# Abstract

For most applications, short reads generated by next-generation sequencing technologies require pre-processing. This SOP describes a basic pre-processing workflow for data generated by CRC projects – including identification of potential issues with FastQC, cleaning of reads using trimmomatic and optional error correction for downstream assembly applications.

# Introduction

Short reads generated by next-generation sequencers will typically include non-genomic/non-transcriptomic nucleotides added by the sequencing library preparation (barcodes, sequencing adapters). In addition, quality of sequenced reads may vary along the length of individual reads, sometimes falling below acceptable values for certain applications. This SOP outlines the relevant steps required to clean read files for downstream processing.

# Requirements

## Input data

Paired-end reads generated on the Illumina Platform (HiSeq 2500, HiSeq 3000, HiSeq 4000, NextSeq and MiSeq).

## Software

Trimmomatic (0.36)

Trowel (0.2.0.4)

FastQC (0.11.3)

# Procedure

## Quality statistics and identification of issues

FastQC analyses individual FASTQ files to identify a range of potential errors, including low quality bases, presence of known adapter sequences and overrepresented kmers (identifying e.g. unclipped barcodes or unknown adapters).

FastQC results are normally delivered by the sequencing provider together with the raw data. If not, the syntax for generating such a report is as follows:

fastqc –o fastqc\_analysis <input file>

where <input file> is either the left or right mate (or single-end) fastq file (preferably in compressed gzip format). The results are available as HTML file (readable by any web browser) as well as a series of log files, compressed into a zip archive. A detailed explanation of how to read the output is available on the FastQC website: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

## Read trimming

Based on the results from the FastQC analysis, it may be necessary to clean the data. We recommend Trimmomatic as it, unlike most other tools, provides a reference file with commonly used sequencing adapters. Trimmomatic can perform a range of correction steps, including clipping of adapters as well as low quality bases. Note that it is NOT recommended to perform quality trimming if the data ist to be used for de-novo assembly (genome or transcriptome). However, if the data will be used for highly sensitive variant calling (or similar), it is recommended to perform stringent clipping of all sub-par bases.

A basic Trimmomatic command will like as follows:

## Error correction

Some assembly programs will benefit from error-corrected reads. This procedure is different from trimming and aims to correct individual positions without clipping. Note that some assemblers already include their own error correction module, like Spades or Allpaths-LG – additional correction will likely decrease assembly performance.

Trowel has proven as the most performant and easy to install option for the purpose of error correcting reads for assembly.

# Implementation

This workflow is implemented as pipeline “ReadPreprocessing.bpipe”, included with the CRC git code repository.

# Related Documents

# Revision history