Lawrence Berkeley National Laboratory
Recent Work

Title
Fosmid Ditags as a New Technology Developed at JGI

Permalink
https://escholarship.org/uc/item/76p7r25f

Authors
Peng, Ze
Malinov, Ilya
Smith, Douglas
et al.

Publication Date
2007-03-28
Fosmid Ditags as a New Technology Developed at JGI

Ze Peng, Ilya Malinov, Doug Smith, Feng Chen, Paul Richardson, Len A. Pennacchio, and Jan-Fang Cheng

Introduction

Paired end reads from large insert DNA libraries are essential for detecting chromosome rearrangements as well as connecting sequence scaffolds of draft genomes. However, fosmid and BAC end sequencing remains challenging as well as expensive. Ditag sequencing of fosmid ends represents a cost effective way to generate paired end sequences from large genome fragments. We present results from several ditag libraries from human, fungi, and bacteria, which were sequenced using 454 technology. Several software tools were developed to analyze the resulted ditag sequences. These tools have been used to (1) create suffix array of the reference genomes; (2) filter, trim, and prepare the paired 18mer ditag sequences for analysis; (3) search for 18mer mismatches for matches; and (4) score the chromosome locations of ditag pairs.

The Foundation of Fosmid Tidag

1. MmeI is a type II restriction endonuclease, it cut 18 or 20 bp away from its recognition site. 2. Fosmid is a F-factor based, propagated phagemid vector system. The variations of insert size is small and only one or two copies in the host offers high stability.

Construction of Fosmid Ditag Vector

Fosmid ditag vector phiPfo5-DT1 was constructed by replacing the phiPfo5 vector’s BamHI-EcoRI-BamHI fragment with BamHI-Mmel-NdeI-Mmel-BamHI at the cloning site and eliminating the 4 existing Mmel sites in the vector.

Work Flow of Generating Fosmid Ditags Sequence

Application 1. Using Fosmid Ditag to Detect Chromosome Rearrangements of Cancer genome.

- Before the ditag analysis is run, a suffix array of the reference genome needs to be created. If there information about SNPs is available, the SNP-library is also created.
- The ditag pairs are extracted from the reads.
- Obtained ditag pairs are mapped to the reference genome using the suffix array of the genome and the SNP-library, if available.
- The mapped ditag pairs are analyzed and grouped:
  - unless both ditags in a pair are mapped, the pair is considered "not found"
  - if among all mapped locations for a pair there is such combination that both ditags are mapped to the same chromosome within the specified distance and are on the same strand, then such pair is considered "normal"
  - if both ditag in a pair is mapped more than once and the pair is not "normal", then it is considered to be a "repeat".
  - if both ditags in a pair are mapped only once and are mapped to different chromosomes, or mapped to the same chromosome, but the distance between them is outside of the specified boundaries, (e.g., 50Kb-50Kb), or they are mapped to opposite strands, then such pair is considered "flagged".
- "positive hits" means that if there are more than one pair of ditags mapped to the same location of different chromosomes (translocation) or mapped to the same chromosome, but the distance is outside the fosmid size range (deletion or insertion), or they are mapped to opposite strands (inversion).

Application 2. Using Fosmid Ditag to Help Bacteria Genome Assembling

We have successfully generated R.Met and R.Pal5 ditag sequence which over 77% or 75% of pairs agree with reference genome, only 2.2% or 1.9% of pairs disagree the reference genome. The screening pairs equal 16 and 10 fold genome clone coverage respectively. We also labeled that poor quality reads are not suitable for generating ditag sequence.

The Graph of Cupriavidus metallidurans CH34 (R.Met) ditag

- Rearranged chromosome
- Scaffolds

<table>
<thead>
<tr>
<th>Segment model</th>
<th># of Unique ditag pairs</th>
<th># of pairs agree with the reference genome</th>
<th># of pairs disagree with the reference genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhodopseudomonas palustris</td>
<td>20194</td>
<td>15074 (75%)</td>
<td>378 (1.9%)</td>
</tr>
<tr>
<td>BisB5 (R.Pal5)</td>
<td>23613</td>
<td>1069</td>
<td>12134</td>
</tr>
<tr>
<td>Dechloromonas aromatica RCB (D.Aro)</td>
<td>20194</td>
<td>15074 (75%)</td>
<td>378 (1.9%)</td>
</tr>
<tr>
<td>Dechloromonas aromatica RCB (D.Aro)</td>
<td>20194</td>
<td>15074 (75%)</td>
<td>378 (1.9%)</td>
</tr>
</tbody>
</table>

Summary:

The ditag technology in conjunction with the 454 sequencing provides a high throughput approach to assist short genome assemblies and characterize cancer genomes.

Reference:
