Rapid Genome-Wide Mapping at Single Molecule Level Using NanoChannel Array for Structural Variation and *De Novo* Assembly

H.C. Cao[1], A.H. Hastie[1], E.L. Lam[1], H.D. Dai[1], T.A. Anantharaman[1], M Rossi[2], M.X. Xiao[3], P-Y.K. Kwok[4]
[1]BioNano Genomics, San Diego, California, USA
[2]Emory University, Atlanta, GA; [3]Drexel University, Philadelphia, PA; [4]UCSF, San Francisco, CA

Abstract

Despite continued cost reduction in raw base generation, improvement in base-calling accuracy, and recent advances in read length, complete *de novo* assembly with accurate genome wide structural variant (SV) analysis of an individual large complex genome remains expensive and challenging. In particular, many disease relevant SVs up to hundreds of kilobases long in the human genome are severely underestimated due to a lack of effective tools.

We present a rapid genome-wide analysis method based on new NanoChannel Array technology (Irys™ System) that temporarily confines and linearizes extremely long DNA molecules for direct image analysis at tens of gigabases per run. This high-throughput platform automates the imaging of genomic DNA hundreds to thousands of kilobases in length at single-molecule level, retaining long-range haplotypes. High-resolution genome maps assembled *de novo* via unique sequence motif labeling preserves native large and small structural variation information (especially highly repetitive regions), which are intractable with current short read NGS platforms. This information is collected independently of sequencing methods and is very valuable to identify structural variants as well as to validate and finish sequencing assemblies.

Here we report the complete *de novo* assembly and analysis of complex regions and whole genomes of several human samples (including a cancer genome) with this approach. Unlike inference from mate-pair library sequencing approaches, hundreds of large structural variants were uncovered without apparent bias (e.g., size or insertion vs deletion) due to its more direct visualization and measurement. We have corrected errors in previous assemblies, spanned and sized many of the remaining gaps, identified known and novel structural variants and phased haplotype blocks, including in the highly variable complex regions related to human immune system functions. We have also discovered abundant previously unknown highly complex large repetitive patterns (greater than 2kb and inverted) spanning large regions of genome and pinpointed foreign genomic components inserted within the host human genome, important for understanding disease and oncogenesis.

Widespread use of this technology will continue to enable new genome discoveries, expand our view of genome architecture, and improve understanding of functional regions.

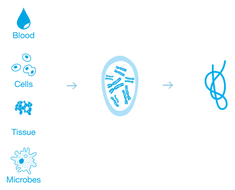
Background

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. 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.

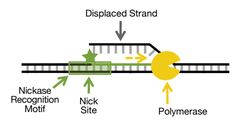
Methods

(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

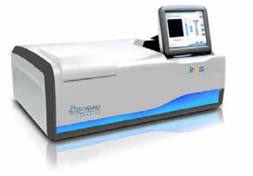


2) IrysPrep reagents label DNA at specific sequence motifs

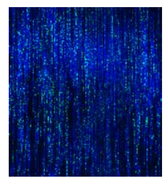


3) IrysChip linearizes DNA in NanoChannels

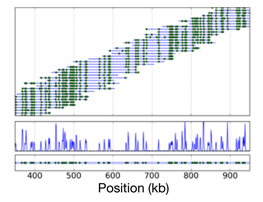
4) Irys automates imaging of single molecules in NanoChannels



5) Molecules and labels detected in images by instrument software

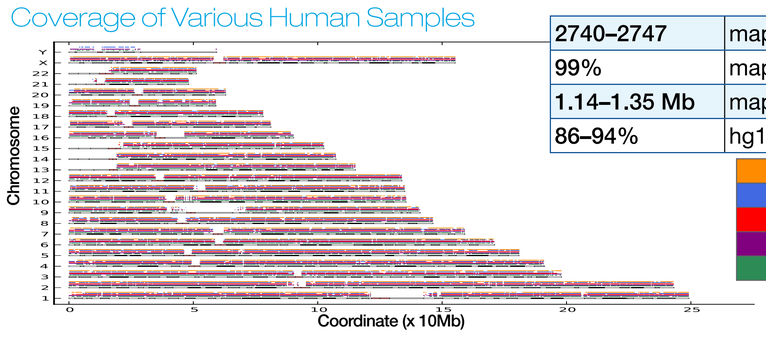


6) IrysView software assembles genome maps



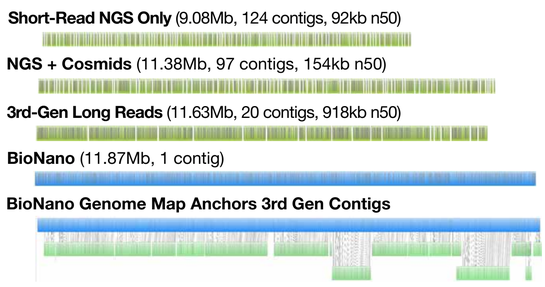
Human *De Novo* Assemblies

Coverage of Various Human Samples



2740–2747	maps per genome
99%	map to hg19
1.14–1.35 Mb	map N50
86–94%	hg19 overlap

