6th August 2010

Novoalign V2.07 & NovoalignCS V1.01 Release Notes

Novoalign V2.07

Illumina Mate Pair Support

We've had a good look at Illumina mate pair libraries and tried to put in the best support that we could.

Illumina mate pairs have a couple of issues resulting from the library preparation. This usually proceeds in several steps:

1. Primary fragmentation to around 3,000bp, the long insert size.
2. Circularisation with addition of Biotin tag at the junction.
3. Secondary fragmentation to 250-500 bp fragments. This steps fragments the circularised DNA and can produce three types of fragments, first is the true mate pairs where the circularisation junction is somewhere near the centre of the fragments, second is some fragments that don't overlap the junction, and third is fragments where the biotin labelled junction is nearer to one end of the fragment and will be inside one of the reads.
4. Enrichment for Biotin labelled fragments. This removes some but not all of the fragments that don't overlap the junction. Typically a mate pair library will have 20-50% of fragments that don't have the junction and map paired end reads with short inserts.

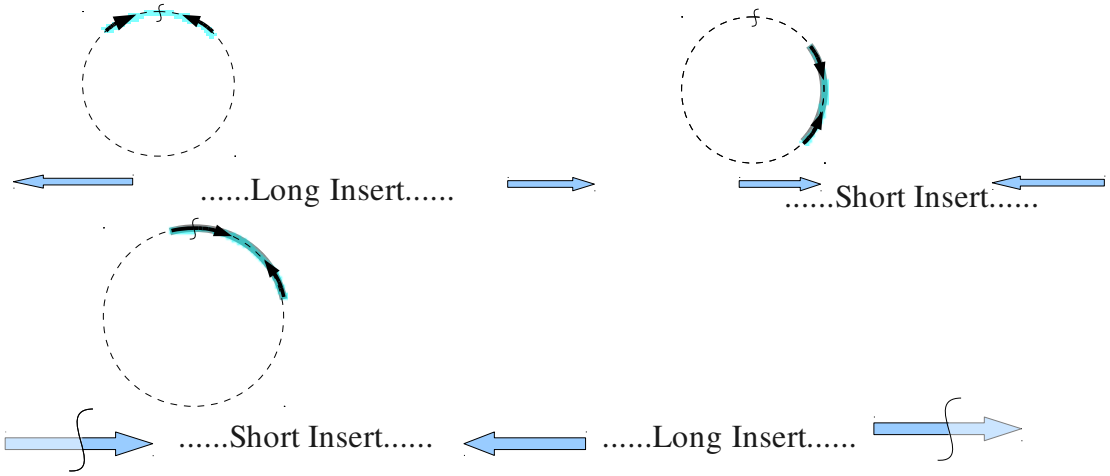
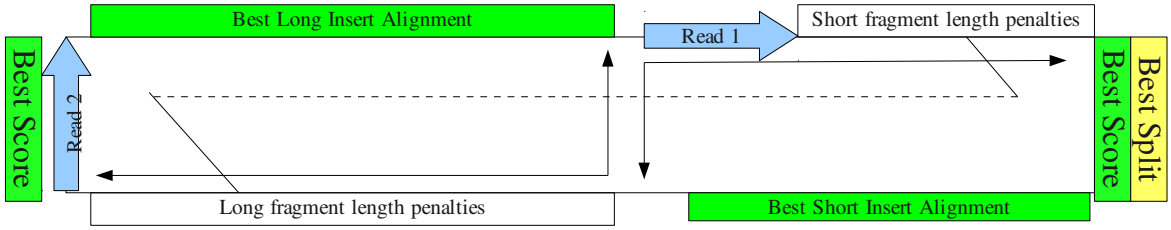


Figure 1. The three types of fragments, long insert, short insert, and fragments where one read overlaps the junction and results in split read alignment.

Split read alignments or split ends occurs when the circularisation junction is located in one of the reads. Illumina recommends using short reads (36bp) with large fragment sizes 500bp to reduce the probability of this happening to 36/500 = ~7% of reads. However, short fragments are not ideal when looking for structural variations or assembling genomes with high repeat content. Extra read length increases specificity however 70bp fragments would result in 14% of mate pairs having split ends, and if secondary fragmentation is reduced to 250bp then we could get 28% of reads with split end alignments.
Novoalign handles split end alignments using a modified Needleman-Wunsch alignment where reads are aligned as both mate pair and paired end and then best row scores are added to find the best split alignment score. The result will be the best of short insert, long insert or a split read alignment.



When a read alignment splits at the circularisation junction the alignment reported is to the longer portion of the mapping and the shorter portion is soft clipped from the alignment.

Mate pair alignment mode is enabled using the -i option. Two variations are available:

-i MP <*long insert mean*>,<*standard deviation*>
: When only one fragment length is specified reads will only be aligned as long inserts mate pairs.

