DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

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 genomic 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 arrays of the reference genomes; (2) filter, trim, and prepare the paired 18mer ditag sequences for analysis; (3) search for 18mer strings for matches; and (4) score the chromosome locations of ditag pairs.

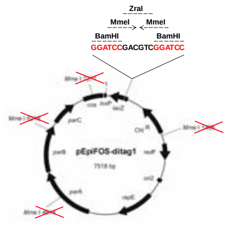
The Foundation of Fosmid Tidag

1. MmeI is a type II restriction endonuclease, it cut 18 or 20 bp away from its recognize sit: 5'...TCCRAC$(N)_{20}$...3'
3'...AGGYTG$(N)_{18}$...5'

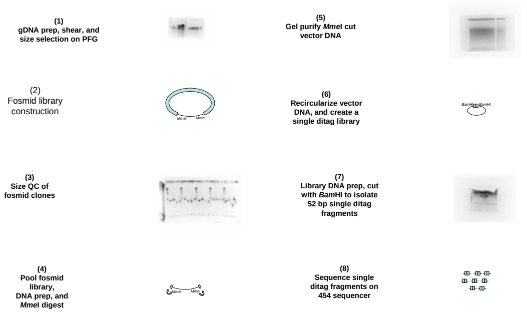
2. Fosmid is a F-factor based, propagated phagemid vector system. The variations of insert size is small and only one or two copies in the host offers high stability

Construction of Fosmid Ditag Vector

Fosmid ditag vector pEpiFos5-DT1 was constructed by replacing the pEpiFos5 vector's BamHI-Eco72I-BamHI fragment with BamHI-MmeI-ZraI-MmeI-BamHI at the cloning site and eliminating the 4 existing MmeI sites in the vector



Work Flow of Generating Fosmid Ditags Sequence


(1) gDNA prep, shear, and size selection on PFG

(2) Fosmid library construction

(3) Size QC of fosmid clones

(4) Pool fosmid library, DNA prep, and *Mme*I digest

(5) Gel purify *Mme*I cut vector DNA

(6) Recircularize vector DNA, and create a single ditag library

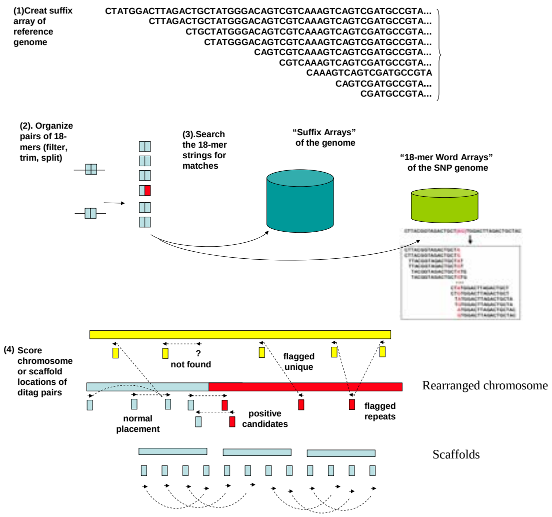
(7) Library DNA prep, cut with *Bam*HI to isolate 52 bp single ditag fragments

(8) Sequence single ditag fragments on 454 sequencer

Reference:

1. Ung-Jin Kim et al (1992) Stable propagation of cosmid sized human DNA insert in an F factor based vector. Nucleic Acids Research 20: 1083-1085
2. 1. U. Manber and G. Myers (1993) *Suffix arrays: A new method for on-line string searches.* SIAM Journal on Computing, 22:935--948

Analysis of Ditag Pairs for Chromosome or Scaffolds Locations


(1)Creat suffix array of reference genome

```
CTATGGACTTAGACTGCTATGGGACAGTCGTCAAAGTCAGTCGATGCCGTA...
      CTTAGACTGCTATGGGACAGTCGTCAAAGTCAGTCGATGCCGTA...
           CTGCTATGGGACAGTCGTCAAAGTCAGTCGATGCCGTA...
             CTATGGGACAGTCGTCAAAGTCAGTCGATGCCGTA...
                   CAGTCGTCAAAGTCAGTCGATGCCGTA...
                      CGTCAAAGTCAGTCGATGCCGTA...
                         CAAAGTCAGTCGATGCCGTA
                            CAGTCGATGCCGTA...
                               CGATGCCGTA...
```

(2). Organize pairs of 18-mers (filter, trim, split)

(3).Search the 18-mer strings for matches

"Suffix Arrays" of the genome

"18-mer Word Arrays" of the SNP genome

(4) Score chromosome or scaffold locations of ditag pairs

• 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 at least one 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 chromosomes, but the distance between them is outside of the specified boundaries, (e.g., 30KBp-50KBp), 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 (invertion).

