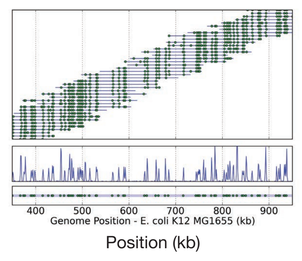
(1) Extremely long DNA is extracted from the source sample and (2) labeled with IrysPrep™ reagents by incorporation of fluorophore-labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules 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.

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

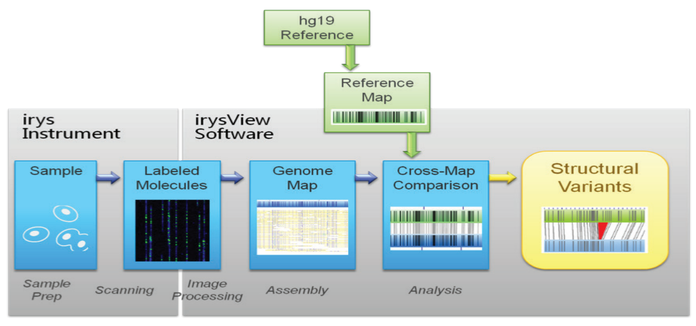








Irys Workflow for SV Discovery by *De Novo* Genome Mapping of a Human



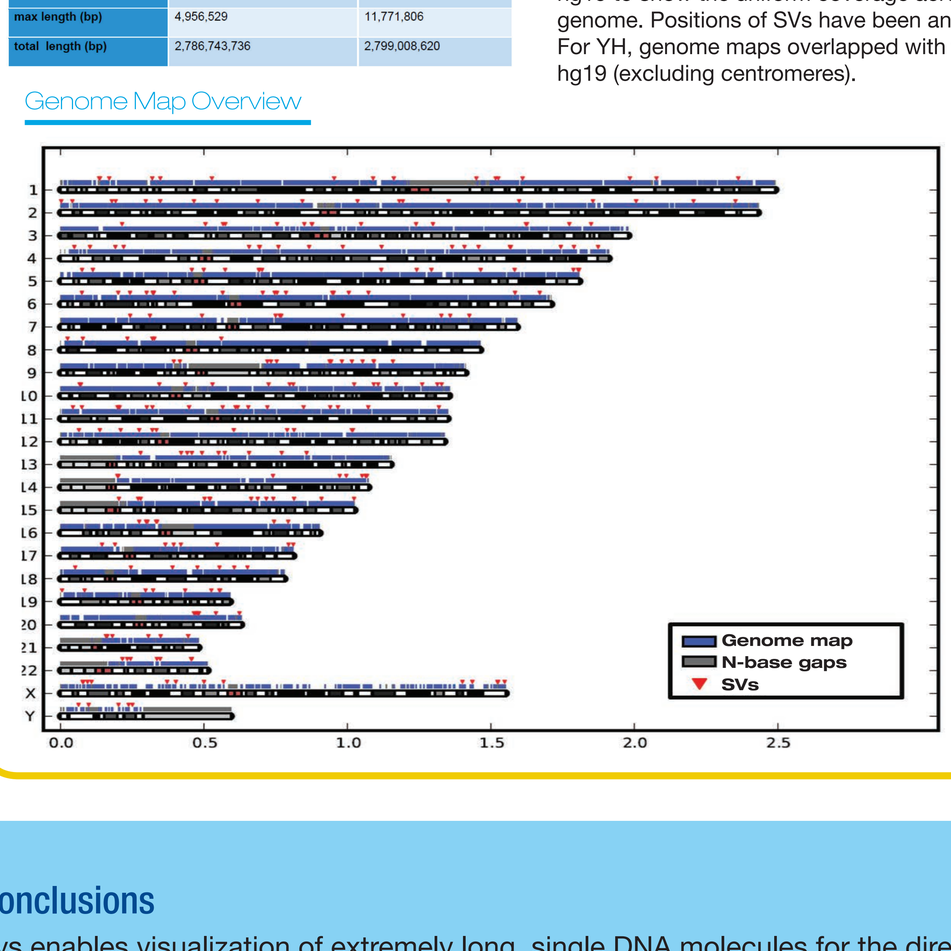
Flow chart of the Irys workflow for SV detection in the human genome. Single molecule images are translated into digital maps used to produce a *de novo* assembly through a overlap-layout-consensus algorithm. The genome map is compared to an *in silico* reference map (or other map) to detect structural variation.

