SegmMan

SegmMan will cut chromosome into pieces in different sizes. And it adds fragments with Gibson, Goldengate, telomere adaptors to them so that they are able to be assembled into whole experimentally. Besides, it adds flanking vector homologous region and enzyme sites for the preservation and excision from vectors.

Plugin Scripts

This module contains three plugins: 01.whole2mega.pl, 02.globalREmarkup.pl and 03.mega2chunk2mini.pl.

2.1 01.whole2mega.pl

This utility can split the whole chromosome (at least 90kbp long) into about 30k segments and add homologous overlap and adaptors, so that these fragments can be integrated into whole experimentally.

2.1.1 Internal operation

First, this utility searches for the location of centromere and ARSs (autonomously replicating site). The minimal distance between centromere and ARS should NOT be larger than a defined megachunk which is about 30k long.

Second, this utility cuts out the first 30k sequence window containing the centromere and its adjacent ARS, and then adds this megachunk with two original markers and left, right telomeres.

Thirdly, this utility continues to cut more megachunks from the original one to both ends. But these megachunks are not independent, they all have about 1kbp overlaps. Moreover, these new splited window can be given only one marker alternately and only left or right telomere.

The output file will be dealed with 02.globalREmarkup.pl

For more information about segmentation design, please refer to the page ASSEMBLY DESIGN PRINCIPLE .

2.1.2 Example (command line)

perl 01.whole2mega.pl –gff sce_chrI.gff -fa sce_chr01.fa -ol 1000 -ck 30000 -m1 LEU2 -m2 URA3 -m3 HIS3 -m4 TRP1 -ot sce_chrI.mega

2.1.3 Parameters

		default	Option
gff	The gff file of the chromosome being restriction enzyme sites parsing		
fa	The fasta file of the chromosome being restriction enzyme sites parsing (The length of the chromosome is larger than 90k)		

ol	The length of overlap between megachunks	1000bp	
ck	The length of megachunks	30kbp	
m1	The first marker for selection alternately	LEU2 (1797bp)	LEU2/URA3/ HIS3/TRP1
m2	The second marker for selection alternately	URA3 (1112bp)	LEU2/URA3/ HIS3/TRP1
m3	The first marker orinally residing in first 30k segmentation	HIS3 (1774bp)	LEU2/URA3/ HIS3/TRP1
m4	The second marker orinally residing in first 30k segmentation	TRP1 (1467bp)	LEU2/URA3/ HIS3/TRP1
ot	The output file	Prefix(fa filename) + suffix(.mega)	

2.1.4 The format of output:

The output file is stored in /the path where you install GENOVO/Result/ 01.whole2mega.

Besides, there is screen output about the process state and result.

Screen output

01.state

Store the segmentation information

Megachunk_ID	Corresponding location in the designed chromosome
Part ID	Location in the segmentation

```
 1 01.whole2mega/sce_chr01_-6.mega 1-29794
 2 left_telomere   1       689
 3 URA3    690     1806
 4 Part of chromosome      1807    31601
 5
 6 01.whole2mega/sce_chr01_-5.mega 28795-58586
 7 left_telomere   1       689
 8 LEU2    690     2491
 9 Part of chromosome      2492    31283
10
11 01.whole2mega/sce_chr01_-4.mega 57587-87378
12 left_telomere   1       689
13 URA3    690     1806
14 Part of chromosome      1807    30598
15
16 01.whole2mega/sce_chr01_-3.mega 86379-116170
17 left_telomere   1       689
18 LEU2    690     2491
19 Part of chromosome      2492    31283
```

*.mega

Store the fasta information of the 30k segments

```
 1 >01.whole2mega/sce_chr01_0.mega 172755-202755
 2 CACACACACCACACCCACACCACACACACACCACACCCACACCAC
 3 ACCACACCCACACCACACACCACACCCACACACCACACCCACACA
 4 ACACACACACACACCACACCCACCACACCCACACACCCACACCCA
 5 CACATAACTTCGTATAATGTACATTATACGAAGTTATCACATCAT
 6 CACGATTATCCACATTTTAATATCTATATCTCATTCGGCGGCCCC
 7 TAACCACTCAATTTATATACACTTATGTCAATATTACAAAAAAAT
 8 AACACATAACTTCGTATAATGTACATTATACGAAGTTATACTACC
 9 ATTCGGTCGAAAAAAGAAAAGGAGAGGGCCAAGAGGGAGGGCATT
10 TATGTCTGTTATTAATTTCACAGGTAGTTCTGGTCCATTGGTGAA
11 GACTTGCTGGGTATTATATGTGTGCCCAATAGAAAGAGAACAATT
12 CAGGCACTCCGAAATACTTGGTTGGCGTGTTTCGTAATCAACCTA
13 TGGAGATGAGTCGTGGCAAGAATACCAAGAGTTCCTCGGTTTGCC
14 TCACAGAAACCTCATTCGTTTATTCCCTTGTTTGATTCAGAAGCA
15 AAGAGAGCCCCGAAAGCTTACATTTTATGTTAGCTGGTGGACTGA
16 AAGCGGAGGTGTGGAGACAAATGGTGTAAAAGACTCTAACAAAAT
17 AAGTATTGTTTGTGCACTTGCCTGCAGGCCTTTTGAAAAGCAAGC
18 ATCATTTGGCTTTTTGATTGATTGTACAGGAAAATATACATCGCA
19 TTCACAGGCGCATACGCTACAATGACCCGATTCTTGCTAGCCTTT
```

3.2 02.globalREmarkup.pl

This utility will parse the exited restriction enzyme sites residing in the chromosome.

