Barcode labeling of short reads for detection of large scale genomic variations

Tao Zhang [1], Kevin Eng[1], Jeff Froula [1], Kanwar Singh [1], Matt Blow [1], Zhong Wang [1], Len Pennacchio [1], Feng Chen [1]

1 Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA

[a]*To whom correspondence may be addressed. E-mail:* tzhang@lbl.gov

March 19, 2012

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California

Barcode labeling of short reads for detection of large scale genomic variations

Tao Zhang, Kevin Eng, Jeff Froula, Kanwar Singh, Matt Blow, Zhong Wang, Len Pennacchio and Feng Chen

DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598

Introduction

Genomic variations, such as insertion, deletion, duplication, inversion and translocation, are common in human population and frequently found in cancer. Identification of genomic variations is important for understanding difference between individuals and diagnosis of diseases. Finding sequence breakpoints is the first step in genetic variation analysis. It can be achieved by making mate pair library and performing optical mapping of DNA through restriction digestion pattern analysis. However, these approaches are either labor intensive or costly. We have developed a method using barcodes to label single large DNA molecules to 40 kb through MDA. One million large DNA molecules can be labeled simultaneously in micro-droplets. By barcode labeling, short sequences derived from same original templates will contain same barcodes. Through barcode analysis, reads are binned according to their barcodes. By mapping barcoded reads into reference genome, breaking points can be detected. Our results suggested that barcode labeling of small genomic DNA fragments can be used to increase footprints of short sequence reads for detection of large genetic variations in complex genome.

Strategy for barcode labeling

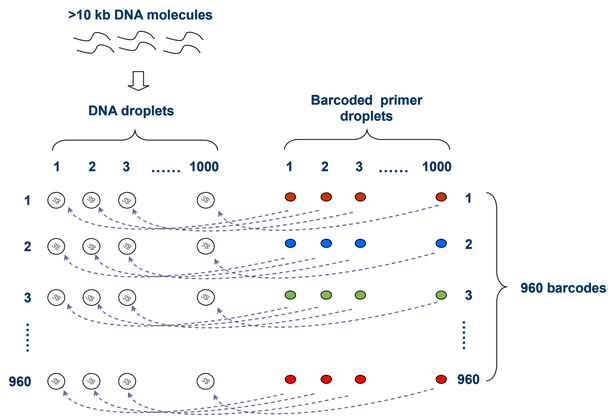

Figure 1. Large DNA molecules are diluted and distributed into micro-droplets, one molecule per compartment. Micro-droplets containing barcoded primers are merged with DNA template droplets. MDA enzymes and reagent are also included in the merged droplets.

Droplets merge in micro-fluidic chip

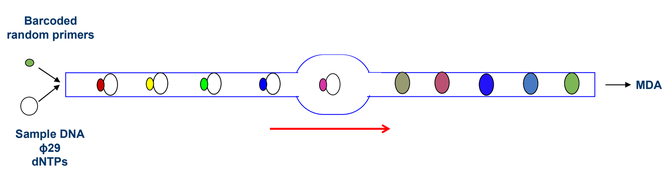

Figure 2. Fusion of DNA template droplets with barcoded primer droplets delivers primer to DNA template. In each compartment, DNA molecules are amplified by MDA. Droplets are destabilized and DNA contents are pooled for making Illumina library and high throughput sequencing.

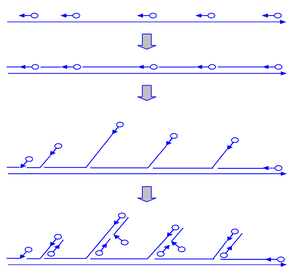

Figure 3. Barcodes are attached to random hexmers for MDA amplification of single DNA molecules which produces branched DNA fragments that containing multiple ends with barcode sequences. These ends can be captured by Illumina sequencing.

Barcoded reads can be grouped based on their barcode sequences

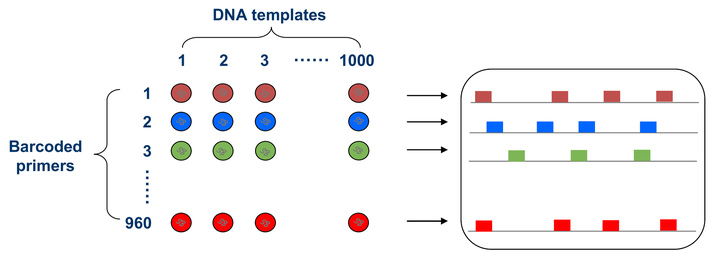

Figure 4. About 1 million of DNA molecules are labeled by 960 barcodes. Reads carrying same barcode may come from different template molecules. This is ~1000 fold reduction of complexity of the library. Mapping reads containing same barcode to reference genome is expected to form clusters randomly distributed alone the genome. Each cluster should corresponding to one original DNA template.

