Setting up a portable metagenome assembly pipeline for CBIO - Support #45

Quality trimming with Trimmomatic

04/26/2018 04:39 PM - Katie Lennard

Status:	New	**Start date:**	08/30/2017
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour

Description

In short after a bit of a struggle to optimize parameters for quality trimming with Trimmomatic I recommend rather using cutadapt, which has better documentation and more predictable behaviour. Nevertheless, the current pipeline test was run with Trimmomatic and below is a short summary of the steps taken to perform quality filtering and remove non-genomic sequences:
*Usually Ulas just performs a fixed length trimming on reads:
"typically, if the quality looks good across the reads, what I
do is to do a fixed length trimming. ~5-10 bp from 5' and ~15bp from
3'end. That is a quick and easy way to both QC and get rid of any
non-genomic sequences (they are usually enriched in 5' to 3' ends) You
can do that prinseq-lite.pl" - so don't even need trimmomatic for this
However, for our test data the quality was poor with huge
variation in the position when the quality degrades, which meant we had to instead do a 'window-based' trimming

*Post-trimming FastQC of the 1st test showed enrichment of (probably) non-genomic reads (NB: If the amount of non-genomic sequences are a big chunk of the input, that usually messes up the assembler heuristics)
*However even when including the Illumina adapters file (NexteraPE-PE.fa) FastQC still lists enrichment with kmers, vastly overrepresented, indicative of non-genomic sequences. This overrepresented sequence (only present in forward reads and not reverse complement of reverse reads) was identified from
https://github.com/tomdeman-bio/Sequence-scripts/blob/master/adapters.fasta
as an Illumina-specific sequence Illumina gnl|uv|NGB00755.1|TruSeq_DNA_HT_and_RNA_HT_i7_xxx, part of the i7 multiplexing technology
*The 'minAdapterLength' parameter (see below) was then reduced to 6 (from the default 8) which seemed to remedy this issue.
*The final run was: java -jar $trimmomatic_path PE -threads $trimmomatic_threads -trimlog $trim_logfile $F_fastq $R_fastq $F_paired $F_unpaired $R_paired $R_unpaired ILLUMINACLIP:$Illumina_adapters:2:30:10:6:true SLIDINGWINDOW:4:15 MINLEN:$MINLEN

- See http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf for details on parameter specifications. (NB: the ILLUMINACLIP parameters and could use further optimization should be adjusted according to the length of the adapter sequence (e.g. a perfect match of a 12 base sequence will score just over 7, while 25 bases are needed to score 15)

History

#1 - 04/26/2018 05:17 PM - Katie Lennard

- Tracker changed from Bug to Support

- Description updated