Leveraging the Flexibility of Multi-color Imaging of Extremely Long Single-Molecules in NanoChannels for Epigenetic Profile Mapping, Centromere Probing by Hybridization and Sample Multiplexing

Authors: A. Hastie[1], DH. Zhang[1], P. Sheth[1], K. Pham[1], J. Reifenberger[1], X. Zhou[1], G. Pfjevaljcic[1], C. Escudé[2], S. Chan[1], E. Lam[1], H. Cao[1]

[1]BioNano Genomics, San Diego, CA, 92121, USA; [2]Muséum National d'Histoire Naturelle, Paris, France

Abstract

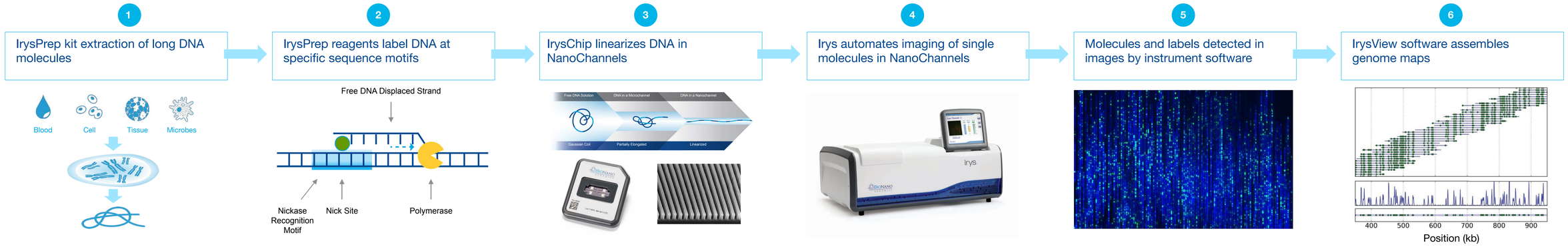
Genome mapping by single molecule imaging of fluorescently labeled motif sites on uniformly elongated long DNA molecules in Nano Channel arrays has been used extensively in the last couple years for improving *de novo* sequence assemblies through scaffolding sequence contigs and for detections of germ line and somatic structural variation (SV). Several improvements have been recently developed to enhance the throughput and richness of genome mapping data. Here we provide data on multiplexing samples for genome mapping, probing high molecular weight DNA with triplex forming oligonucleotides for mapping and assembly of centromeres as well as for mapping of CpG methylation status in conjunction with genome mapping.

As throughput has improved for single molecule data collection in nanoarrays, a single flow cell is now able to collect far more data than needed for smaller genomes resulting in a need for multiplexing. By coding samples with different fluorophore combinations, it is possible to mix several samples together in a single flow cell for data collection and demultiplex the samples bioinformatically for downstream analysis. Genome mapping is primarily done by creating single strand nicks using modified restriction enzymes. In order to target other regions of interest that may not have enough information by nick labeling, a hybridization approach was undertaken. This has been applied to the centromere of chromosome 17. From this, the pattern of centromere higher order repeats in single molecules are directly observed and measured. Finally, a new method for CpG methylation status will be presented. Single molecules are labeled at nicking motif sites and, using a second color, at a subset of CpG sites, the unmethylated sites are labeled. This method allows visualization of methylation status in phase with other genomic positions including SVs. These powerful new methods leverage the flexibility of the Irys® System for single molecule linearization and imaging for improved throughput and deeper biological investigation.

Background

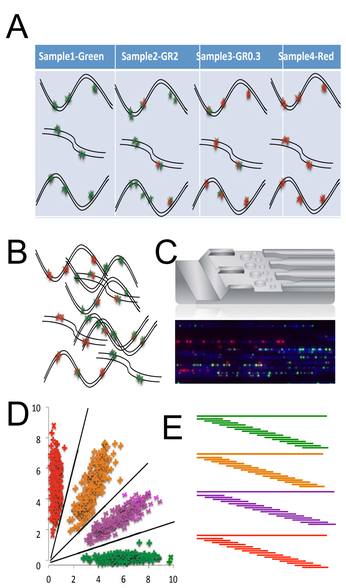
Irys has been commercialized for whole genome mapping by *de novo* assembly using very long single molecule reads. The platform has been widely adopted for *de novo* sequence assembly and structural variation analysis but is also very flexible for additional assay development. Here we show data on new applications: epigenetic profiling, assembly of centromere DNA by hybridization labeling and multiplexing of multiple samples. (1) The throughput of the Irys System has sufficiently surpassed the need for coverage depth for many smaller genomes so we have developed a multiplexing method to better exploit the high throughput. (2) Centromeres have thus far been intractable for assembly; therefore, we developed a labeling method that allows for assembly of centromeres by genome mapping. (3) Human epigenetic regulation is coordinated, to a large extent, by methylation of CpG sequences at the cytosine site. These modifications can be interrogated by several methods such as bisulfite sequencing and other NGS methods but none are compatible with SV analysis, potentially missing important regulatory information.

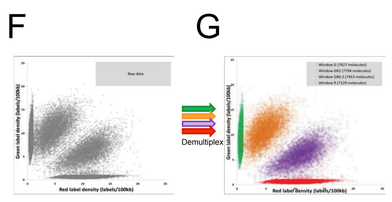
Methods



(1) Long molecules of DNA are 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.

