1. stdchc.svg diagram

(1) Coverage

Harnessing *SOAPaligner 2.21*（http://soap.genomics.org.cn/soapaligner.html），the clean sequencing reads are aligned to the reference sequence (removed Chr09 right arm of BY4741 and added IXR_BACseq). The sequencing depth and coverage are calculated based on the alignments. The single base sequencing depth is calculated by using *soap.coverage 2.27*.

Blue column chart shows the relative sequencing coverage per 10 bases: Relative sequencing coverage = $\frac{\sum_1^{10} \mathrm{Dx}}{10*\mathrm{M}}$ (Dx: the single base sequencing depth; M: the maximum single base sequencing depth)

The coverage information is very important for duplication and deletion variation.

(2) Axis and Loxp site

The axis shows the position of IXR_BACseq and the Loxp sites are marked in green line.

(3) SV visualization

All the structure variations, including deletion (DEL), inversion (INV), tandem duplication (TDUP), non-tandem duplication (NTDUP), and deletion & insertion (D&I), are displayed in the figure. The rectangles indicate the region of variation. The DEL is marked in gray and the INV, TDUP, NTDUP, D&I are marked in red, dark blue, royal blue and indigo respectively. For NTDUP and D&I, the target position is showed in an arrow and the direction is showed in “+” or “-”. “+” means the variation region inserts target position in forward sequence; on the contrary, “-” means the variation region inserts target position in complementary sequence. For the SV details, please refer to the *.lsv file.

(4) Loxp mapping reads

For each Loxp site, the regular mapping single reads with loxp sequence are counted based on the alignments. The Loxp mapping reads of breakpoint would be significantly less than others.

(5) Split reads

The unmapped reads with loxpsym sequence (split reads) are split into two ends (split read ends) by removing the loxpsym sequence. Each ends, which is more than15 base pair, are aligned to the IXR_BACseq sequence by using *bowtie2-2.0.0*. In the picture, the arrows indicate the mapping position and direction of the both split ends. If the both split ends have the same mapping direction, they are connected with blue dotted line and the Loxp sites are marked in blue line. On the contrary, the both split ends aren’t connected and the loxp sites are marked in red line. Based on this information, the breakpoint and connected relation of both side of breakpoint can figure out easily.

(6) Irregular mapping reads on UnSCRaMbLEd reference

The pair end reads with irregular mapping include the pair end reads with the same mapping direction, and the reads with mapping distance more than 625bp or less than 375bp. These reads may have important information on structure variation. The pair end reads with the same mapping direction are marked in red line, the reads with mapping distance > 500 are marked in blue, and the rest are marked in gray. There exists a background due to library construction. The irregular mapping reads on UnSCRaMbLEd reference contain important information. According to these irregular pair ends mapping, we can further confirm the breakpoint and connected relation.

(7) SV Validation

The SCRaMbLEd sequence (IXR_BACseq_scb) is refactored based on the

variation that we found. Using the *SOAPaligner2*, the clean sequencing reads are aligned to the reference sequence (removed Chr09 right arm of BY4741 and added IXR_BACseq_scb). The regular mapping and irregular mapping reads on SCRaMbLEd sequence based on the alignments.

In theory, a perfect SCRaMbLEd sequence, which was refactored by total variation, is mapped by the sequencing reads with no fragmentation in regular mapping and no any clusters in irregular mapping. And all irregular pair end mapping reads with variation information and split reads with loxp sequence on UnSCRaMbLEd reference can map to it normally (reads mapping with regular transformation).

The mapping reads with regular transformation (transformation reads) are marked in colorful line in the regular mapping. The split reads are marked in green and the rest reads are marked in the same color with irregular mapping reads on UnSCRaMbLEd reference. In addition, we also count the reads possible containing variation information and transformation reads.

2. *.lsv

This file records the structure variation for each sample.

