NONACUS
Advancing Non-Invasive Healthcare

Cell3 Target Data Analysis Guidelines

This guide provides instructions for bioinformaticians on batch processing data files following sequencing of libraries prepared using Nonacus' Cell3 Target NGS kit and subsequent de-multiplexing using Illumina's bcl2fastq tool.

Bcl2fastq De-multiplexing

Input requirements

1. Illumina bcl2fastq software – (BCL) Base Call Files to Fastq conversion software http://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html
2. Basecalls directory
3. Samplesheet.csv – ensure there are no N's present in the sample sheet (examples provided on request)
4. Check RunInfo.xml to ensure the correct number of cycles have been performed for the indexes
    a. i7 = 17 cycles
    b. i5 = 8 cycles

Expected file outputs per sample

1. Read 1 fastq file (R1)
2. Read 2 fastq file (R3)
3. fastq file containing information on a per fragment molecular tag (R2)
4. Index 1 and 2 fastq files (these are not required)

Procedural steps

1. Open a command terminal
2. Move into the basecalls directory of the data to be de-multiplexed
3. Run the following bcl2fastq command

    bcl2fastq --create-fastq-for-index-reads --mask-short-adapter-reads 0 --use-bases-mask Y*,I8Y9,I8,Y* --no-lane-splitting (optional)

    * replace with the number of cycles performed for read 1 and 2

NonacusTools a Consensus BAM File Preparation Tool

NonacusTools automates the preparation of consensus bam files utilising open source tools. The following is a list of tools used and steps implemented within the script.

Step No.	Step Name	Tool	Version	Input Files	Output Files	Additional Settings	Extra Files
1	Alignment	bwa mem	0.7.13	R1/R3	.sam	-M -t 16	GRCh38.fasta.gz
2	Annotate BAM with UMI	fgbio AnnotateBamWithUmis	0.4.0	.bam / I1.fastq	fgtag_.bam	-f I1.fastq	
3	Sort bam by queryname	fgbio SortBam	0.4.0	fgtag_.bam	fgsort_.bam	-s queryname	
4	Reset mate information	fgbio SetMateInformation	0.4.0	fgsort_.bam	setmate_.bam		
5	Group read by UMI	fgbio GroupReadsByUmi	0.4.0	setmate_.bam	histogram.txt fggroup_.bam	-f family_size_histogram.txt -s adjacency	
6	Consensus read build	fgbio CallMolecularConsensusReads	0.4.0	fggroup_.bam	fgcon_.bam	-M 2	
7	Covert Consensus to Fastq	PICARD SamtoFastq	2.2.2	fgcon_.bam	con_R1.fastq / con_R2.fastq	FASTQ=con_R1.fastq SECOND_END_FASTQ=con_R2.fastq VALIDATION_STRINGENCY=LENIENT	
8	Align consensus bam	bwa mem	0.7.13	con_R1.fastq / con_R2.fastq	con_.sam		GRCh38.fasta.gz
9	Sort consensus bam	PICARD SortSam	2.2.2	con_.sam	consensus_.bam	SORT_ORDER=coordinate	GRCh38.fasta.gz
10	Index consensus bam	samtools index	1.3.1	consensus_.bam	consensus_.bai		

NonacusTools Usage

1. Contact nonacus support for a copy of NonacusTools.tar.gz (5.4Gb) the download includes the GRCh38 reference bundle
2. Unzip the file using tar

   ```
   tar -xvzf file.tar.gz -C path/to/where/to/save/tool
   ```

3. cd (what does cd mean?) into the NonacusTools directory
4. Create a tab delineated text file or batch file in the below format and store in the same folder as the fastq files you want to process

   R1_file_name R3_file_name R2_file_name ID

5. Run the following command including the full file path to the batch file and specify the number of cores to assign to the processing through option -t

   ```
   Python consensus.py -i full/path/to/batch_file.txt -t 24
   ```

6. The output will be a file for each sample containing the following files:
   a. Final_Consensus_fileName.bam + Final_Consensus_fileName.bai
   b. Tag_fileName.bam + Tag_fileName.bai - UMI placed in RX tag
   c. Stats_fileName.txt – contains information relating to the family size, count and fraction