Spectral Multiplexing of Four Samples per Irys Flowcell



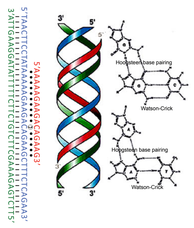


H

	BL21	DH10b	MR-1	SalmTyphiL	Total purity
Window G	16	41	***3959***	18	99.96%
Window GR2	190	***5882***	153	9	94.2%
Window GR0.3	221	22	4	***5015***	95.1%
Window R	***5022***	263	2	23	94.3%
Not used	89	115	185	55	

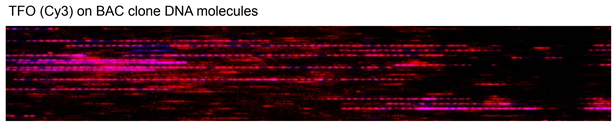
The strategy for spectral multiplexing: (A) Label four DNA samples with different mixtures of green and red fluorophores, mix the molecules together (B) Collect data on Irys (C) Different samples will have different ratios of red to green label density (D), which allows for demultiplexing and *de novo* assembly of each genome separately (E). X-Y scatter plot of four genomes labeled with multiplex label mixes and run in one Irys flowcell (F). Demultiplexed DNAs are color-coded for each pool. Mapping of demultiplexed molecules to original genomes shows purity of at least 94%.

Centromere Labeling by Triplex-forming Oligo (TFO) Hybridization

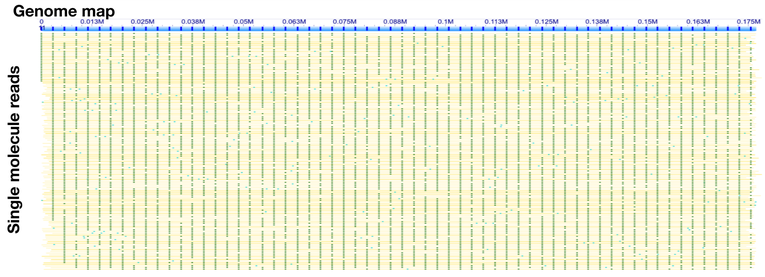


The triplex forming oligo binds dsDNA by Hoogsteen base pairing at stretches of purine bases. We have used a probe that targets a segment of D17Z1 with high efficiency and tested this on a BAC clone (RP11-352P13, images below). These molecules could be used to assemble a sequence motif map for the centromere segment. In this clone, the target sequence occurs every 2.8 kb at uniform intervals. This is consistent with individual higher order repeat models for D17Z1. This labeling can be combined with standard nick labeling for human *de novo* assemblies.

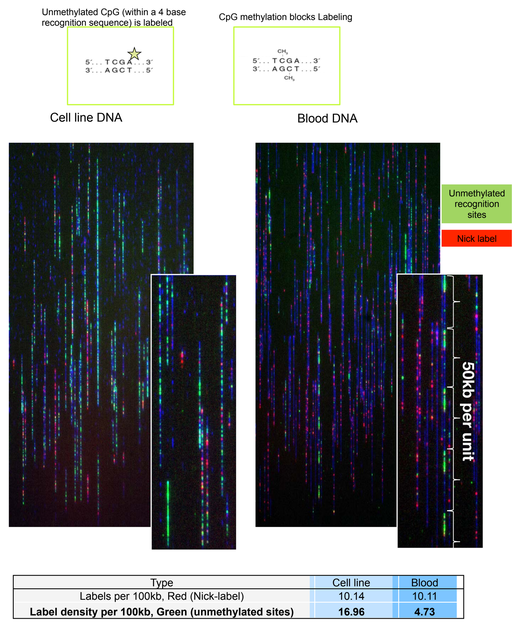
TFO (Cy3) on BAC clone DNA molecules



Genome map



Epigenetic-state Detection Assay



Type	Cell line	Blood
Labels per 100kb, Red (Nick-label)	10.14	10.11
Label density per 100kb, Green (unmethylated sites)	**16.96**	**4.73**

The methyltransferase M. TaqI is able to transfer a labeled cofactor to sequence motifs but is blocked by overlapping methyl-CpG. We have taken advantage of this property to differentially label TCGA sites which are unmethylated while methylated sites remain blank. In the images, we can see that there is significantly more unmethylated TCGA sequences in the cell line DNA than in the human blood DNA. From the cell line DNA, ~17 sites/100 kb could be resolved while in the blood, only ~5 sites/100 kb were detected. Interestingly, a long repeated motif was highly labeled in the blood cells. This tandem repeat would be difficult to assay using short-read methylation detection sequencing methods.

Conclusions:

- It is now possible to efficiently multiplex and demultiplex several genomes in a single flow cell of an IrysChip using fluorophore mixture.
- TFO-based labeling provides a flexible method to label genomic DNA for targeted interrogation by Irys.
- Epigenetic status can now be measured on the Irys System and observed in conjunction with structural variation analysis.

Reference:

1) Cao, H., *et* al., Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience (2014); 3(1):34
2) 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.
3) Lam, E.T., *et* al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
4) Bénédicte, G-L., *et* al. Sequence-specific fluorescent labeling of double-stranded DNA observed at the single molecule level. Nucleic Acids Research (2003); 31(20): e125
5) Dalhoff, C., *et* al., Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. Nature Chemical Biology (2006); 2(1): 31-32
6) Xiao, M., *et*. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.