Comprehensive Analysis of Large Structural Variants in Well-Characterized Human Genomes

ET Lam[1], AH Hastie[1], ACY Mak[2], YYY Lai[2], HZ Cao[3], DD Cao[3], W Andrews[1], H Dai[1], M Austin[1], F Trintchouk[1], M Saghbini[1], T Anantharaman[1], K Haden[1], X Xu[3], P-Y.K. Kwok[2], H Cao[1]

[1]BioNano Genomics, San Diego, CA, USA; [2]University of California, San Francisco, San Francisco, CA, USA; [3] BGI-Shenzhen, China

Abstract

Genome mapping in nanochannel arrays (BioNano Genomics) represents a new single-molecule platform complementary to short-read sequencing for genome assembly and structural variation analysis. Extremely long molecules of hundreds of kilobases fluorescently labeled at sequence motifs and elongated in nanochannels enable direct interrogation of genome structure at a high resolution.

The high throughput of the BioNano Irys System has made possible, for the first time, rapid analysis of multiple genomes and cross-sample comparison to identify genome structural variation at high resolution. To date, we have *de novo* assembled more than 20 normal and diseased human genomes and analyzed their structural variation content. Our genome map assemblies cover at least 90% of non-N-base portions of the genome and also extend into subcentromeric and subtelomeric regions of the genome.

Here, we present results from extensive analysis of an Asian genome and a CEPH trio. We detected hundreds of large structural variants per genome and haplotype differences in these genomes. In the YH genome, we found 708 insertions/deletions and 17 inversions larger than 1 kb. Without considering 59 SVs that overlap with N-base gaps in hg19, 609 out of 666 (90%) are supported by orthogonal experimental methods (resequencing- and/or fosmid assembly-based) or historical evidence in public databases. For the CEPH trio, we identified novel and previously reported structural variants consistent with Mendelian inheritance. We also used publically available sequence read data to confirm and refine our SV calls.

Overall, our genome map assemblies provide valuable structural information otherwise difficult or impossible to decipher with short-read sequencing data alone.

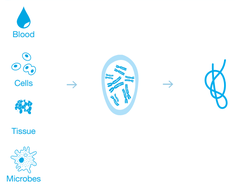
Background

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.

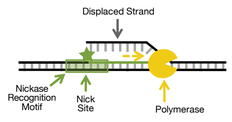
Methods

(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.

1) IrysPrep Kit extraction of long DNA molecules



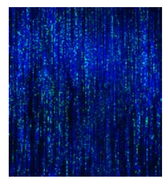
2) IrysPrep reagents label DNA at specific sequence motifs



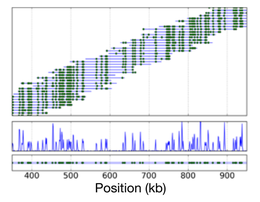
3) IrysChip linearizes DNA in NanoChannels

4) Irys automates imaging of single molecules in NanoChannels

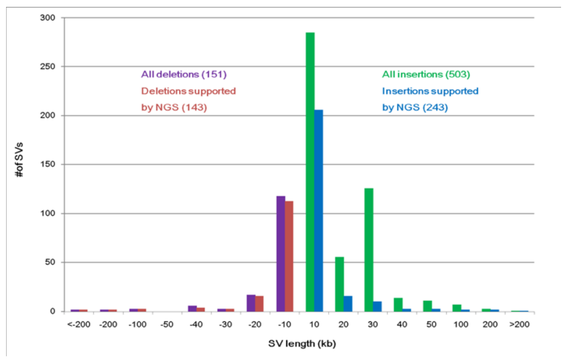
5) Molecules and labels detected in images by instrument software

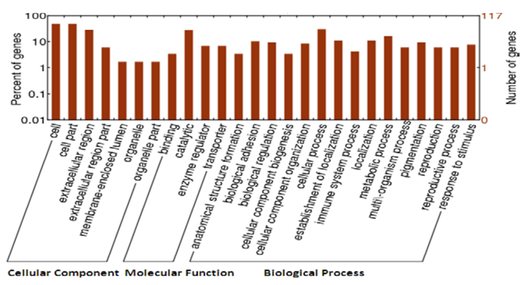


6) IrysView software assembles genome maps



Structural Variation Detection in YH



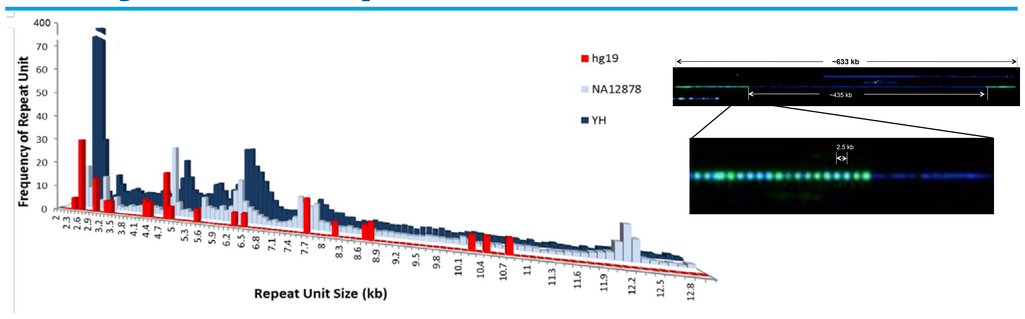


Sizes of detected large insertions (green) and deletions (purple) using genome mapping are shown. The comparative histogram bars in red and blue respectively represent deletions and insertions supported by NGS data.

