

# Mapping and *De Novo* Assembly of Long Barcoded Molecule of DNA from a Cancer Genome



**J. Reifenger, Z. Dzakula, V. Dergachev, T. Anantharaman, A. Hastie, S. Chan, H. Cao**  
BioNano Genomics, San Diego, CA, USA

**Abstract** We present a rapid genome-wide analysis method based on new NanoChannel Array technology (Irys® System) that confines and linearizes individual molecules of DNA to nearly 85% of their respective contour length. Once confined, extremely long DNA molecules (100 to 1,000 kilobases) are imaged with a custom three-color epifluorescence microscope. Genomic DNA is stained with YOYO and can be labeled specifically at the 'GCTCTTC' or 'CCTCAGC' sequence with different colored fluorescent probes allowing for each molecule to have a unique pattern and mapped to its corresponding location in a reference. By imaging hundreds of gigabases per run, a high-resolution genome map was assembled *de novo* from the extremely long single molecules, which retain the original context and architecture of the genome. This allows for the cataloging of complex repetitive regions and large-scale rearrangements. The advantages of genome mapping for rearrangement analysis are demonstrated here by several examples in which we identify hallmark fusions as well as rearrangements from cancer cells.

## Background and Methods

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

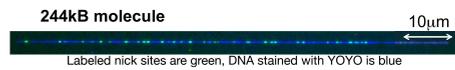
(1) Long molecules of DNA is labeled with IrysPrep reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® nanochannels and single molecules are imaged by Irys. (4) Single molecule 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.



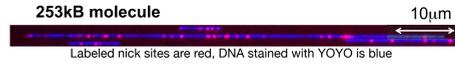
### Individual Molecules of Long DNA Provide a Unique Barcode

Once DNA is nicked, the nick sites can be extended through incorporation of modified nucleotides with either dyes similar to cy3, cy5, or both. DNA is always stained with YOYO. Nick sites provide a unique barcode pattern that relates to the recognition sequence of the respective nicking enzyme.

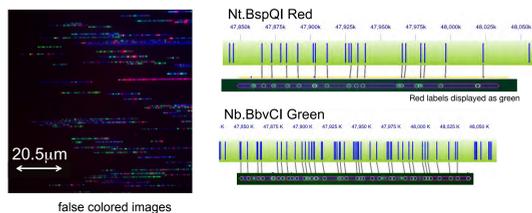
Example of DNA nicked with Nt.BspQI (GCTCTTC) and labeled with dye similar to cy3



Example of DNA nicked with Nt.BspQI (GCTCTTC) and labeled with dye similar to cy5



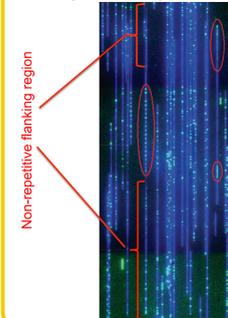
Example of dual nicked DNA. Red labels from Nt.BspQI (GCTCTTC) and green labels from Nb.BbvCI (CCTCAGC).



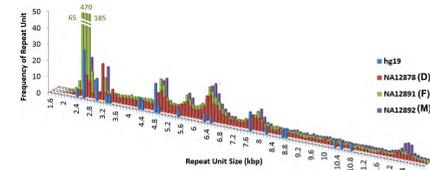
### Long Repetitive Regions are Easily Identified

Long Molecules often capture an array of repeat segments that current sequencing technologies struggle to categorize. Molecules that span the repeat or extend past it on either side can be mapped to a reference or even assembled. Irys technology allows for the detection of both the copy number and the locations of these regions in the genome.

Example of Repeats detected on DNA



#### Prevalence of Repeats in Human Datasets



Comparison of frequency of repeat motifs found in hg19 reference and trio assemblies (normalized to 1X coverage). Only repeats with 5 or more tandem repeat units were considered. Prominent peak at 2.5kb is seen in NA12891 (father) and has been found in other male samples.

### De Novo Assembly of Cancer Genome Detects Mutation Events

A cancer cell line used to model Myeloma was mapped and assembled

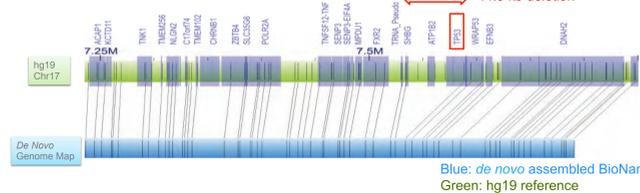
#### Raw Molecule Stats

Molecule n50	Values
Molecule n50	177kB
Total Number of Molecules	709,930
Total Size of Molecules	158 GB (49x)

