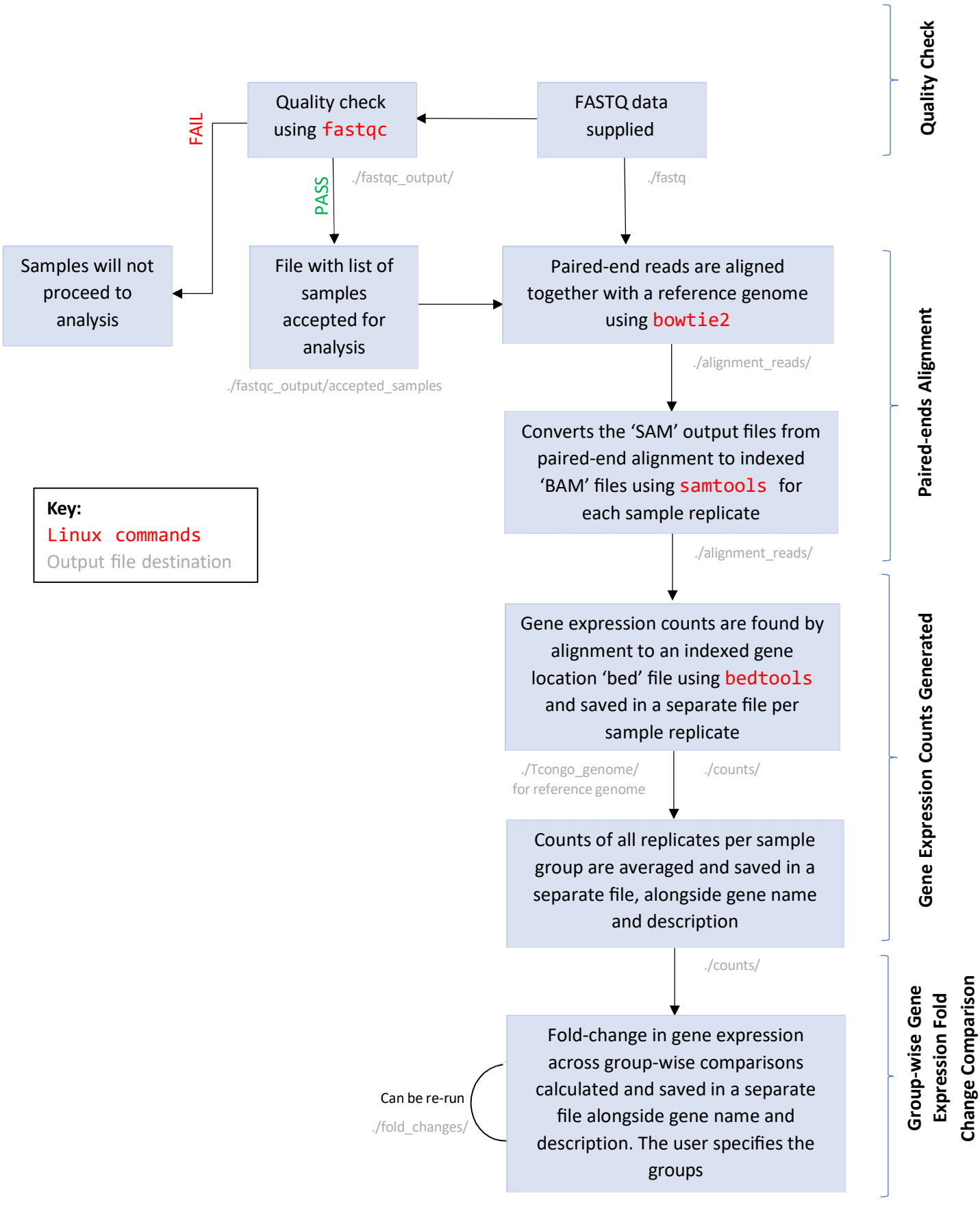
Pipeline Flowchart:



Programme Parameters:

fastqc -o: Saves all the outputs of the quality check to a separate folder

bowtie2-build: Indexes the reference genome into .bt2 format to be used for alignment using bowtie2

bowtie2 -x -1 -2: Used to specify which files are the reference genome, and which are the two pair ends

samtools view -bS: Indicates that the input file is in SAM format, while the output will be BAM format.

samtools sort -o: Sorts the BAM alignments by the reference genome, and outputs the file to a specified directory

samtools index: Creates indexes for the generated BAM alignment files

bedtools multicov: BAM focused tool of bedtools that can take many BAM inputs at once to be able to generate one file with counts of all replicates per sample group

Instructions for user:

All script files to be run on Bash on Linux. 'general_analysis.sh' and 'fc_calculator.sh' contain the code for the pipeline analysis of the FASTQ data. The file 'general_analysis.sh' should be run first by executing "./general_analysis.sh", and completes all the steps of copying the required data to the local directory, performing quality check, aligning the data and generating average gene expression counts across each sample group. This file will ask for the maximum number of fails in quality check you would allow for the sample to proceed to analysis, to identify the threshold.

Secondly, the file 'fc_calculator.sh' should be executed through "./fc_calculator.sh", and this deals with calculating the fold change across two sample groups, and saves a newly generated file with these values for each gene. Running this file requires the user to input the two sample groups they wish to compare, in the format involving the two sample types, two treatments, and two time conditions. For example if wanting to compare Clone1 Induced 24h to Clone1 Uninduced 24h, then Clone1,Clone1,Induced,Uninduced,24,24 should be inputted. Please note that the average counts of the first set of conditions will be divided by the second set of conditions to find the fold change in gene expression. If you are sure that the supplied samples have passed quality check and have been inputted in the correct format but you have still received an error, please **copy and paste** the exact same Clone1,Clone1,Induced,Uninduced,24,24 after being prompted when rerunning the file, as this seems to work better. This file can be re-run for desired group-wise comparisons.

Furthermore, some commands may take a long time to run. Once executed, more directories and files will be created in the local directory of the user.