Genome Map Assembly Statistics

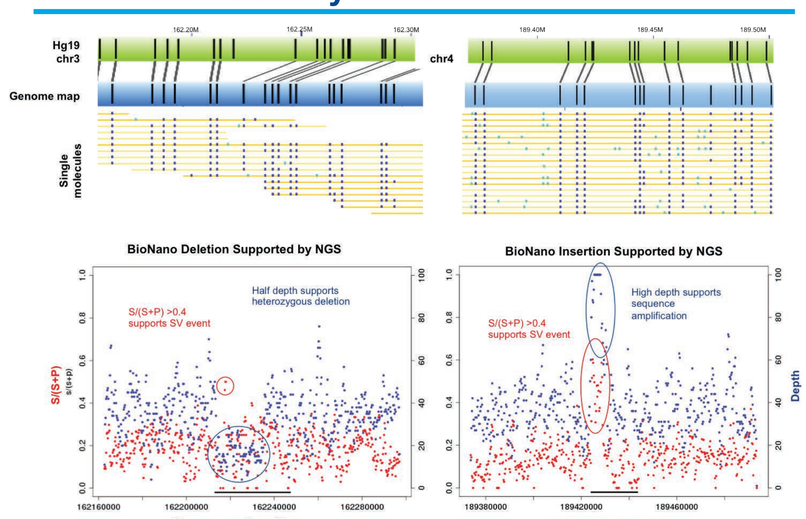
	Pre-stitch	Post-stitch
# maps	3,565	1,634
min length (bp)	90,350	90,350
median length (bp)	599,630	1,096,601
mean length (bp)	781,695	1,712,980
N50 length (bp)	1,027,446	2,868,628
max length (bp)	4,956,529	11,771,806
total length (bp)	2,786,743,736	2,799,008,620

The table show the statistics of *de novo* genome map assembly and finished genome map assembly. The finished genome map has adjacent contigs merged at positions where they were broken by nicking on both strands. Below, genome maps have been aligned to hg19 to show the uniform coverage across the genome. Positions of SVs have been annotated. For YH, genome maps overlapped with 93% of hg19 (excluding centromeres).

Genome Map Overview

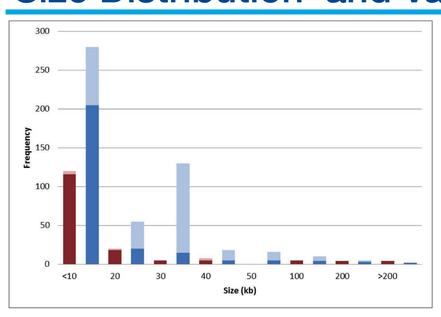


CNV Validation by NGS



Examples of CNV calls in YH are shown above, hg19 in silico map is shown in green with the *de novo* genome map below and single molecule support below that. NGS methods were used to support genome map findings by plotting S/(S+P) and coverage depth (shown) and *de novo* assembly of fosmid sequences (not shown). Examples of a validated deletion and insertion are shown. Because of the very high noise of these methods, they perform poorly in stand alone experiments but can help to support calls by genome mapping.

Size Distribution and Validation of CNVs



Graph shows the net size of 666 SV events found by genome mapping in the YH genome. In total, 60% of SVs were verified by NGS methods. Interestingly, the validation rate for deletions was 93% and 48% for insertions. The lower validation rate for insertions could be attributed to the difficulty of insertion detection by NGS. A total of 88% of all SVs from genome mapping overlap with entries in the DGV database.

Long Labeled and Unlabeled Tandem Repeats

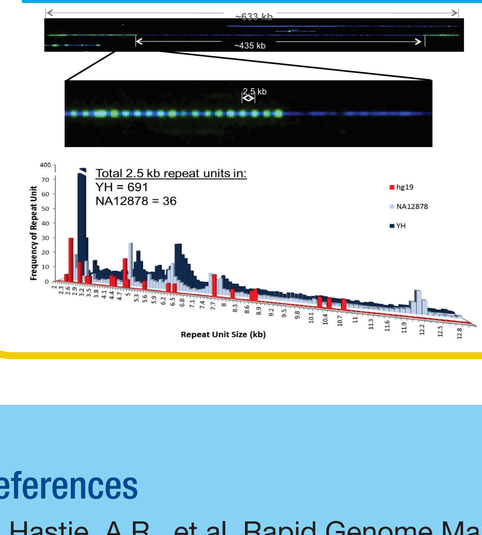


Image of a long DNA molecule with 79 labeled DNA units with the same distance intervals, representing portions of two tandem arrays of 2.5 kb repeats. These arrays are separated by 435 kb of unlabeled DNA which is likely another repetitive DNA unit. The graph shows various labeled repeats in the human genome extracted from single molecules. The 2.5 kb repeat is predominant in males (YH) and at very low level in females (NA12878).

Conclusions

Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. Genome mapping in NanoChannel arrays is shown to be a rapid, accurate, powerful and robust method for detection of structural variation and the study of complex regions in the human genome.

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:
4) Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
5) http://dgv.tcag.ca/dgv/app/home