Using genome mapping, we obtain 708 insertions/deletions and 17 inversions larger than 1 kb. Without considering 59 SVs (54 insertions/deletions, 5 inversions) that overlap with N-base gaps in the reference assembly hg19, 396 out of 666 (60%) are verified by paired-end data from WGS based re-sequencing or *de novo* assembly sequence from fosmid data. In the remaining 270 SVs, 260 are insertions and 213 overlap known SVs in the DGV database. Overall, 609 out of 666 (90%) are supported by experimental orthogonal methods or historical evidence in public databases.

GO annotations suggest that a number of SVs are associated with genes important for cellular functions; these SVs may have important functional consequences.

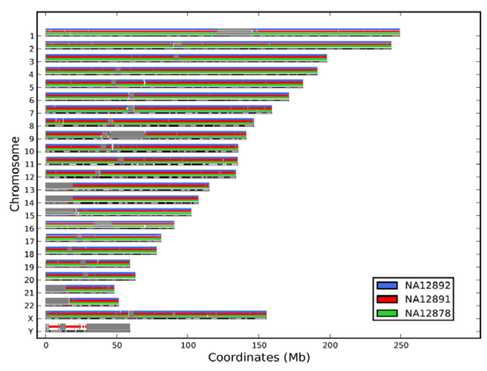
Analysis of Repetitive Elements in YH



Highly repetitive regions of the human genome are known to be nearly intractable by NGS. We found repeat units across the entire spectrum of sizes in YH and NA12878 while there were only sporadic peaks in hg19. Furthermore, we have found a very large peak of approximately 2.5 kb repeats in YH (male, 691 copies) but not in NA19878 (female, 36 copies). As an example, we show a raw image of an intact long molecule of 630 kb with two tracts of at least 53 copies and at least 21 copies of 2.5 kb tandem repeats (each 2.5 kb unit has one nick label site, creating the evenly spaced pattern) physically linked by another label-absent putative tandem repeat spanning over 435 kb. Unambiguously elucidating the absolute value and architecture of such complex repeat regions is not possible with other short fragment or hybridization-based methods.

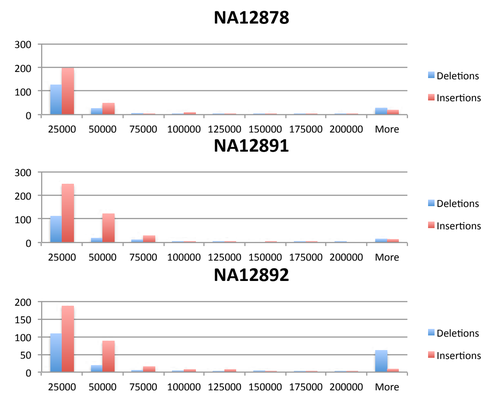
Analysis of Trio Samples

Genome Map Coverage of Reference



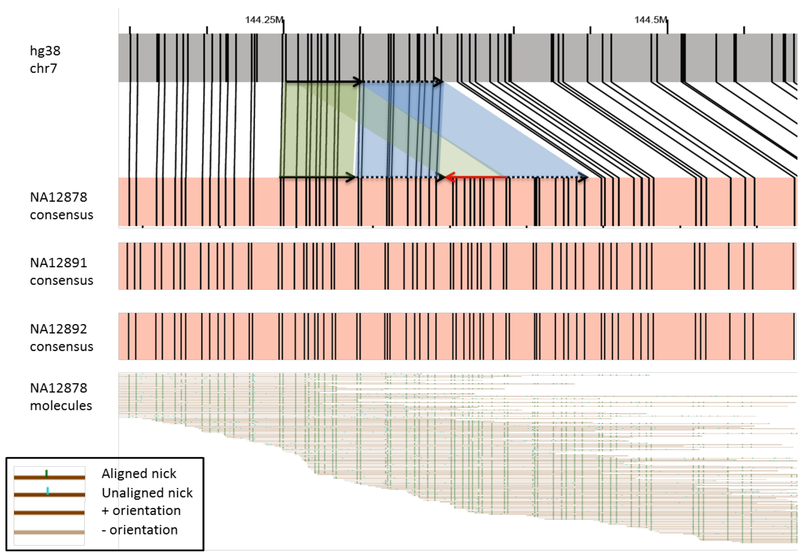
Using single molecules more than 250 kb on average, we generated *de novo* assemblies with N50s ranging from 4.6 to 5.0 Mb. 99% of the genome maps can be aligned to the reference, and 96% of the reference is covered by genome maps.

Structural Variation Detection



SVs across different size ranges were identified in all three trio samples. Their size distributions were similar, and in all three samples, there were more insertions and deletions. Over 70% of 1000 Genomes insertions and deletions reported in DGV were identified. Most of the genome mapping calls were novel and manually verified using single molecule maps.

Structural Variation Detection



Complex structural variation detected at chr7:144.25M where an inversion was previously reported. Our consensus maps further show that more complex structural variation events, including both duplication and inversion, have occurred at this locus.

Conclusions

Single-molecule genome mapping allows for complete characterization of a genome. *De novo* assembly of genome maps is performed without the use of any reference. It enables detection of large SVs difficult for NGS technologies. We have taken advantage of well-characterized genomes to validate SV detection using genome mapping data. Genome mapping of a human individual can currently be accomplished today with three chips or less in a few days. The throughput will improve such that a single chip would generate enough data for analysis of a human genome in less than one day, making population-based studies of structural variation possible.

References

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