-i MP <*long insert mean*>,<*standard deviation*> <*short insert mean*>,<*standard deviation*>
: When two insert lengths are specified then reads will be aligned as long inserts, short inserts or with split ends with both long and short inserts.

Example:

```
novoalign -d strep.nix -f SRR034554_1.fastq SRR034554_2.fastq -i MP 3500,600 250,80 -o SAM
```

Mate Pairs and Short Fragments

Secondary fragmentation can produce short fragments such as two reads overlap or even short enough that the reads extend into adapter sequence. Novoalign includes an option -a that detects short fragments and trims any adapter sequence. Normally this process leaves any overlap between the reads. The split end mapping process described above will not work if both reads of a pair are split across the circularisation junction, to avoid this situation the short read adapter trimming function now trims reads to remove any overlap as well as the adapter sequence.

454 Paired End Alignments

Along with changes to support mate pairs and ABI SOLiD colour space reads in NovoalignCS we

revisited code that pairs alignments and forms proper pairs. In Novoalign is pairing is specified as '+ +' then 454 paired end reads can be aligned.

Aligning 454 paired end reads is as simple as adding option the following to the novoalign command.

-i ++ *<mean insert size>,<standard deviation>*

Example using datset SRR001355 from NCBI short read archive:

novoalign -d ecolik12.nix -f SRR001355_1.fastq.gz SRR001355_2.fastq.gz **-i ++ 3300,1000**

Run log:

```
# novoalign (2.07.00MT - Aug  5 2010 @ 18:45:42) - A short read aligner
with qualities.
# (C) 2008 NovoCraft
# Licensed to Novocraft Internal Use
#  novoalign -d ecolik12 -f SRR001355_1.fastq SRR001355_2.fastq -i ++
3300,1000
# Interpreting input files as Sanger FASTQ.
# Index Build Version: 2.7
# Hash length: 9
# Step size: 1
#       Paired Reads:    62654
#      Pairs Aligned:    55810
#     Read Sequences:   125308
#            Aligned:   124027
#   Unique Alignment:   123305
#   Gapped Alignment:    13026
#     Quality Filter:        0
# Homopolymer Filter:        0
#       Elapsed Time: 85.824 (sec.)
#           CPU Time: 2.7 (min.)
```

Soft-clipped Alignments

The default for SAM report format is to soft trim alignments back to the best local alignment. This removes mismatches and indel s from near the ends of alignments and generally improves SNP & Indel calling with samtools and avoids some issues with Broad GATK suite of programs. This can be disabled with option **-o FullNW**

Read Segment Quality Control Indicator

Illumina uses a base quality of 2 to indicate a problem with calling of a read.

"*The Read Segment Quality Control Indicator: At the ends of some reads, quality scores are unreliable. Illumina has an algorithm for identifying these unreliable runs of quality scores, and we use a special indicator to flag these portions of reads with a quality score of 2, encoded as a "B", is used as a special indicator. A quality score of 2 does not imply a specific error rate, but rather implies that the marked region of the read should not be used for downstream analysis. Some reads will end with a run of B (or Q2) base calls, but there will never be an isolated Q2 base call*"

A quality of 2 is probability of error Perr = 0.63 and in earlier releases of Novoalign were scored as 4 for a match and 7 for a mismatch so they contributed slightly to alignment scores. This could cause a slight increase in SNP noise. In this release all bases with quality of 2 or less are treated as N's.

We also added an option, **-H** , to hard clipping of trailing bases with quality <=2 from reads. This just removes the clipped bases from the output reports and should have no affect on mapping.

NovoalignCS V1.01

Polyclonal Filter

This filter is designed to detect low quality reads and to stop any attempt to align them. This release changes how the option is specified and the default values.

-p 99,99 [0.9,99]	Sets thresholds for polyclonal filter. This filter is designed to remove reads that may come from polyclonal clusters or beads. Please refer to paper: ***Filtering error from SOLiD Output, Ariella Sasson and Todd P. Michael.*** The first pair of values (n,t) sets the number of bases and threshold for the first 20 base pairs of each read. If there are n or more bases with phred quality below t then the read is flagged as polyclonal and will not be aligned. The alignment status is 'QC'. The second pair applies to the entire read rather than just the first 20bp and is specified as fraction of bases in the read below the given quality. Setting **-p -1** disables the filter. Default for Novoalign is **off**. Default for NovoalignCS is -p 7,10 0.3,10. i.e 7 of first 20bp below Q10 or 30% of all bases below Q10 will be flagged as a low quality read. Low quality reads may still be used in paired end mode if the mate is not low quality.

Soft Clipped Alignments

As per Novoalign, the default for SAM report format is to soft trim alignments back to the best local alignment. This generally improves SNP & Indel calling with samtools and avoids some issues with Broad GATK suite of programs. This can be disabled with option **-o FullNW**

MPI Version of NovoalignCS

This release includes an MPI version of NovoalignCS allowing multiple servers in a cluster or network to be used to align one file of reads.