3.2.1 Internal operation

This utility searches the exited restriction enzyme sites along the chromosome both plus strand and minus strand, after users define the list of enzymes.

Besides, we tried to find out all the potential restriction enzyme sites, so that maybe some unusual restriction enzyme sites can be created and let segmentation go. But because it had low efficiency, we're still working on that.

The output file will be dealed with 03.mega2chunk2mini.pl

For more information about segmentation design, please refer to the page ASSEMBLY DESIGN PRINCIPLE .

3.2.2 Example (command line)

perl 02.globalREmarkup.pl -sg 01.whole2mega/sce_chrI.mega -re standard_and_IIB -ct Standard.ct –ot sce_chrI.mega.parse

3.2.3 Parameters

		default	Option
sg	The fasta file of the chromosome being 30k segmentated, the output of 01.whole2mega.pl		
re	The restriction enzyme sites list. It is devided by different standards, type (IIP, IIA, IIB), cost (standard, nonexpensive) and etc.	Standard_and_IIB	IIP/IIA/IIB/Standard/ Nonexpensive/ Standard_IIB Nonexpensive_IIB
ct	The codon table file of operated organism.	Standard.ct	See the list below.
out	The output file	Prefix(fa filename) + suffix(.parse)	

Codon table list
1 The Standard Code
2 The Vertebrate Mitochondrial Code
3 The Yeast Mitochondrial Code
4 The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
5The Invertebrate Mitochondrial Code
6 The Ciliate, Dasycladacean and Hexamita Nuclear Code
7 The Echinoderm and Flatworm Mitochondrial Code

8 The Euplotid Nuclear Code
9 The Bacterial, Archaeal and Plant Plastid Code
10 The Alternative Yeast Nuclear Code
11 The Ascidian Mitochondrial Code
12 The Alternative Flatworm Mitochondrial Code
13 Blepharisma Nuclear Code
14 Chlorophycean Mitochondrial Code
15 Trematode Mitochondrial Code
16 Scenedesmus Obliquus Mitochondrial Code
17 Thraustochytrium Mitochondrial Code
18 Pterobranchia Mitochondrial Code
19 Candidate Division SR1 and Gracilibacteria Code

3.2.4 The format of utput

The output file is stored in /the path where you install GENOVO/Result/. 02.globalREmarkup.

Besides, there is screen output about the process state and result.

Screen output

*.parse

Store the exited enzyme recognition site in the megachunks

Enzyme ID	Start	End	Recognition site	Real site

1	BsrGI	318	323	TGTACA	TGTACA
2	TatI	318	323	WGTACW	TGTACA
3	CviQI	319	322	GTAC	GTAC
4	RsaI	319	322	GTAC	GTAC
5	RsaI	319	322	GTAC	GTAC
6	MslI	340	349	CAYNNNNRTG	CATCATTATG
7	MslI	340	349	CAYNNNNRTG	CATCATTATG
8	HpyCH4V	348	351	TGCA	TGCA
9	HpyCH4V	348	351	TGCA	TGCA
10	BceAI	351	355	ACGGC	ACGGC
11	BbvCI	361	367	CCTCAGC	CCTCAGC
12	Bpu10I	361	367	CCTNAGC	CCTCAGC
13	MnlI	361	364	CCTC	CCTC
14	BseMII	362	366	CTCAG	CTCAG
15	BspCNI	362	366	CTCAG	CTCAG
16	DdeI	362	366	CTNAG	CTCAG
17	MspA1I	364	369	CMGCKG	CAGCGG
18	MspA1I	364	369	CMGCKG	CAGCGG
19	AciI	366	369	CCGC	GCGG

3.3. 03.chunk_30k_10k_2k.pl

This utility can produce 2k minichunks with Gibson adaptors and 10k chunks with goldengate adaptors.

3.3.1 Internal operation

This utility will segment the mega chunk produced by 03.mega2chunk2mini.pl into 2k minichunks with Gibson assembly adaptors, so that they can be put together into 10k chunks.
First, this bin will search the inexistent restriction enzyme sites locally, and then decide the size of the minichunks according to the requirements from users, and add two same Gibson adaptors to each sides of minichunks.
Secondly, the second part of this bin will define the start and end point of the chunks as users asked and design goldengate assembly adaptors for the chunks.
The output file can be sent in gene synthesis company after human attention and double check.
For more information about segmentation design, please refer to the page ASSEMBLY DESIGN PRINCIPLE .

