NovoalignCS Quick Start

NovoalignCS is our aligner for colour space reads and can be found in version 3 and earlier releases. 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 *dbname*	Full pathname of indexed reference sequence from novoindex -c
-f F3*_seqfile1* [R3*_seqfile2*]	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 [99,99]	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*

	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.
-o *format* [*readgroup*]	Specifies the report format. **Native, SAM, Pairwise.** Default is Native. eg. novoalign -o SAM
-i 99 99	Sets approximate fragment length and standard deviation. Default [2500, 500]
-k	Enables quality calibration. This is worth trying!
-K [file]	Colour Error counts are written to the named file after all reads are processed. This file is useful for charting colour errors by base position in the read.

File Formats

CSFASTA

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

```
>2_14_26_F3
T0112131222002211230321112210212101313322221O1
>2_14_192_F3
T1100212211003100301200220322211132102211222 3
```

*_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
```

Color Space FASTQ

There are two variations of colour space fastq files being used by other aligners.

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.

Novoalign supports both formats and can automatically distinguish the two types based on the length of the quality string on the first line. You can also specify format on the command line using -F BFASTQ or -F CSFASTQ

Colour space qualities are phred quality plus ascii '!'.

BWA Type CSFASTQ

```
@SRR015241.1 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_34_F3 length=50
T3232213330000233003100102223002023200220322203023 1
+SRR015241.1 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_34_F3 length=50
!21(()+%'+%40*.%%**)&%&*&%%%&%%%%%%%%%%%%%%%%(+%%%%'
@SRR015241.2 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_269_F3 length=50
T0121212033322332202002232223223223222022232033230
+SRR015241.2 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_269_F3 length=50
!,*+*()+*(%'+)%%%&%+&%%'%%%%%%%%%%%%%%%%%%%%'+%%%%%
```

BFAST Type CSFASTQ

```
@SRR015241.1 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_34_F3 length=50
T3232213330000233003100102223002023200220322203023 1
+SRR015241.1 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_34_F3 length=50
21(()+%'+%40*.%%**)&%&*&%%%&%%%%%%%%%%%%%%%%(+%%%%'
@SRR015241.2 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_269_F3 length=50
T0121212033322332202002232223223223222022232033230
+SRR015241.2 CLARA_20071207_2_CelmonAmp7797_16bit_26_88_269_F3 length=50
,*+*()+*(%'+)%%%&%+&%%'%%%%%%%%%%%%%%%%%%%%'+%%%%%
```

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.

Requirements

HDF5

1. Download the source code of HDF5 1.8.8 (we haven't tested later releases) from HDF5 web site
2. Open terminal and extract the tar
3. ./configure –prefix=/usr/local –enable-cxx
4. make
5. make check
6. sudo make install