#### Assembly stats

Total Assembled Contig Length	Values
Total Assembled Contig Length	2.63 Gb
Total Number Contigs	3,227
Contig n50	1.014 Mb
% hg19 Overlap	86%

Once assembled, a large deletion was detected at the P53 Locus which removes one copy of P53, a common tumor suppressor

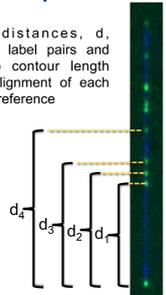


### Confinement Results in Uniform Stretching DNA in Nanochannels

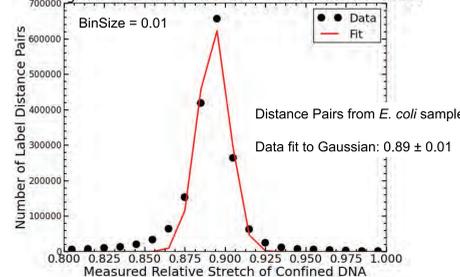
Once labeled and stained, DNA is loaded into nanochannels through electrophoresis. By measuring the distance pairs between all labels on a molecule aligned to its respective genome, a histogram of relative stretch (measured length to theoretical contour length) can be generated. Individual DNA molecules are typically stretched to > 85% of their respective contour length.

#### Example of Molecule

Measure distances,  $d_i$ , between all label pairs and compare to contour length based on alignment of each molecule to reference



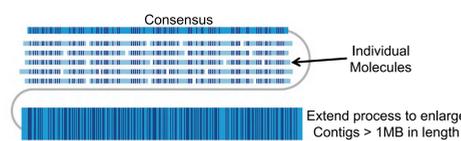
#### Histogram of Relative Stretch of DNA in Nanochannels



### De Novo Assemble by Consensus of Individual DNA Barcodes

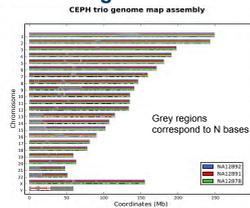
Once enough individual molecules are imaged (> 30x), a *de novo* assembly can be generated through matching individual molecules and determining the consensus barcode pattern

Algorithms Developed to 'stack' molecules with similar barcode patterns in order to generate a consensus *without a reference*



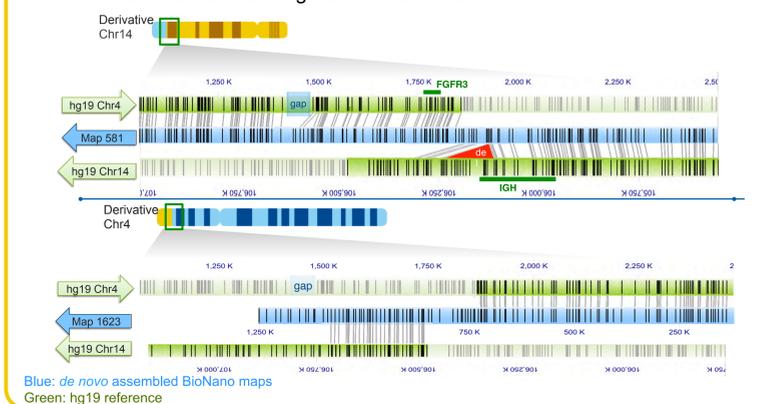
#### Examples of *de novo* Assemblies on 3 human genomes

	#Maps	#Mapped maps*	Map N50 (Mb)	hg19 coverage
NA12878 (D)	2,791	2,769 (99%)	1.23	92%
NA12891 (F)	2,716	2,689 (99%)	1.35	94%
NA12892 (M)	2,747	2,728 (99%)	1.22	89%



### Long Assembly Detects Translocation Event

Assembled contigs (map 581 and map 1623) from Myeloma sample also found a balanced translocation causing FGFR3-IGH Fusion Event



## Conclusions

Nanochannels allow for long DNA molecules to be consistently stretched to over 85% of their contour length. Once stretched fluorescent probes at sequence specific locations allow detection of long repeat regions as well as the mapping and *de novo* assembly of a reference that can highlight large scale rearrangements commonly found in most cancers.

## References

- Cao, H., et al. Rapid Detection of Structural Variation in a Human Genome using Nanochannel-based Genome Mapping Technology. *Giga Science* (2014); 3(December 2014): 34
- O'Bleness, M., et al. Finished Sequence and Assembly of the DUF1220-rich 1q21 Region Using a Haploid Human Genome. *BioMed Central* (2014); 15(July 2014): 387
- 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
- Lam, E.T., et al. Genome mapping on Nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology* (2012); 30(8):771