Full reproducible codes

1) Create directory 'data', 'src', 'results' in a CA2 folder

(base) weiqi@LSLab5-22:~$ cd Desktop

(base) weiqi@LSLab5-22:~/Desktop$ cd CA2

(base) weiqi@LSLab5-22:~/Desktop/CA2$ mkdir data src results

2) Download relevant packages

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install curl

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install fastqc

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install trimmomatic

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install picard-tools

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install bowtie bowtie2

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install samtools

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install libbz2-dev

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install zlib1g-dev

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install libncurses5-dev

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install libncursesw5-dev

(base) weiqi@LSLab5-22:~/Desktop/CA2$ sudo apt-get install liblzma-dev

(base) weiqi@LSLab5-22:~/Desktop/CA2$ wget https://github.com/samtools/samtools/releases/download/1.9/(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools-1.9.tar.bz2

(base) weiqi@LSLab5-22:~/Desktop/CA2$ tar -vxjf samtools-1.9.tar.bz2

(base) weiqi@LSLab5-22:~/Desktop/CA2$ cd samtools-1.9

(base) weiqi@LSLab5-22:~/Desktop/CA2/samtools-1.9$ make

(base) weiqi@LSLab5-22:~/Desktop/CA2/samtools-1.9$ sudo make install

(base) weiqi@LSLab5-22:~/Desktop/CA2/samtools-1.9$ sudo apt-get install bedtools

3) Download mutated yeast clone (A0158131M.tgz) from nBox and put it in data folder

4) Prepare data by extracting out all the files from the "tarball"

(base) weiqi@LSLab5-22:~/Desktop/CA2/samtools-1.9$ cd ../data

(base) weiqi@LSLab5-22:~/Desktop/CA2/data$ tar -zxvf A0158131M.tgz

5) Quality control - Assessing quality using FastQC

(base) weiqi@LSLab5-22:~/Desktop/CA2/data$ cd ../results

(base) weiqi@LSLab5-22:~/Desktop/CA2/results$ mkdir fastqc

(base) weiqi@LSLab5-22:~/Desktop/CA2/results$ cd ../

(base) weiqi@LSLab5-22:~/Desktop/CA2$ fastqc data/*.fq -o results/fastqc

(base) weiqi@LSLab5-22:~/Desktop/CA2$ cd results/fastqc

(base) weiqi@LSLab5-22:~/Desktop/CA2/results/fastqc$ for filename in *.zip; do unzip $filename; done

(base) weiqi@LSLab5-22:~/Desktop/CA2/results/fastqc$ cd ../..

(base) weiqi@LSLab5-22:~/Desktop/CA2$ mkdir -p docs

(base) weiqi@LSLab5-22:~/Desktop/CA2$ cat results/fastqc/*/summary.txt > docs/fastq_summaries.txt

6) Check for any failed category: No failed results

(base) weiqi@LSLab5-22:~/Desktop/CA2$ grep FAIL docs/fastq_summaries.txt

7) Assess HTML files and decided not to trim as all categories passed.

8) Align reads to reference genome - prepare by downloading yeast genome (sacCer.fa) and transposon (ty5_6p.fa) from IVLE, and put it in data folder

(base) weiqi@LSLab5-22:~/Desktop/CA2$ mkdir combined

(base) weiqi@LSLab5-22:~/Desktop/CA2$ mkdir -p results/sam results/bam

(base) weiqi@LSLab5-22:~/Desktop/CA2$ bowtie2-build 'data/sacCer3.fa,data/ty5_6p.fa' combine.indexing

9) Put combine.indexing files into a new folder called “combined”

10) Next, prepare 3 bash(.sh) scripts and put them in “src” folder. The scripts can be found in the github repository too.

11) Align paired-ends to reference genome

(base) weiqi@LSLab5-22:~/Desktop/CA2$ bash src/run-bowtie2.sh

12) Convert SAM file format to BAM file format

(base) weiqi@LSLab5-22:~/Desktop/CA2$ bash src/convert-to-bam.sh

13) Sort BAM file by coordinates

(base) weiqi@LSLab5-22:~/Desktop/CA2$ bash src/sort-bam-files.sh

14) Learn more about the BAM file

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools flagstat results/bam/data-sorted.bam

15) Using SAM flags to filter to get improperly paired, and read and mates mapped

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools view -h -f 1 -F 14 results/bam/data-sorted.bam | grep -e 'TY5' -e '^@' > data-sorted.sam

(base) weiqi@LSLab5-22:~/Desktop/CA2$ head data-sorted.sam

16) Make it into a .txt file so as to open on excel sheet to further filter and narrow down the regions to find the paired-ends and insertions on IGV later on

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools view -h -f 1 -F 14 results/bam/data-sorted.bam | grep -e 'TY5' -e '^@' > data-sorted.txt

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools view -S -b data-sorted.sam > data-sorted.bam

17) Index BAM files to allow visualisation of alignment on IGV

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools index results/bam/data-sorted.bam

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools index data-sorted.bam

18) Check the number of reads; 150

(base) weiqi@LSLab5-22:~/Desktop/CA2$ samtools view data-sorted.bam | wc -l