3.3.2 Example (command line)

perl 03.mega2chunk2mini.pl -re standard_and_IIB -sg 01.whole2mega/sce_chr01_0.mega -ps 02.globalREmarkup/sce_chr01_0.parse -ot 03.mega2chunk2mini

3.3.3 Parameters

		default	Option
sg	The fasta file of the 30k segmentation, the output of 01.wh2mega.pl		
ps	The markup file of the 30k segmentation, the output of 02.globalREmarkup.pl		
re	The restriction enzyme sites list. It is devided by different standards, type (IIP, IIA, IIB), cost (standard, nonexpensive) and etc.	Standard_and_IIB	IIP/IIA/IIB/Standard/ Nonexpensive/ Standard_IIB Nonexpensive_IIB
a2	2k to 10k assembly strategy (Gibson or Goldengate)	Gibson	Gibson/ Goldengate
a10	10k to 30k assembly strategy	Goldengate	Gibson/ Goldengate

	(Gibson or Goldengate)		
ckmax2	The maximum length of minichunks	2200 bp	
ckmin2	The minimum length of minichunks	1800 bp	
cknum	The number of minichunks in a chunk	5	

If parameter a2 is Gibson, then there are additional parameters:

ol2	The length of overlap	40 bp	
tmax2	The maximum melting temperature of the overlap of minichunks	60℃	
tmin2	The minimum melting temperature of the overlap of minichunks	56℃	
fe2	The minimum free energy of the overlap of minichunks	-3	
ex2	The type of exonuclease used for minichunks	T5	T5/T3
lo2	The minimum distance between minichunks overlap and loxpsym	40 bp	
en2	The type of enzyme flanking minichunks	IIP	
et2	The temperature of enzyme used in minichunks digestion	37℃	
ep2	The maximum unit price of enzyme used in minichunks digestion	0.5 $/unit	

If parameter a10 is Goldengate, then there are additional parameters:

en10	The type of enzyme flanking chunks	IIB	IIA/IIB
et10	The temperature of	37℃	

	enzyme used in chunks digestion		

3.3.4 The format of ouput

The output file is stored in /the path where you install GENOVO/Result/. 03.mega2chunk2mini.

Besides, there is screen output about the process state and result.

Screen output

*.2kstate

Store the minichunks states.

Left IIP enzyme site	Right IIP enzyme site	Start	End	Size of minichunks	Melting temperature of overlap

1	*	BamHI	1	2000	2000	67.21
2	BamHI	BamHI	1961	3953	1993	69.04
3	BamHI	BamHI	3914	6069	2156	67.04
4	BamHI	BamHI	6030	8000	1971	67
5	BamHI	*	7961	9970	2010	68.56
6	*	BamHI	9931	11956	2026	66.32
7	BamHI	BamHI	11917	14006	2090	66.95
8	BamHI	BamHI	13967	16063	2097	68.98
9	BamHI	BamHI	16024	18008	1985	67.8
10	BamHI	*	17969	19881	1913	69.05
11	*	EcoRI	19842	22067	2226	68.52
12	BamHI	BamHI	22028	24057	2030	66.89
13	HindIII	HindIII	24018	26049	2032	67.48
14	HindIII	HindIII	26010	28080	2071	68.26
15	BamHI	*	28041	30107	2067	67.17
16	*	BamHI	30068	32078	2011	68.83
17	EcoRI	EcoRI	32039	33878	1840	71.74
18	EcoRI	*	33839	34630	792	

*.10kstate

Store the chunks states.

Left IIB enzyme site	Right IIB enzyme site	Start	End	Size of chunks

1	*	AjuI	1	9970	9970
2	AjuI	AjuI	9931	19881	9951
3	BaeI	BaeI	19842	30107	10266
4	BaeI	*	30068	34630	4563

*.mini

Store the fasta of designed minichunks.

```
>*      BamHI   1       2000    2000    67.21
CACACACACCACACCCACACCACACACACACCACACCCACACCACA

>BamHI  BamHI   1961    3953    1993    69.04
GGATCCTTTGCATAAACACCATCAGCCTCAAGTCGTCAAGTAAAGA

>BamHI  BamHI   3914    6069    2156    67.04
GGATCCATTTCAACTACAGTGGCACCTAGAGACCAAATGTCGCTGA

>BamHI  BamHI   6030    8000    1971    67
GGATCCATACCTGTACAGGTTTCATTCGTAAAGCAGGGACTCTAGT

>BamHI  AjuI    7961    9970    2010
GGATCCATTCGATCCTCATGCAGCCCTCGTTAATATGCTAAAATGG

>AjuI   BamHI   9931    11956   2026    66.32
7GATTAGTAGTATAGCAAAAGTAACACTTGTCCACCGCAGACTCCA

>BamHI  BamHI   11917   14006   2090    66.95
GGATCCAGTAAAAAAAAATAACGACAACTGCAGGACTCGAACCTGC
```