Eg:

```
DEL         34623     36856
INV         13922     16400
NTDUP     46561     88994     54158     54159     -
TDUP       54159     87190     54158     54159     +
D&I         86193     87190     13921     13922     +
#
```

Column ID: Meaning

[1]: type of the SV, which including:

DEL (deletion),

INV (inversion),

TDUP (tandem duplication),

NTDUP (non-tandem duplication),

D&I (deletion & insertion)

[2]: the start site of variation region;

[3]: the end site of variation region;

[4]: the start site of insertion for variation region;

[5]: the end site of insertion for variation region;

[6]: strand (+/-)

"+" means the variation region inserts target position in forward strand;

"-" means the variation region inserts target position in complementary sequence.

The Column [4], [5] and [6] only exist for TDUP, NTDUP and D&I. The lines beginning with # is some description information.

2. lxpchc.svg diagram

(1) Reads with loxp site that can regularly map to reference before Scrambling

By applying *SOAPaligner 2.21*（http://soap.genomics.org.cn/soapaligner.html），the clean sequencing reads are aligned to the reference sequence (removed Chr09 right arm of BY4741 and added *IXR_BACseq*). The mapping of reads with complete loxp sequence is shown. Each pair end reads are marked in blue line. In addition, the number of mapping reads for each loxp site is counted and shown at each loxp site in purple.

(2) Reads with loxp site that can regularly map to reconstructed sequence

Based on the structure variations we identified, new scrambled chrIXR sequence was constructed and named as *Reconstructed_IXR.* Again, with *SOAPaligner2.21*, all reads are aligned to the new reference sequence (removed Chr09 right arm of BY4741 and added *Reconstructed_IXR*). Based on the alignments, the mapping of reads with complete loxp sequence on *Reconstructed_IXR* sequence is shown.

Reads that can both map to *IXR_BACseq* and *Reconstructed_IXR* sequence are marked in blue, and reads that can only map to *Reconstructed_IXR* sequence are marked in red. And the number of mapping reads for each loxp site is shown in purple.

(3) Summary of loxp reads mapping

This part records the statistics of loxp reads, including total number of reads with complete loxp site, the number of these reads can/cannot map to reference, and the ratio of reads with recombination event happened that can regularly map back to new construction reference.

Those loxp reads with recombination event happened have to meet following criteria:

> have complete loxp site,

> can NOT map to *IXR_BACseq*,

> with no indel happened,

> with less than 2 SNP,

> with high sequencing quality (Ns<4, low quality base<15).

Theoretically, if the structure variation prediction is correct, then all loxp sites in

reconstructed reference should be covered by all reads used for alignment with significant supporting number. In addition, those loxp reads with recombination event happened can regularly map back to the reconstructed reference.

(4) Unmapping loxp reads check

This part records the further analysis of all loxp reads that cannot map back to *Reconstructed_IXR* sequence. To figure out the reason, these reads are aligned to the new construction sequence by *bwa0.5.6* (http://bio-bwa.sourceforge.net/). Since strict mapping criteria (no gap is allowed, mismatch <=2) is applied in SOAP alignment, some reads containing indels or more than 3 SNPs will be identified as unmapping reads. We suspect this might be one of the reasons, so we apply a less stringent criteria (bwa aln -o 3 -e 64 -i 2 -L -l 31 -k 4 -t 2 -M 1 -O 8 -E 2) in BWA alignment to identify the unmapping reason.

It is worth paying attention that an insertion (IXR_BACseq, at 91012, a G base is inserted) exists in the segment 44 after 8 bases of the last Loxp site. This will also cause unmapping result in SOAP alignment.

Since some those unmapping reads still cannot find out the reason after BWA alignment analysis, the bowtie alignment is performed again to check if the reconstructed sequence misses a connection. We abstract the reads, which are marked with “*”, from BWA alignment analysis results. These reads are split into two segments by removing the loxp sequence. Each ends, which is more than 15 base pair, are aligned to the *IXR_BACseq* sequence with *bowtie2-2.0.0*. Generally, no any pair-end (PE) cluster exists if the reconstructed *sequence* is actual a sequence.

