Appendix

B3.1—*Spo11Mapper* (*v2.7*)

Aim: Alignment, filtering and analysis of genome-wide Spo11-mapping data

Input(s): Paired-end _R1 and _R1 FASTQs, indexed reference genome (FASTA), user configuration

Output(s): 1bp Histograms, molecule sizes, molecule frequencies, alignment logs

Req(s): Perl 5.25.9, BioPerl, Bash 4.1

Spo11Mapper constitutes a novel, low memory software package for the automated batch processing of *sae2*Δ Spo11 DSBs, Topo II DSB/SSBs, Spo11-oligos or any library containing informative Read-1/2 5' ends. *Spo11Mapper*, run on the command line, requires two main arguments:

```
Usage: Spo11Mapper -i [INPUT FOLDER] -c [CONFIGURATION FILE]
-i INPUT: Input data folder containing paired-end FASTQ files
-c CONFIG: Configuration file specifying user-parameters (Spo11Mapper.config)
```

B3.1.1—Configuration

Spo11Mapper is initially configured via an external .config file specifying key variables:

(i) *CALL_MODE* (*SINGLE/DOUBLE/OLIGO*)—specification of library type which in turn differentiates the type of coordinates called and the analyses performed (single cuts, double cuts, oligos)

(ii) *SPACE_SAVER* (*Y/N*)—when enabled, *Spo11Mapper* will reduce the disk footprint of the pipeline, progressively deleting non-essential files including .SAM and .FASTQ files

(iii) *CORE*—no. of CPU cores available

(iv) *READ1/2_EXT*—FASTQ file extension for automated detection of paired end samples

(v) *GLOBAL_OPTIONS—s*pecification of Bowtie2 parameters for end-to-end alignment. Default settings, utilised throughout this chapter, are (-X 1000 --no-discordant --very-sensitive --mp 5,1)

(vi) *LOCAL_OPTIONS*—specification of Bowtie2 parameters for local alignment (default as above)

(vii) *GENOME_DIR*, *GENOME–NAME*—directory and filename of a Bowtie2 indexed FASTA reference genome. Data throughout this chapter was aligned against a modified S288c reference (SGD Jan 2015 - R64-2-1) containing sequence for the exogenous hotspots, *HIS4*::*LEU2* and *LEU2*::*HISG*.

(viii) *TRIM*, *TRIM_LEN*—when enabled, *Spo11Mapper* will perform a two step alignment (see: Section 3.4), 3'→5' trimming unmapped mates by a length specified by TRIM_LEN.

Spo11Mapper.config - Example

```
##############################################################
## Program Settings
##############################################################
CALL_MODE = DOUBLE
SPACE_SAVER = N
CORE = 4
##############################################################
## Input Data
##############################################################
READ1_EXT = _R1
READ2_EXT = _R2
##############################################################
## Alignment Options
##############################################################
GLOBAL_OPTIONS = -X 1000 --no-discordant --very-sensitive --mp 5,1
LOCAL_OPTIONS = -X 1000 --no-discordant --very-sensitive --mp 5,1
GENOME_DIR = /usr/local/Genomes/Cer3H4L2/
GENOME_NAME = Cer3H4L2
TRIM = Y
TRIM_LEN = 10
```

B3.1.2—SAM files

Tab delimited SAM files, the primary output of *Bowtie2* alignment, specify (i) 1-based leftmost coordinates (by Chr, Position) for all mapped or partially mapped read pairs (ii) the mapped or unmapped read sequence with associated quality scores (iii) Numerical "flags" denoting the aligned identity of each read pair:

Paired, *fully aligned* (*SAM Flags*)

99 - Read-1, Watson | 147 - Read-2, Crick

83 - Read-1 Crick | 163 - Read-2, Watson

Paired, *partially aligned* (*one–mate*)

