Fast and Comprehensive Analysis of Large Structural Variants in Human

E. Lam, A. Hastie, W. Andrews, H. Dai, M. Austin, F. Trintchouk, M. Saghbini, T. Anantharaman, K. Haden, H. Cao
BioNano Genomics, San Diego, California, USA

Abstract

Irys genome mapping represents a recent 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 specific sequence motifs span across and enable direct interrogation of structural variants.

The short turnaround time to comprehensively analyze a human genome has given us the ability to rapidly analyze multiple genomes and perform cross-sample comparison to identify variation. To date, we have *de novo* assembled more than 20 normal and diseased human genomes and analyzed their structural variation contents. Genome map assemblies cover the majority of non-N base portions of the genome but also extend into subcentromeric and subtelomeric regions of the genome. We have expanded our analysis pipeline to include detection and validation of inversions and translocations in addition to deletions and insertions. We detected hundreds of large structural variants per genome and haplotype differences. Furthermore, we constructed highly accurate copy number profiles that are free from amplification bias and are particularly informative for analysis of cancer genomes.

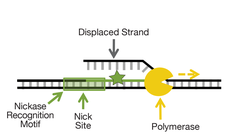
Background

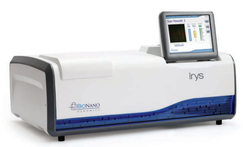
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing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

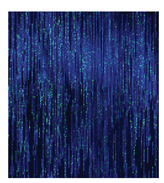
Methods

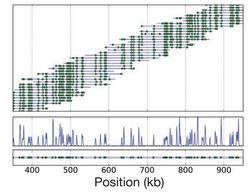
(1) DNA is labeled with IrysPrep™ reagents by incorporation of fluorophore-labeled nucleotides at a specific sequence motif throughout the genome. (2) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules are imaged by Irys. (3) Irys performs automated data collection and image processing. (4) Molecules are labeled with a unique signature pattern that is uniquely identifiable. (5) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite. (6) Downstream analyses include detection of structural variation.

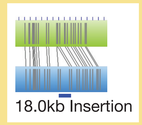
1) IrysPrep reagents label DNA at specific sequence motifs → 2) IrysChip linearizes DNA in NanoChannels → 3) Irys automates imaging of single molecules in NanoChannels → 4) Molecules and labels detected in images by instrument software → 5) IrysView software assembles genome maps → 6) Automated Structural Variation Detection Algorithm





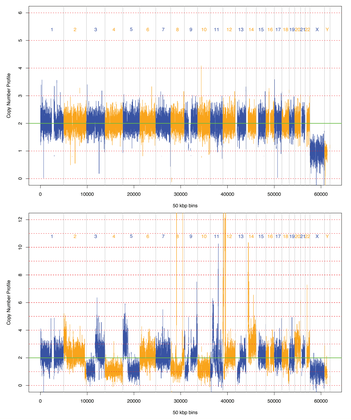






To align genome maps to each other or to a reference, a dynamic programming approach similar to classic Smith-Waterman is employed. However, instead of the units of comparison being single base-pairs, they are instead intervals between neighboring motif sites. Structural variations are called at regions that are discordant, having relatively low similarity. To gain confidence in such calls, it is required that the structural variations be flanked on both sides by high scoring regions.

Normalized Copy Number Profiles

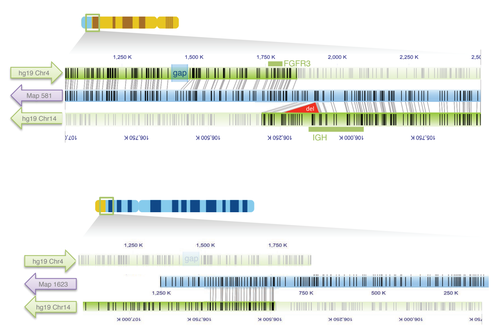


Copy number per bin was derived based on single molecule alignment to reference. Copy number profiles can reveal large-scale aberrations, localize breakpoints of large SV events, guide SV detection to regions of interest, and detect aneuploidy.

Single molecules were aligned to hg19. The genome was partitioned into 50 kb bins, and counts of molecules aligning to each bin were normalized and plotted.

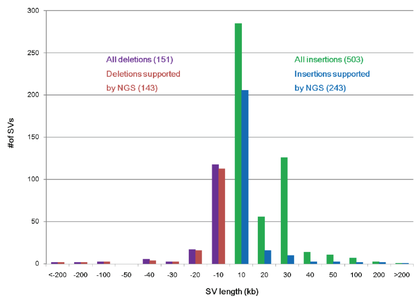
As shown, the copy number centers at around 2 throughout the genome in the diploid control sample. Since it was male, there is half the coverage of chrY. In the cancer sample, the copy number fluctuates dramatically compared to the control sample.

Detection of Balanced Translocations



Genome mapping in a cancer sample reveals a t(4;14)(p16.3;q32.3) reciprocal translocation event that results in a FGFR3/IGH fusion and MMSET dysregulation. This is a known translocation, important for cancer prognosis. FGFR3 and IGH are juxtaposed in the rearranged genome. We also see a deletion in IGH. This could be disease-related or simply an SV in IGH, which is highly variable in humans.

Complete Characterization of an Asian Individual



Utilizing the genome mapping technology, 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 hg19, 396 out of 666 (60%) are verified by paired-end data from re-sequencing or *de novo* assembly sequence of fosmids. Of the remaining 270 SVs, 260 of them 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.

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 not only observe insertions and deletions, but also inversions and (intra- and inter-chromosomal) translocations. 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) Nat Biotechnol. 2012 Aug;30(8):771-6. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY.