

NovoalignCS Quick Start

NovoalignCS is our aligner for colour space reads. Operation is similar to standard Novoalign and the two programs share quite a bit of code. NovoalignCS changes the file handling and alignment routines.

Novoindex

You need to build a colour space index for colour space reads. This index uses a hash table with colour space seeds rather than nucleotide seeds.

To construct a colour space index just add option -c to the Novoindex command, as in

```
novoindex -c genome.ncx *.fa
```

NovoalignCS

NovoalignCS command line is very similar to Novoalign. There are a few Novoalign features and options missing from NovoalignCS: adapter stripping, miRNA mode & Bisulphite mode.

Common Options:

Option	Description
-d <i>dbname</i>	Full pathname of indexed reference sequence from novoindex -c
-f <i>F3_seqfile1</i> [<i>R3_seqfile2</i>]	NovoalignCS accepts ABI Solid *.csfasta files with _QV.qual quality files or .csfastq files.
-t 99	Sets the threshold or highest alignment score acceptable for the best alignment. A default threshold is calculated from read length and genome size such that an alignment to a non-repeat should have a quality higher than 30.
-s 1	If a read is unaligned then shorten by 1 base and try again. This is useful for aligning short RNA reads. Suggested parameters for short RNA against Human are: novoalignCS -d -s 1 -l 14 -t 40 -f
-p 99,99 [<i>99,99</i>]	Sets thresholds for polyclonal filter. This filter is designed to remove reads that may come from polyclonal clusters or beads. Please refer to paper: <i>Filtering error from SOLiD Output, Ariella Sasson and Todd</i>

	<p><i>P. Michael</i></p> <p>Sets polyclonal filter thresholds. 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 the fraction of bases below a base quality. Setting -p -1 disables the filter. Default is -p 7,10 0.3,10 for 7 of first 20bp below Q10 or 30% of all bases below Q10.</p>
-o format [readgroup]	<p>Specifies the report format. Native, SAM, Pairwise. Default is Native.</p> <p>eg.</p> <p style="text-align: center;">novoalign -o SAM</p>
-i 99 99	Sets approximate fragment length and standard deviation. Default [2500, 500]
-k	Enables quality calibration. This is worth trying!
-K [file]	<p>Colour Error counts are written to the named file after all reads are processed.</p> <p>This file is useful for charting colour errors by base position in the read.</p>

File Formats

CSFASTA

If a csfasta file is specified as input NovoalignCS will look in the same fold for a quality file by replacing the .csfasta file extension with _QV.qual

```
>2_14_26_F3
T011213122200221123032111221021210131332222101
>2_14_192_F3
T110021221100310030120022032222111321022112223

*_QV.qual
>2_14_26_F3
24 24 22 27 23 10 13 13 20 19 19 18 24 20 22 12 14 5 20 17 14 20 18 17 19 11 21 19 13 13 12 25 9 19
19 6 5 12 20 13 11 8 12 7 14
>2_14_192_F3
14 19 21 13 24 17 18 18 25 21 8 12 21 8 7 11 14 7 19 23 11 24 7 11 29 12 28 17 7 19 7 11 5 11 5 14
13 9 24 8 7 20 0 8 9
```



1. BWA uses a csfastq format that includes a quality value for the primer base. This is typically coded as a '!' and is not used in alignment scoring.
2. BFAST has a fastq format that is similar to BWA format except that it does not have a quality for the primer base and hence the quality line is letter shorter than the read line. NovoalignCS does not support paired end reads in a single BFAST fastq file, it requires two files for paired end.

Colour space qualities are phred quality plus ascii '!'.
 The first column is the read ID, the second column is the read length, the third column is the read quality, the fourth column is the read sequence, the fifth column is the read position, the sixth column is the read quality, the seventh column is the read sequence, the eighth column is the read position, the ninth column is the read quality, the tenth column is the read sequence, the eleventh column is the read position, the twelfth column is the read quality, the thirteenth column is the read sequence, the fourteenth column is the read position, the fifteenth column is the read quality, the sixteenth column is the read sequence, the seventeenth column is the read position, the eighteenth column is the read quality, the nineteenth column is the read sequence, the twentieth column is the read position, the twenty-first column is the read quality, the twenty-second column is the read sequence, the twenty-third column is the read position, the twenty-fourth column is the read quality, the twenty-fifth column is the read sequence, the twenty-sixth column is the read position, the twenty-seventh column is the read quality, the twenty-eighth column is the read sequence, the twenty-ninth column is the read position, the thirtieth column is the read quality, the thirty-first column is the read sequence, the thirty-second column is the read position, the thirty-third column is the read quality, the thirty-fourth column is the read sequence, the thirty-fifth column is the read position, the thirty-sixth column is the read quality, the thirty-seventh column is the read sequence, the thirty-eighth column is the read position, the thirty-ninth column is the read quality, the fortieth column is the read sequence, the forty-first column is the read position, the forty-second column is the read quality, the forty-third column is the read sequence, the forty-fourth column is the read position, the forty-fifth column is the read quality, the forty-sixth column is the read sequence, the forty-seventh column is the read position, the forty-eighth column is the read quality, the forty-ninth column is the read sequence, the fiftieth column is the read position, the fifty-first column is the read quality, the fifty-second column is the read sequence, the fifty-third column is the read position, the fifty-fourth column is the read quality, the fifty-fifth column is the read sequence, the fifty-sixth column is the read position, the fifty-seventh column is the read quality, the fifty-eighth column is the read sequence, the fifty-ninth column is the read position, the sixtieth column is the read quality, the sixty-first column is the read sequence, the sixty-second column is the read position, the sixty-third column is the read quality, the sixty-fourth column is the read sequence, the sixty-fifth column is the read position, the sixty-sixth column is the read quality, the sixty-seventh column is the read sequence, the sixty-eighth column is the read position, the sixty-ninth column is the read quality, the seventieth column is the read sequence, the seventy-first column is the read position, the seventy-second column is the read quality, the seventy-third column is the read sequence, the seventy-fourth column is the read position, the seventy-fifth column is the read quality, the seventy-sixth column is the read sequence, the seventy-seventh column is the read position, the seventy-eighth column is the read quality, the seventy-ninth column is the read sequence, the eightieth column is the read position, the eighty-first column is the read quality, the eighty-second column is the read sequence, the eighty-third column is the read position, the eighty-fourth column is the read quality, the eighty-fifth column is the read sequence, the eighty-sixth column is the read position, the eighty-seventh column is the read quality, the eighty-eighth column is the read sequence, the eighty-ninth column is the read position, the ninetieth column is the read quality, the ninety-first column is the read sequence, the ninety-second column is the read position, the ninety-third column is the read quality, the ninety-fourth column is the read sequence, the ninety-fifth column is the read position, the ninety-sixth column is the read quality, the ninety-seventh column is the read sequence, the ninety-eighth column is the read position, the ninety-ninth column is the read quality, the hundredth column is the read sequence.

BWA Type CSFASTQ

BFAS_T Type CSFAS_TQ

Two files can be specified for paired end mode. In this case Novoalign parses the header records looking for a header in standard ABI format (e.g. >2_14_26_F3). If found then headers from the two files are assumed to be in order and matched for purpose of identifying paired reads. Reads that exist in only one file will be aligned in single end mode.

Report Formats

SAM Format

SAM format follows SAM specifications including colour space specific tags.



Native Format

Two extra columns are added for the read in Nucleotide space and the nucleotide qualities.