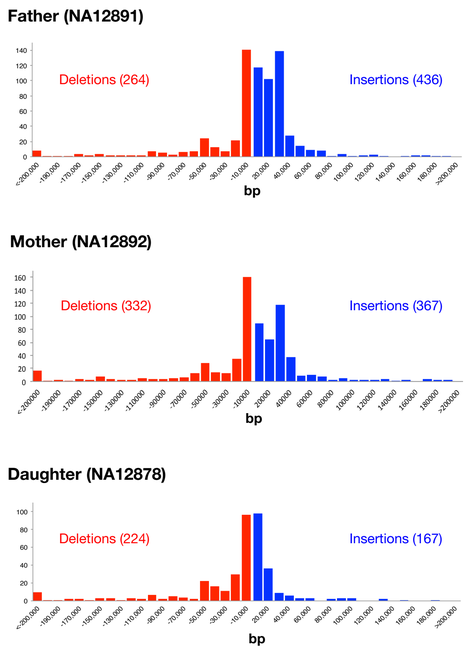
Genome Maps Are More Complete



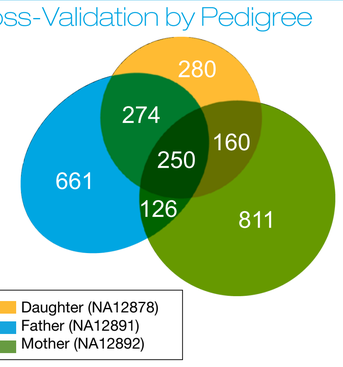
A Streptomyces genome was sequenced and assembled by a combination of different sequencing platforms (green maps). In contrast to these fragmented assemblies, one intact contiguous genome map (blue) was assembled *de novo* by the Irys system. The genome map anchored 19 of the 20 3rd-Gen Long Reads contigs.

Genome-Wide Structural Variation

Comparison of SV Size Distribution

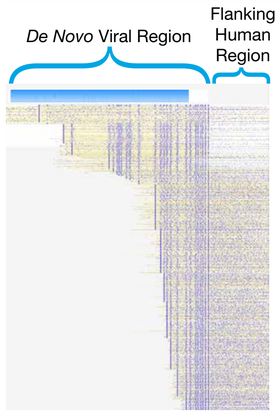


Cross-Validation by Pedigree



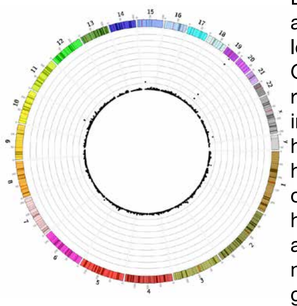
Structural variation across a broad range of sizes refractory to many high throughput and short-read technologies was detected. Insertions are called by the presence of novel label sites and expansion of adjacent labels. Deletions are evident by the absence of label sites or narrowing of inter-label segments. Maps also identify variation in difficult-to-sequence highly repetitive regions, such as those involved in immune function (such as MHC) and near the telomeres and centromeres. Concordance between the trio supports the validity of the detected SVs.

Viral Insertion Site Identification



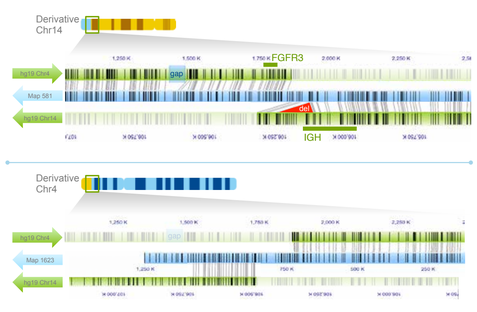
Long molecules preserve physical links between integrated viral components and host genome. The flanking regions physically adjacent to virus insertion sites were mapped throughout the human genome, with various copy numbers.

Genome-Wide Insertion Sites



Black dots represent depth and location of insertion location in host genome. Grey concentric ring each represents 2X coverage increments. For example high virus load (1865X) vs human depth (45X) was detected. This result would help to understand the roles and functions of genomic mobile elements in evolution, genome instability, cancer, and improve gene therapy and transgenic engineering.

Balanced Transolcation in MML



Translocations can manifest as aberrant expression of existing genes, nascent expression of resulting new fusion genes, and potential oncogenic activity. These are difficult to detect with current methods. For example, the t(4;14)(p16.3;q32.3) is found in 15% of multiple myeloma (MM) cases and is associated with poor prognosis. Irys Genome Map detected and confirmed precisely the t(4;14)(p16.3;q32.3) reciprocal translocation event in KMS11 cells. Translocation results in FGFR3/IGH fusion and MMSET dysregulation; it is difficult to detect by NGS due to telomeric position of IgH locus.

Conclusions

BioNano Genomics Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. This system permits rapid accurate genome-wide *de novo* assembly and detection of structural variants that typically confound short-read genome assembly and comparative genomic analysis. Here we demonstrate *de novo* human Genome Map assembly capabilities of the IrysChip nanochannel arrays and the Irys imaging system to characterize genome-wide structural variation in the human genome. By comparing *de novo* assemblies of a trio family, we show that genome mapping is able to detect large structural variants with very good cross-validation. We are also able to map regions of the genome that are refractory to assembly by other methods, such as reciprocal translocations in cancer and foreign DNA insertions.

References

1) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
2) 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
3) Xiao, M et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.