Command Workflow

Max Barclay

2022-11-15

Creating the environment

This is the set of commands used to create the environment used within the CLIMB server for this analysis. The environment itself is specifically a conda environment using mamba as a package installer. Once installed, necessary channels were added and ordered as needed, then set the channel priority to strict. Then, the software needed installed

```
conda create -n strep_project python=3.10 mamba -y
conda activate strep_project

conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
conda config --set channel_priority strict

mamba install snippy trimmomatic fastqc -y
mamba install -c bioconda snpEff=4 -y
```

Using fastqc and trimmomatic

On the CLIMB server, fastqc cannot open its GUI. To circumvent this the query files were added into the command instead of opening within the GUI, and they were output into their own fastqc_report folder.

Still within the CLIMB server, using trimmomatic was used for trimming these reads, specifically the paired end outputs. Its own directory was made 'trimmed_reads', and parameters were set for 4 threads, using 2 computer CPUs. ILLUMINACLIP was used to cut adapters and other Illumina-related sequences from the read, using the TruSeq3.fa file for identifying where the adapters are. This is followed by the 'seed mismatch value', which species the maximum mismatch count where full matches are still allowed, and then the 'palindrome clip threshold', which specifies the match accuracy between the two reads before consideration as a paired end palindrome read for alignment. The final value is the 'simple clip threshold', which specifies the how accurate the match between the adapter and the sequence its read against must be.

LEADING is responsible for removing the low quality bases seen at the start, as long as it is a value below the value that is set. TRAILING, does the same, except for low quality bases at the end. MINLEN is repsonsible for removing reads that fall below the minimal length value set, normally carried out as the last trimmomatic step.

The PHRED quality was not specified between either 33 and 64, with trimmomatic running with TOPHRED33 in this case.

These commands were all carried out within the individual directories that required the analysis'.

```
mkdir fastqc_reports
fastqc -outdir fastqc_reports Y1_P188_02_FDSW220035690-1r_HMYW7DSX2_L2_1.fq.gz \
Y1_P188_02_FDSW220035690-1r_HMYW7DSX2_L2_2.fq.gz
```

```
mkdir -p trimmed_reads
trimmomatic PE -threads 4 \
  Y1_P188_02_FDSW220035690-1r_HMYW7DSX2_L2_1.fq.gz \
  Y1_P188_02_FDSW220035690-1r_HMYW7DSX2_L2_2.fq.gz \
    trimmed_reads/Y1_P188_02_forward_paired.fq.gz \
      trimmed_reads/Y1_P188_02_forward_unpaired.fq.gz \
    trimmed_reads/Y1_P188_02_reverse_paired.fq.gz \
      trimmed_reads/Y1_P188_02_reverse_unpaired.fq.gz \
    ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36
```

Mapping out trimmed reads, and calling variants

The trimmomatic reads were mapped against the progenitor NCBI strain and the differences between the generationally different strains and the progenitor strain using snippy. These were put into their own directory to follow tidy data practices, and all working on the command line was carried out in the CLIMB server and within the necessary directory were analysis was being carried out.

```
snippy --cpus 2 --outdir variant_calling \
  --ref reference_strain_data/M1152_NC_003888_genbank_JTM_edit_SNIPPY_progenitor.gb \
  --R1 trimmed_reads/Y1_P188_02_forward_paired.fq.gz \
  --R2 trimmed_reads/Y1_P188_02_reverse_paired.fq.gz
```

This process was then carried out another two times for the other Y1_P188_05, and Y1_P188_15 strains, following the same steps except for replacing the name to match the specific strain (i.e Y1_P188_05 instead of Y1_P188_02) and working within their matching directories instead.

Further analysis

Beyond this, the output snps.tab, snps.html, snps.bam, and ref.fa files were all used to analysis the identified variants from the variant calling. The snps.bam and ref.fa were specifically put into Tablet for visualisation of this.