For the reads mapped in BWA alignment, the format specifications of each line as follow:

Eg:

1 FCC0CJEACXX:8:2102:21060:30063#CGTAGGAC/2 BWA Mismatch - Reconstructed_IXR 163176 100M MD:Z:0C5T0G0T61G29
2 FCC0CJEACXX:8:1108:10720:29186#CGTAGGAC/1 BWA Mismatch + Reconstructed_IXR 164147 100M MD:Z:97A0G0A0
3 FCC0CJEACXX:8:2102:4187:200041#CGTAGGAC/2 BWA Indel - Reconstructed_IXR 164149 94M1I5M MD:Z:99
4 FCC0CJEACXX:8:1305:20730:188868#CGTAGGAC/1 BWA Indel - Reconstructed_IXR 164155 88M1I11M MD:Z:99

Column ID: Meaning

[1]: read number;

[2]: read ID;

[3]: mapping tools;

[4]: reason of the read which can NOT map to new constructed sequence,

“Indel”: a short insertion or deletion in read

“Mismatch”: >=2 SNP in read

[5]: strand + or -;

[6]: chromosome ID;

[7]: coordinate of read mapping;

[8]: match pattern (CIGAR format)

A CIGAR string is comprised of a series of operation lengths plus the operations when reads are performed alignment. The conventional CIGAR format allows for three types of operations: M for match or mismatch, I for insertion and D for deletion. This column records indels information.

[9]: mismatch positions

This string is used for recording mismatched positions of reads and reference bases in the format of [0-9]+(([ACGTN]|\^[ACGTN]+)[0-9]+).

For the reads unmapped in BWA alignment, the format specifications of each line as follow:

Eg:

39 FCC0CJEACXX:8:1107:21198:74413#CGTAGGAC/2 Ns 11 23 73
40 FCC0CJEACXX:8:1207:21206:113023#CGTAGGAC/2 Ns 13 24 68
41 FCC0CJEACXX:8:2106:15180:139026#CGTAGGAC/2 Low_quality 0 26 74
42 FCC0CJEACXX:8:2305:21049:95667#CGTAGGAC/2 Low_quality 0 20 79
44 FCC0CJEACXX:8:1304:14355:184810#CGTAGGAC/2 * 0 0 100
FCC0CJEACXX:8:1304:14355:184810#CGTAGGAC/2_2 Bowtie SE - IXR_BACseq 54208 66M MD:Z:66
43 FCC0CJEACXX:8:2106:5423:73555#CGTAGGAC/2 * 0 0 96
FCC0CJEACXX:8:2106:5423:73555#CGTAGGAC/2_1 Bowtie SE - IXR_BACseq 54176 51M8I4M MD:Z:53C1

Column ID: Meaning

[1]: read number;

[2]: read ID;

[3]: “-”, nonsense;

[4]: reason of the read which can NOT map to new construction sequence,

“Ns”: number of N >=4 in read

“Low_quality”: number of Low quality base >=15 in a read. (Phred score of the low quality base is < 7)

“*”: no appropriate reason and the following line shows bowtie alignment of split reads. For the meaning of these lines, please refer to the format specification of Bowtie alignment.

[5]: number of N

[6]: number of low quality base

[7]: number of phred score >=20 base

For the split reads in Bowtie alignment, the format specifications of each line as follow:

Eg:

FCD0JN9ACXX:1:1205:1718:99157#GCGGAACT/2_2 SE + IXR_BACseq 45380 57M MD:Z:2G1G0A51

FCD0JN9ACXX:1:1301:20767:17415#GCGGCACT/2_1 SE - IXR_BACseq 45633 45M MD:Z:45

Column ID: Meaning

[1]: a space;

[2]: read ID;

[3]: mapping tools;

[4]: SE/PE

SE: single-end, one of two ends of split read mapping

PE: pair-end, both of two ends of split read mapping

[5]: strand + or -;

[6]: chromosome ID;

[7]: coordinate of read mapping;

[8]: match pattern (CIGAR format)

[9]: mismatch positions