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

Human breast cancer cell line BT474 has been used for generating fosmid ditag sequences because this cell line has a BAC library end sequence for compare. Two 454 bulk run have generated 575453 pair reads, among them 235394 unique reads made 3.1 fold genome coverage. Total 86 positive unique hit has been found, 14 of those were also detected in the Collins's End Sequence Profiling (ESP) data. (Colin **Collins** lab at UCSF Cancer Center **)**

454 bulk runs 1&2 of BT474 fosmid ditag	# of ditag	%	%
total reads =	**575453**		100.0%
unique reads with flanking vector =	235394	100%	40.9%
genome coverage =	**3.13859**		
Not found (including missing- missing, missing- unique and missing- repeat)	28603	12%	
Normal placement (including unique- unique, and unique- repeat)	186779	79%	
Flagged repeats (including Repeat-repeat and Unique-repeat)	15803	7%	
Flagged unique	4209	2%	
Positive hits	**274**	**0.1%**	
Positive unique hits	**61**		
Insertion:	**2**		
Translocation:	**13**		
Deletion:	**23**		
Inversion:	**23**		

14 out of 61 ditag unique positive hit were detected in the ESP data

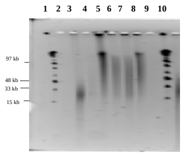
Data from Collins's ESP:

- **1X clone depth**
- **132 rearrangements**
- **16 multiple hits**
- **? false positives**

chr1	120553521	+	SD	chr1	146517584	+		inversion?
chr11	71496704	-		chr11	72375726	-		inversion?
chr17	32944877	+		chr17	46273871	+		inversion?
chr17	35473272	+		chr17	43739226	-		deletion?
chr17	35085806	-		chr17	52107671	-		inversion?
chr17	35160915	+		chr20	56424446	+		translocation?
chr17	43808873	-		chr17	44231573	-		inversion?
chr17	44143568	-		chr17	60362061	-	SD	inversion?
chr20	31522219	+		chr20	53209029	-		deletion?
chr20	33391478	+		chr20	42342847	-		deletion?
chr20	43274238	+		chr20	50174082	-		deletion?
chr20	46030002	+		chr20	52078586	-		deletion?
chr20	50446610	+		chr20	52268629	-		deletion?
chr20	53723166	+		chr20	57745534	+		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 agreeing pairs equal 101 and 109 fold genome clone coverage respectively. We also learned that poor quality source DNA is not suitable for generating ditag sequence.



The picture shows source DNA QC gel, Lane1 is PFG MidRangeI marker, Lane3 is A.EII, shows DNA degraded; lane 4 is D.Aro, shows little amount of DNA; lane5 is R.Met and Lane7 is R.Pal5. The table below shows their ditag sequence results.

organism's name	**# of Unique ditag pairs**	**# of pairs agree with the reference genome**	**# of pairs disagree with the reference genome**
Cupriavidus metallidurans CH34 (R.Met) (6.5 Mb)	21396	16454 (77%)	468 (2.2%)
Rhodopaseudomonas palustris BisB5 (R.Pal5) (5.5 Mb)	20194	15074 (75%)	378 (1.9%)
Acidobacterium sp.Ellin 345 (A.EII)	23613	1069	12134
Dechloromonas aromatica RCB (D.Aro) (4.5 Mb)	1454	216	120

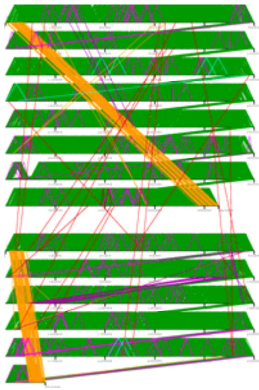
The Graph of Cupriavidus metallidurans CH34 (R.Met) ditag pairs mapped to the reference genome.

Genome:

Cupriavidus metallidurans CH34 has one 3.92 Mb circular chromosomes and one 2.58 Mb circular megaplasmid.

Legend:

GREEN - normal pair
RED - translocated pair
MAGENTA - short pair
DARK RED - short inverted pair
ORANGE - long pair
LILAC - long inverted pair
CYAN - inverted pair



Definitions:

"normal pair" – the distance between the two ditags in a pair is between 30KBp and 50KBp and both ditags are mapped to the same strand
"short pair" - the distance between the two ditags in a pair is less than 30kb
"short inverted pair" - short pair and the ditags are mapped to the opposite strands
"long pair" - the distance between the two ditags in a pair is more than 50kb
"long inverted pair" - long pair and the ditags are mapped to the opposite strands
"inverted pair" - the distance between the two ditags in a pair is between 30kb and 50kb and and the ditags are mapped to the opposite strands
"translocated pair" - the ditags in a pair are mapped to two different chromosomes, scaffolds or contigs.

Summary:

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

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.