73 - Read-1, Watson, Mapped | 133 - Read-2, Unmapped

89 - Read-1, Crick, Mapped | 133 - Read-2, Unmapped

69 - Read-1, Unmapped | 137 - Read-2, Watson, Mapped

69 - Read-1, Unmapped | 153 - Read-2, Crick, Mapped

(iv) Alpha-numeric CIGAR codes describing base by base alignment (5’->3’) and detailing the presence of INDELs. SNPs/mismatches are not included. For example, a CIGAR code of *5M2I30M1D25M* denotes:

- 5bp reference match (5M) (“M” may contain unspecified mismatches)
- 2bp insertion in the read (relative to the reference) (2I)
- 30bp reference match (30M)
- 1bp deletion in the read (relative to the reference) (1D)
- 25bp reference match (25M)

(v) Alpha-numeric MD:Z tags denoting the position and base composition of any SNPs/deletions present in the read relative to the reference. Insertions (relative to the reference) are not specified. For example, an MD:Z-tag of 0T0C25A5^T10 denotes (from left to right):

- An initial 2bp mismatch (reference specifies TC, read contains alternative bases) (0T, 0C)
- 25bp of precise reference:read match (25) followed by a 1bp mismatch (25A)
- 5bp of precise reference:read match (5) followed by a 1bp deletion in the read (reference contains a T) (5^T)
- 10bp of precise reference:read match (10)

ATGAGCGTACCTGTAAATAAGAAGATC**G**ATCGA**_**GGTACATACT — *READ* (*0T0C25A5^T10*)

TCGAGCGTACCTGTAAATAAGAAGATC**A**ATCGA**T**GGTACATACT — *REF*

Spo11Mapper collectively reads and interprets this information read-by-read for each .SAM file to accurately filter and extract high quality coordinate lists and facilitate read trimming.

B3.1.3—Orientation and ambiguous end filtering

Coordinate calling is only performed for properly paired, fully aligned 99-147 (Read-1 Watson—Read-2 Crick) or 83-163 (Read-1 Crick—Read-2 Watson) read pairs that pass ambiguous end and orientation checks. Atypical read orientations can arise through the sequencing of self circles or incomplete sequencing runs, generating asymmetrically overlapping read pairs. SAM "flags", MD:Z-tags or CIGAR codes contain no information pertaining to this phenomenon—thus, *Spo11Mapper* employs a custom filter using the leftmost positional information held in .SAM files. Any 99/147 or 83/163 read pairs where the 3' end of the (-) Crick read is to the left (upstream) of the 5' end of the (+) Watson read—signifying asymmetric overlap—is discarded. Ambiguous ends are determined through parsing and interpretation of MD:Z-tags, detecting mismatches at any informative end. In SINGLE mode, 2bp of mismatch or more at the Read-1 5' end disqualifies a read and enters it into a separate, ambiguous coordinate dataset. A single terminal base of mismatch is permitted to accommodate alignment of data from divergent strains (e.g. SK1 to S288c). In DOUBLE mode, the same threshold is applied to both Read-1 and Read-2 5' ends. If either end is ambiguous, the entire read pair is disqualified and separated. Non-informative 3' ends are not considered. A MD:Z-tag of *0T0C73* (2bp 5' mismatch) would thus fail the ambiguous end check, while a tag of *0C74* (1bp 5' mismatch) or 74*T* (1bp 3' mismatch) would pass.

B3.1.4—Coordinate calling

For properly oriented, unambiguous 99-147 and 83-163 read pairs, coordinate positions of the informative (e.g. Spo11) ends are calculated. SAM files specify 1-based leftmost coordinates—thus, for Watson (+) reads (99 or 163) the 5' end is readily called by *Bowtie2*. In contrast, for Crick (-) reads (83 and 163), the leftmost base is the 3' end of the read. To call 5' Crick (-) coordinates, CIGAR codes are parsed and scored to determine the mapped read length—according to the following rules: (M = 1, D = 1, I = 0)—which is then added to the 3' coordinate. Insertions (I) (in the read) are ignored in order to call coordinates accurate to the utilised reference. As an example, a Crick (-) read with a CIGAR code of 75M and a leftmost coordinate is 10200 is called as 102074

(10200+75-1). A 1bp adjustment is made as the leftmost base is included as part of 75M. A more complex Crick (-) read with a CIGAR code of 35M2D10M3I30M and a leftmost coordinate of 10200 is called as 10276 (10200 + 35 + 2 + 10 + 30 -1).