Mapping of barcoded reads formed clusters in reference genome

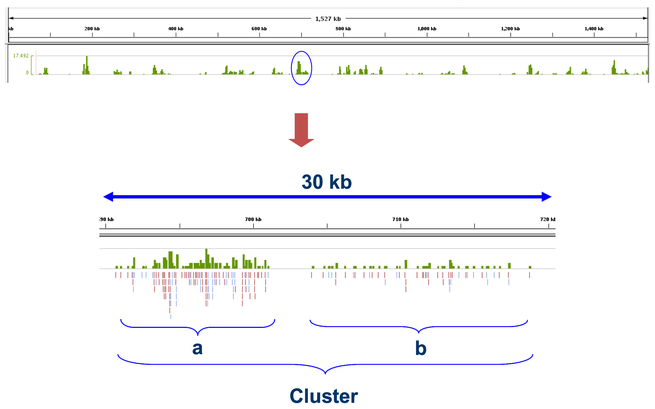

Figure 5. Reads containing one barcode were trimmed and mapped into reference genome. The mapped reads formed clusters randomly distributed along the reference genome. Each cluster reflected one original DNA template in one micro-droplet. Average size of clusters is 16kb (range from 5-40kb).

Distribution of mapped barcoded reads in reference genome

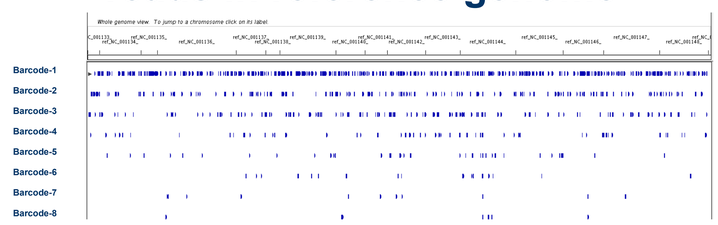

Figure 6. Reads containing different barcodes form multiple clusters some are overlapping to each other. The number of clusters are different depends on the type of barcode primers used in MDA.

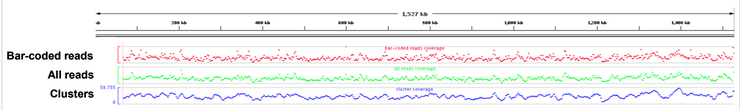

Figure 7. Even distribution of MDA amplified reads across the reference genome.

Detection of breaking points

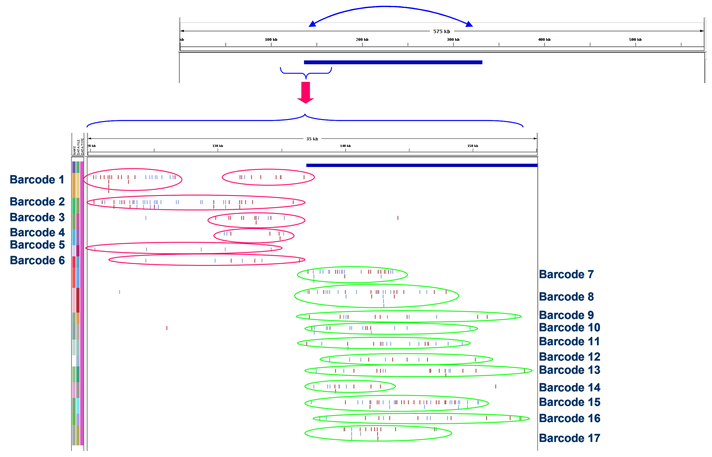

Figure 8. The boundary of clusters formed by barcoded reads is right at the sequence breakpoint.

Metagenome complexity reduction

Figure 9. Droplet based MDA could reduce the complexity of metagenome sample and generate high sequence coverage for assembly of large contigs.

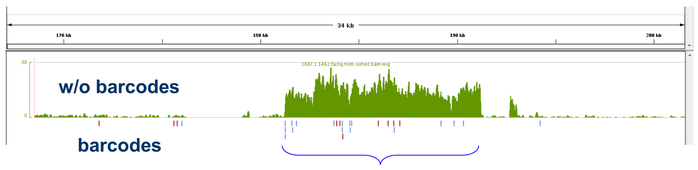

Figure 10. Barcoded reads can be used to evaluate quality of metagenome assemblies.

Summary of barcode labeling in genomic variation detection

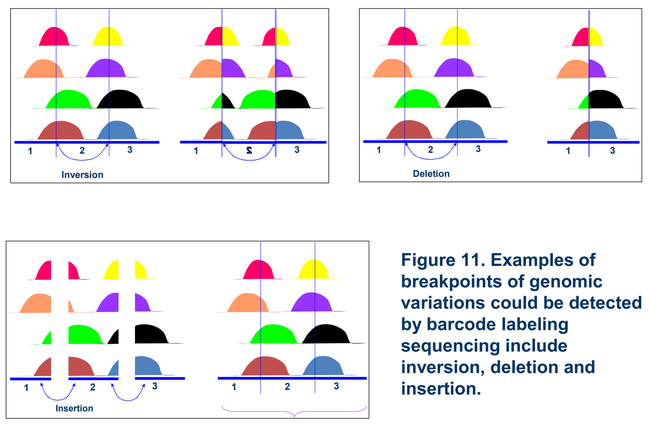

Figure 11. Examples of breakpoints of genomic variations could be detected by barcode labeling sequencing include inversion, deletion and insertion.

Acknowledgment: Darren Link (RainDance Technologies)