B3.1.5 —3’->5’ Trimming

If enabled (TRIM=Y), *Spo11Mapper* additionally handles SAM flags 73-133, 89-133, 69-137 and 69-153—all of which denote pairs of reads where only a single mate mapped. During first round processing of SAM files, trimmed FASTQ files are reconstructed by *Spo11Mapper* in a standard, four line format:

```
@M00561:9:000000000-ALBWJ:1:1101:15097:1775 1:N:0:1 - Read Header
TAATGAATTAATCAACTTCAACTCATCACTGCCCAATGATTCGTCGGGTTTCACTATTTTTAGATAATCTTCCCT - Seq
+
@—A---CE,,CC,,;EEE,,;CC,C;,,;C,<;;,,,;,,,<,<,;@+++,886,<C,<@CEF,,,<,,<<@C,;CC - Quality
```

Mapped mates (73, 89, 137 and 153) are sorted into their respective Read-1 or Read-2 trimmed FASTQ files “as is”, without trimming. SAM files store mapped read sequences based on the top (+) strand, regardless of which strand the read aligned to. Sequences for mapped Crick (-) mates (89, 153) are thus reverse complemented before addition to a FASTQ file. Read sequences for unmapped mates (133, 69) are trimmed from the 3’ end as are the associated quality lines. SAM files store the actual read sequence for unmapped mates, thus no further processing is required. Untrimmed FASTQ files are subsequently auto detected for all samples and entered into *-local Bowtie2* alignment. Local alignment mode is less strict, however it also permits Bowtie2 to trim reads from *either* end if it improves MAPQ (quality) scores. Any *-local* trimming is recorded into the CIGAR code as ’S’. As *-local* trimming at the 5’-end lowers the integrity of coordinate calling, ambiguous end filtering is expanded to disqualify any reads with ’S’ CIGAR entries at an informative end (e.g. 4S71M). Any read pairs that now sufficiently align will be marked by 99-147 or 83-163 flags, undergo coordinate calling if unambiguous and properly oriented as above and are appended onto the main dataset.

B3.1.6—Aligning Spo11-oligos

Alignment of ~25-65bp Spo11-oligos requires additional pre-alignment processing. Any oligo or double cut molecule shorter than the sequencing run length (i.e. 75bp) will contain poly(G/C) tails —added during Spo11-oligo library prep—and portions of illumina adaptor at both ends of the associated read. In order to improve mappability of Spo11-oligos, *Spo11Mapper* includes a utility script (*OligoTrim.pl*) capable of trimming FASTQ files according to user specified sequence patterns:

```
perl OligoTrim.pl –1 <Read 1 FASTQ> –2 <Read 2 FASTQ> –U <5’ Upstream Pattern> –L <3’ Downstream Pattern>
```

Multiple patterns may be specified. All Spo11-oligo libraries analysed throughout this chapter were trimmed using -U CCCC -U CCC -L GGGG -L GGG, to accommodate situations where poly(G) tailing was incomplete.

B3.1.7—Log Files

Spo11Mapper records several log files: (i) Individual alignment reports per sample, including a summary of coordinate calling (ii) Batch, strain summary, detailing stats for each sample processed (see: Figure 3.4) (iii) System log, detailing errors and recording the command-line output.

Alignment Report (Individual Sample) - Example

```
MJ315_WT_2A_6h
----------------------
GLOBAL
----------------------
3512089 reads; of these:
  3512089 (100.00%) were paired; of these:
    164323 (4.68%) aligned concordantly 0 times
    3190282 (90.84%) aligned concordantly exactly 1 time
    157484 (4.48%) aligned concordantly >1 times
    164323 pairs aligned 0 times concordantly or discordantly; of these:
      328646 mates make up the pairs; of these:
        215163 (65.47%) aligned 0 times
        89083 (27.11%) aligned exactly 1 time
        24400 (7.42%) aligned >1 times
96.94% overall alignment rate
```

```
----------------------
TRIMMED
----------------------
68175 reads; of these:
  68175 (100.00%) were paired; of these:
    14152 (20.76%) aligned concordantly 0 times
    27312 (40.06%) aligned concordantly exactly 1 time
    26711 (39.18%) aligned concordantly >1 times
    ----
    14152 pairs aligned 0 times concordantly or discordantly; of these:
      28304 mates make up the pairs; of these:
        10822 (38.23%) aligned 0 times
        10900 (38.51%) aligned exactly 1 time
        6582 (23.25%) aligned >1 times
92.06% overall alignment rate

----------------------
CALL STATS
----------------------
Total Hits:         3401787
----------------------
Valid Hits:         3377298
    Global:         3341250
   Trimmed:         36048
----------------------
Ambig Hits:         24489
```

System Log - Example

```
--------------------------------------------
FASTQ Alignment (Global --end-to-end)
--------------------------------------------
Currently aligning:
MJ315_WT_2A_6h
-------------------------------------
Calculating Coordinates....
-------------------------------------
Currently processing:
MJ315_WT_2A_6h
----------------------
Completed
Runtime: 0:05:50
______________________


------------------------------------------
FASTQ Alignment (Trimmed --local)
------------------------------------------
Currently realigning:
MJ315_WT_2A_6h
-------------------------------------
Calculating Coordinates....
-------------------------------------
Currently processing:
MJ315_WT_2A_6h
----------------------
Completed
Runtime: 0:00:14
----------------------
```

B3.1.8—Analysis and output files

Spo11Mapper generates several key output files: (i) all raw unambiguous, coordinate calls are recorded into tab delimited .txt files, which are in turn utilised for downstream analysis:

Single Cut Library:

Strand	Chr	Pos	ReadLength	CIGAR	Adjustment	Read-Flag
w	9	204982	75	75M	0	99
c	15	1059148	75	75M	74	83
c	6	36193	75	75M	74	83
c	6	221802	73	73M	72	83
c	6	226858	75	43M1D32M	75	83

Double Cut Library (*Paired W–C Lines*, *Molecule Size Added*):

PairID	Strand	Chr	Pos	ReadLength	CIGAR	Adjustment	Read-Flag	mSize
1	w	7	307254	19	19M	0	99	21
1	c	7	307274	19	19M	18	147	21
2	w	13	577501	23	23M	0	99	25
2	c	13	577525	23	23M	22	147	25
3	w	15	770237	25	25M	0	163	27
3	c	15	770263	25	25M	24	83	27

Ambiguous, filtered calls are stored in separate but identically formatted files (ii) Sparsely-formatted, 1bp histograms are produced for Read-1 5’ ends, and separately, Read-2 5’ ends (if detected) and stored as tab delimited .txt files, detailing the total number of Watson (+) and Crick (-) per base pair on each chromosome (iii) In DOUBLE mode, *Spo11Mapper* calculates the frequency of molecule sizes as the absolute distance between Read-1 5’- and Read-2 5’ ends for each unambiguous read pair. Moreover, the system records the frequency with which any given pair of 5’ coordinates is observed. Molecule sizes and frequencies are stored as tab delimited .txt files:

1bp Histogram:

Chr	Pos	Watson	Crick
1	6457	32	0
1	6458	0	32
1	6463	0	1
1	6468	2	10
1	6469	0	1

Molecule Sizes:

Size	Freq
20	42104
21	53755
22	62272
23	69923
24	82517

Molecule Frequencies:

Chr	Coord-A	Coord-B	R1W Freq	R1C Freq
1	2434	2460	1	5
1	2434	2461	0	6
1	2434	2462	5	12
1	2434	2464	0	2
1	2434	2465	2	0